Examination of statistical methods to analyse interrupted time series studies in public health

Simon Lee Turner

BSc(Hons), BTeach(Hons), MSc(Astronomy), MBiostat

A thesis submitted for the degree of
Doctor of Philosophy at Monash University in 2020
School of Public Health and Preventive Medicine

Copyright notice

Contents

I. Abstract

Introduction:

Interrupted time series (ITS) designs are frequently used in public health research to examine whether an intervention or exposure has impacted on health outcomes. However, little guidance is available for choosing the most appropriate statistical method for analysis or how best to visually report the results of the analysis. One key aspect necessary to account for in ITS studies is the correlation of data points over time, known as autocorrelation. The magnitude of autocorrelation impacts on sample size calculations for planning ITS studies in addition to the performance of statistical methods, yet there is little guidance in the literature as to its typical values.

Aim:

The aim of the research presented in this thesis was to assess the design, reporting quality and statistical methods used in ITS studies, and to subsequently provide tools and guidance that may facilitate improvement. The thesis focused on ITS studies in public health.

Methods:

Several methodologies were used in this thesis. A review was undertaken to examine the design characteristics and statistical methods used in ITS evaluating public health interruptions. Information was extracted on study characteristics, statistical models, estimation methods, effect metrics, parameter estimates, and presentation of results. Seminal data visualisation resources were examined to inform the development of recommendations for graphically presenting data and results from ITS studies. Using the graphs identified in the review, an assessment was made as to whether the graphs met the recommendations. A simulation study was undertaken to investigate the performance of statistical methods commonly used for analysing ITS data. Data were simulated under a range of conditions, which were informed by the review. An empirical study was undertaken to investigate the impact of using the different statistical methods on real-world data, and estimates of autocorrelation were calculated.

Results:

The review identified 200 ITS studies, comprising 230 separate time series, and determined that the reporting of ITS studies could be substantially improved, due to statistical analysis methods (especially the handling of autocorrelation) often being inadequately described. Assessment of the quality of the time series graphs included in the review showed that they were frequently missing key components. Findings from the simulation study showed that autocorrelation was underestimated by all statistical methods. Furthermore, while all statistical methods yielded unbiased estimates of the effects of the interruption under study, they differed in their ability to accurately estimate the standard errors, particularly in the presence of autocorrelation. Findings from the empirical study demonstrated that autocorrelation is frequently present.

Conclusion:

The research in this thesis demonstrates that there is substantial need to improve the quality of reporting of statistical analysis methods in ITS studies. Implementation of the proposed graphing recommendations may facilitate improved interpretation. Autocorrelation was found to be frequently present, and as such, this should be considered in the design, analysis and interpretation of findings of ITS studies in public health. Further research is required to develop statistical methods for ITS analysis that handle autocorrelation more appropriately.

II. Thesis including published works declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes three original papers published in peer reviewed journals and two submitted publications. The core theme of the thesis is the examination of statistical methods to analyse interrupted time series studies in public health. The ideas, development and writing of all the papers in the thesis were the principal responsibility of myself, the student, working within the School of Public Health and Preventive Medicine under the supervision of Associate Professor Joanne McKenzie, Professor Andrew Forbes and Dr Amalia Karahalios.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.

In the case of chapters two to six my contribution to the work involved the following:

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution	Co-author(s), Monash student Y/N
2	Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: protocol for a review	Published in peer reviewed journal: *BMJ Open*	80%. Led the design of the review in collaboration with JEM. Led the creation of data collection items with contributions from JEM, AK, ABF. Drafted the manuscript with contributions from JEM. Made subsequent revisions based on critical review from all co-authors.	1) Joanne E. McKenzie 4%. Contributed to the conception and design of the study and drafting of the manuscript, and provided critical review of the manuscript. 2) Amalia Karahalios 4%. Contributed to the design of the study and drafting of the manuscript, and provided critical review of the manuscript. 3) Andrew B. Forbes 4%. Contributed to the design of the study and drafting of the manuscript, and provided critical review of the manuscript. 4) Monica Taljaard 2% Contributed to the design of the study, and provided critical review of the manuscript. 5) Jeremy M. Grimshaw 2%. Contributed to the design of the study, and provided critical review of the manuscript. 6) Allen C. Cheng 2%. Contributed to the design of the study, and provided critical review of the manuscript. 7) Lisa Bero 2%. Contributed to the design of the study, and provided critical review of the manuscript.	No for all

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution	Co-author(s), Monash student Y/N
3	Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: A review	Published in peer reviewed journal: *Journal of Clinical Epidemiology*	80%. Led the design of the review, searching, screening and data extraction with JEM, AK, ABF. Drafted the manuscript with contributions from JEM. Made subsequent revisions based on critical review from all co-authors.	1) Joanne E. McKenzie 4%. Contributed to the design of the study and drafting of the manuscript, screened and extracted data, and provided critical review of the manuscript. 2) Amalia Karahalios 4%. Contributed to the design of the study and drafting of the manuscript, screened and extracted data, and provided critical review of the manuscript. 3) Andrew B. Forbes 4%. Contributed to the design of the study and drafting of the manuscript, screened and extracted data, and provided critical review of the manuscript. 4) Monica Taljaard 2%. Contributed to the design of the study and provided critical review of the manuscript. 5) Jeremy M. Grimshaw 2%. Contributed to the design of the study and provided critical review of the manuscript. 6) Allen C. Cheng 2%. Contributed to the design of the study and provided critical review of the manuscript. 7) Lisa Bero 2%. Contributed to the design of the study and provided critical review of the manuscript.	No for all

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution	Co-author(s), Monash student Y/N
4	Creating effective interrupted time series graphs: review and recommendations	Published in peer reviewed journal: *Research Synthesis Methods*	80%. Conceived and designed the study, reviewed data visualisation resources, proposed the first set of recommendations and generated the computer code. Led the extraction of the data with AK, EK, JEM. Drafted the manuscript with contributions from JEM. Made subsequent revisions based on critical review from all co-authors.	1) Joanne E. McKenzie 3%. Contributed to the design of the study and drafting of the manuscript, assessed the recommendations, extracted data and provided critical review of the manuscript. 2) Amalia Karahalios 3%. Contributed to the design of the study and drafting of the manuscript, assessed the recommendations, extracted data and provided critical review of the manuscript. 3) Andrew B. Forbes 3%. Contributed to the design of the study and drafting of the manuscript, assessed the recommendations, and provided critical review of the manuscript. 4) Monica Taljaard 2%. Assessed the recommendations, and provided critical review of the manuscript. 5) Jeremy M. Grimshaw 2%. Assessed the recommendations, and provided critical review of the manuscript. 6) Allen C. Cheng 2%. Assessed the recommendations, and provided critical review of the manuscript. 7) Lisa Bero 2%. Assessed the recommendations, and provided critical review of the manuscript. 8) Elizabeth Korevaar 3%. Assessed the recommendations, extracted data and provided critical review of the manuscript.	1) - 7) No 8) Yes

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution	Co-author(s), Monash student Y/N
5	Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study	Submitted to *Biometrical Journal*	80%. Led the design of the study in collaboration with JEM and ABF. Designed the computer code, ran and analysed the simulations. Drafted the manuscript with contributions from JEM. Made subsequent revisions based on critical review from all co-authors.	1) Joanne E. McKenzie 6%. Contributed to the design of the study and drafting of the manuscript, and provided critical review of the manuscript. 2) Amalia Karahalios 6%. Contributed to the design of the study and drafting of the manuscript, and provided critical review of the manuscript. 3) Andrew B. Forbes 6%. Contributed to the design of the study and drafting of the manuscript, and provided critical review of the manuscript. 4) Monica Taljaard 2%. Provided critical review of the manuscript.	No for all
6	Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series	Submitted to *BMC Medical Research Methodology*	Led the design of the study in collaboration with JEM. Contacted authors and digitally extracted data to create data sets. Designed and ran the computer code, analysed the data. Drafted the manuscript with contributions from JEM. Made subsequent revisions based on critical review from all co-authors.	1) Joanne E. McKenzie 6%. Contributed to the design of the study and drafting of the manuscript, and provided critical review of the manuscript. 2) Amalia Karahalios 5%. Contributed to the design of the study and drafting of the manuscript, and provided critical review of the manuscript. 3) Andrew B. Forbes 5%. Contributed to the design of the study and drafting of the manuscript, and provided critical review of the manuscript. 4) Monica Taljaard 2%. Provided critical review of the manuscript. 5) Jeremy M. Grimshaw 2%. Provided critical review of the manuscript.	No for all

Student name: Simon Lee Turner

Student signature: **Date: 30/11/2020**

I hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor name: Joanne McKenzie

Main Supervisor signature: **Date: 30/11/2020**

III. List of research outputs

Listed below are the candidate's first-author publications and conference proceedings that are relevant to the period of candidature.

Publications relevant to the thesis:

Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, McKenzie JE. Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series. Submitted to *BMC Medical Research Methodology*.

Turner SL, Forbes AB, Karahalios A, Taljaard M, McKenzie JE. Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study. Submitted to the *Biometrical Journal*, preprint available (doi.org/10.1101/2020.10.12.20211706).

Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Korevaar E, Cheng AC, Bero L, McKenzie JE. Creating effective interrupted time series graphs: review and recommendations. *Research Synthesis Methods*, 2020, 1-12.

Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Cheng AC, Bero L, McKenzie JE. Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: a review. *Journal of Clinical Epidemiology*, 2020. **122**: p. 1-11.

Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Cheng AC, Bero L, McKenzie JE. 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.

Conference proceedings:

Turner SL, Karahalios A, Forbes AB, McKenzie JE. "The use of interrupted time series designs in the evaluation of public health interventions." Oral presentation at the Joint International Society for Clinical Biostatistics and Australian Statistical Conference, 26-30 August 2018, Melbourne, Australia.

IV. Acknowledgements

I would like to wholeheartedly thank my supervisors. Associate Professor Jo McKenzie, Dr. Amalia Karahalios and Professor Andrew Forbes. Jo recommended a project that suited my interests and enabled this amazing journey, thank you for being patient, supportive and so highly skilled (especially at editing my writing) – I look forward to continuing to work in this fascinating research area. Amalia, thank you for all your guidance and knowledge, your eye for detail has been especially appreciated! Finally, Andrew, I am in awe of your expertise and want to thank you for sharing it with me! I could not have asked for a better supervisory team thank you all so much.

I am also very grateful to the co-authors of my publications: Associate Professor Monica Taljaard, Professor Jeremy Grimshaw, Professor Allen Cheng, Professor Lisa Bero and Elizabeth Korevaar. Not only did they provide excellent feedback, advice and direction during the project and publication drafts, they were also very encouraging. I also want to thank Steve McDonald for his invaluable search strategy expertise.

This research was supported by an Australian Government Research Training Program Scholarship. I was also supported by my colleagues, the cheer squad from Cochrane Australia, the Methods for Evidence Synthesis team, SPHPM Graduate Research and my fellow PhD students! Thank you Cathie Smith, Kathryn Daly, Kelsey Grantham, Kim Jachno, Matt Page, Miranda Cumpston, Sam Brilleman, Professor Sally Green, Shauna Hurley and Sue Brennan.

When I tore my brain away from my research, I found many friends willing and able to lend me emotional sustenance and encouragement. I want to particularly acknowledge Clayton for the many games of Destiny, Mandy for the many cups of tea, PJ for all those miniature battles, Dave and Dean for the computer gaming and Amy for rocking with us in the Disco Den.

Mum and Dad, thank you for always encouraging me to follow where my heart, faith and brain directed me. You taught me that God wants us to make the world a better place, and that we could all go about that in different ways.

My penultimate thanks go to my son Oscar. You were six years old when I started, and now you are nine! You have been so patient all those times I was too busy to spend time with you. You are an amazing kid and I hope that in the future you will achieve something really big that I will get to help you through in the way that you have helped me.

Finally, I want to thank my wonderful wife Tari. You have been outstanding, always there for me, ready to listen, read, congratulate, “oooh” at graphs and animations, and especially always ready with a hug when required. I love starting every day knowing you are on my side and finishing every day with you right there with me. This has totally been a team effort and would never have happened at all without your partnership. I love you so very much. Thank you most of all.

V. List of abbreviations

Abbreviation	Definition
ARIMA	AutoRegressive Integrated Moving Average
CI	Confidence Interval
CO	Cochrane-Orcutt
CRT	Cluster Randomised Trial
d.f.	Degrees of freedom
DW	Durbin-Watson
GEE	Generalised Estimating Equation
GLMM	Generalised Linear Mixed Model
GLS	Generalised Least Squares
HAI	Healthcare-Associated Infection
ICC	Intra-Correlation Coefficient
IQR	Inter-Quartile Range
ITS	Interrupted Time Series
LOESS	LOcal regESSion
OLS	Ordinary Least Squares
MeSH	Medical Subject Headings
MCSE	Monte-Carlo Standard Error
ML	Maximum Likelihood
NHMRC	National Health and Medical Research Council
NW	Newey-West
PW	Prais-Winsten
RCT	Randomised Controlled Trial
REML	REstricted Maximum Likelihood
RMSE	Root Mean Square Error
Satt	Sattherthwaite
SD	Standard Deviation
SE	Standard Error

Chapter 1. Introduction

1.1 Introduction

Randomised controlled trials (RCTs) are the gold standard study design for investigating the impact of an intervention (1-4). When RCTs are well designed and conducted, they minimise bias, and as such, any observed differences in the outcome between groups can be more confidently ascribed to the intervention. For this reason, RCTs have a pivotal role in informing clinical decision making through their inclusion in systematic reviews and clinical practice guidelines. However, RCTs cannot always be used to study public health questions, such as those examining policy or programmes targeted at the population level, or the impact of exposures such as natural disasters or pandemics (1-8).

The interrupted time series (ITS) design provides an alternative study design for assessing the impact of interventions and exposures and is considered one of the strongest non-randomised experimental designs (1, 3, 5, 8-10). In an ITS study, data are collected at multiple time points before and after an interruption (intervention or exposure). Data from the pre-interruption interval can be used to create a counterfactual outcome for what would have occurred in the absence of the interruption, analogous to the control group in an RCT.

The aim of the research presented in this thesis was to assess the design, reporting quality and statistical methods used in ITS studies, and to subsequently provide tools and guidance that may facilitate improvement. The outputs of this thesis include a review of ITS studies, examining the design characteristics, statistical models, estimation methods and quality of reporting; recommendations for graphing ITS data; and recommendations for statistical methods to analyse ITS. It is hoped that that these outputs may facilitate improvement in the design, conduct and reporting of ITS studies.

This first chapter contextualises the research presented in this thesis by giving an overview of ITS study design, describing key considerations in the design and implementation of these studies; summarising previous reviews of the design, analysis and reporting of ITS studies; and providing an overview of statistical methods commonly used to analyse ITS data. The chapter concludes with the aims and objectives of the research and presents an outline of the thesis.

1.2 Overview of the interrupted time series study design and analysis

In an ITS study, data for an outcome are collected over a period of time. The data may be collected at different levels (e.g. individual, organisation, country), continually or at specific points in time, and be of different data types (e.g. binary, count). Often the data are aggregated using a summary statistic for the purpose of analysis (e.g. proportion of individuals that have been vaccinated, infection rate per 1000 patient days) within a time interval (e.g. weekly, monthly, annually). The time series is interrupted at a defined time by an intervention or exposure, separating the series into pre-interruption and post-interruption segments. The interruption may be intended (e.g. an anti-smoking mass media intervention), or unintended (e.g. an exposure such as a global pandemic).

An ITS design is often adopted when it is not possible to use randomisation, which can arise for many reasons. Ethical considerations may preclude randomisation. For example, when investigating the effects of introducing smoking bans in public locations it would be unethical to randomise a group to a location with known health risks (11). Some interruptions may preclude separating a population into two groups because being exposed to the intervention or not can never occur contemporaneously. Examples of this include government policies that impact an entire population (12), an interruption that occurred historically (e.g., a retrospective evaluation in which the researcher has no control over the target population group or timing (13)), a natural event such as a global financial crisis (14) or a natural disaster like an earthquake (15) or pandemic (16). Furthermore, it may be economically or practically infeasible to use an RCT, for example, when examining the effects of an interruption on a rare outcome requiring populations on a state or national scale in order to have a sufficient sample size (14).

The impact of the interruption can be investigated by the comparison of statistics between the pre- and post-interruption segments. A simple but naïve analysis involves comparing the mean levels of the data points before the interruption to those after the interruption, which is known as a before-after or pre-post analysis. However, this type of analysis does not account for any trends over time in the data and, in the presence of a trend, may over- or under-estimate the impact of the interruption. An ITS analysis allows for the pre-interruption trend to be accounted for in estimating the impact of the interruption. When the pre-interruption trend is modelled correctly and extrapolated into the post-interruption period, a “counterfactual” (an estimate of what would have happened in the absence of the interruption) can be established. The impact of the interruption can then be investigated by the comparison of statistics between the counterfactual trend and the observed post-interruption trend.

Statistical methods can be used to estimate a range of effect measures that characterise the impact of the interruption. Two commonly used effect measures are the immediate level change (representing the change immediately following the interruption); and change in slope (representing the difference in trends pre- and post-interruption) (17) (Figure 1). Segmented regression methods are commonly used in the analysis of ITS designs and are the focus of this thesis (9, 17-19).

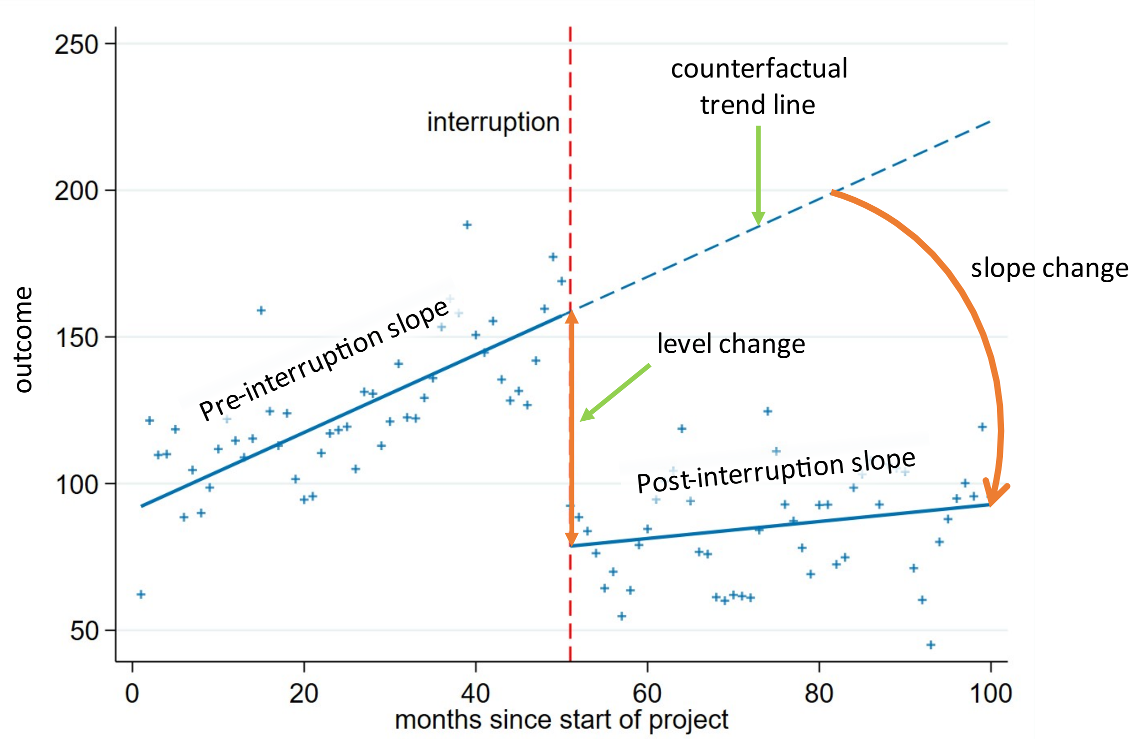


Figure 1: A two-period interrupted time series design with a segmented linear regression model. Trends (solid blue lines) based on the observed data (blue crosses) can be estimated in the segments before and after the interruption (vertical dashed red line). A counterfactual (extrapolation of the pre-interruption trend line shown as a dashed blue line) enables different effect measures (e.g. changes in level or slope) between the expected and observed to be estimated over a range of times.

1.3 Complexities of interrupted time series designs and analyses

There are a range of issues that need to be considered in the design and analysis of ITS studies. Key design issues include consideration of the potential impact of events other than the interruption being investigated and the length of the series. Key analysis issues include the need for valid estimation of the counterfactual and consideration of the potential for correlation between the data points.

Co-interruptions, extraneous events or changes to the environment that occur around the time of the interruption may bias the estimated effects from an ITS study. One way to guard against these factors is to include one or more control time series in the analysis (20, 21). If a large effect (e.g. level change) is observed in the interruption series but not in the control series, this provides more certainty that the interruption has caused the impact. If the effect is seen in both series, this suggests that another factor may be responsible. A range of types of controls are available including: location (e.g. a different area where the interruption is absent), outcome (e.g. an outcome not affected by the interruption), behaviour (e.g. a group of individuals who never performed the behaviour being investigated), characteristic (e.g. a group not targeted by an interruption designed to target a group holding a certain characteristic) and, historical (e.g. comparing a previous age group to a current age group) (20). While uncontrolled ITS are the focus of this thesis, the frequency with which controls are used in ITS evaluating public health interruptions is investigated.

The length of the time series is an important design consideration because it will impact how precisely the trends, and therefore effects, are estimated (22) (Figure 2). Various recommendations for the length of time series have been proposed, including a minimum of 8 in each segment (8), 9 in each segment (9),12 in each segment (10, 23) and a general acknowledgement that longer series are better (3, 23, 24). A focus of this thesis is to examine the impact of series length on the accuracy of effect.

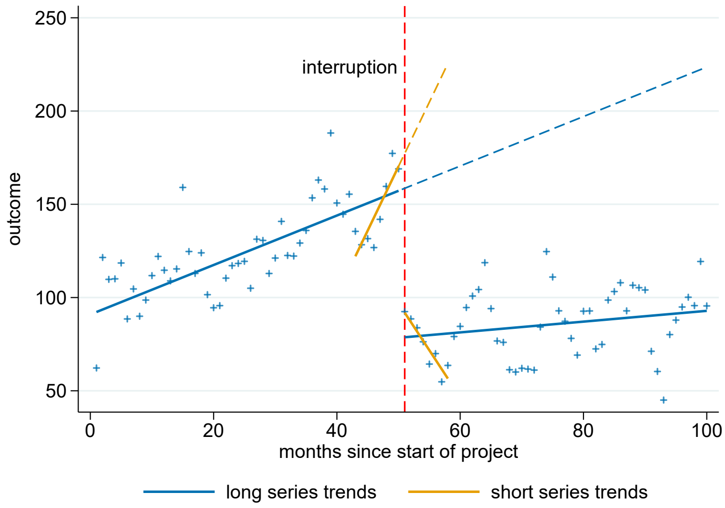


Figure 2: Interrupted time series graph showing an example of the difference in trend lines when estimated using a short data series (orange) versus a longer data series (blue).

The validity of effect estimates is dependent on the validity of the counterfactual. The key assumption is that in the absence of any interruption, the pre-interruption trend continues without deviation into the post-interruption period (1, 3). This may be not always be a sensible assumption, particularly in situations when trends cannot assume to hold over indefinitely long periods of time (1, 10). For example, when assessing the impact of social distancing measures on the increase in pandemic cases it cannot not be assumed that the number of infections would increase indefinitely in the absence of an interruption (16).

A feature of data collected over time is that the data points tend to be correlated. This is known as autocorrelation or serial correlation (25). This association could be positive (whereby data points close together in time are more similar than data points further apart) or, rarely, negative (whereby data points close together are more dissimilar than data points further apart). If positive autocorrelation is present and not accounted for, this may lead to standard errors (SEs) that are too small, with resulting confidence intervals (CIs) that are too narrow and p-values that are too small (26). A specific type of autocorrelation that may be observed in ITS designs with data collected over a long period of time is seasonality, which refers to periodic, repetitive, and predictable patterns in the levels of the time series. Influenza rates, for example, may demonstrate patterns of higher levels in winter and lower levels in summer months, recurring every calendar year (Figure 3). The time gap between autocorrelated data points is often denoted by lag-x (e.g. lag-1 refers to correlation with the previous time point, lag-12 could be used for annual, 12-monthly correlation). The magnitude of autocorrelation is a key parameter in sample size calculations; in the presence of autocorrelation, a larger sample size is required to achieve a given power (27). Autocorrelation may be handled in different ways in the analysis. For example, by inclusion of covariates in the model to try and explain the autocorrelation (e.g. time-varying predictors, terms for seasonality), adjustment of the standard errors to account for autocorrelation, or by directly modelling the error term. The latter two analysis approaches are the focus of the research presented in this thesis.

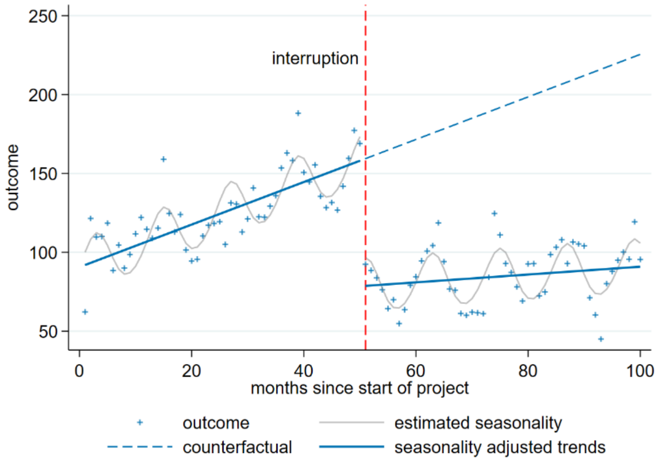


Figure 3: Interrupted time series graph displaying seasonality.

1.4 Overview of reviews examining design, analysis and reporting characteristics of ITS studies

Five reviews have been undertaken examining the design characteristics, statistical methods and reporting of ITS studies. Two of these were published at the time of undertaking this thesis (9, 24). The five reviews differ in the types of interruptions investigated and the years in which the included ITS studies were published. Ramsay et al (24) included ITS studies published from 1976 to 1998 from two systematic reviews (one examining the impact of a mass media campaign, one examining the dissemination of clinical guidelines and implementation strategies). Jandoc et al (9) included 220 ITS studies investigating healthcare interventions on drug utilisation (published between 1994 and 2013). Hudson et al (17) included 116 ITS studies published in 2015 that assessed the impact of healthcare interventions. Ewusie et al (19) performed a review of 1365 studies that used an ITS design or analysis for assessing a health-related intervention that had been published prior to September 2017. Hategeka et al (18) included 120 ITS studies examining the effects of health system quality improvement interventions that had been published prior to June 2018.

The design characteristics examined in the reviews included use of control groups and series length (Table 1). Approximately one-fifth to one-third of studies included a comparison or control series (9, 18, 19), though no details of the type of control group were reported in the reviews. Series lengths tended to be short, with the median number of points less than 20 per segment (9, 17-19, 24). Details of multiple segment studies (i.e. those with more than two segments), which may arise, for example, from multiple interruptions or the inclusion of an additional segment to allow time for the interruption to take effect, were rarely examined in these reviews.

Table 1: Study characteristics identified by ITS reviews. Abbreviations IQR: inter-quartile range, CI: confidence interval, SE: standard error.

Design Characteristics			
Control series	Included a control series		
Jandoc et al (9)	35% (77/220)		
Ewusie et al (19)	17% (237/1365)		
Hategeka et al (18)	18% (22/120)		
Series lengths	Number of data points (pre-interruption median)	Number of data points (post-interruption median)	Number of data points (other)
Ramsay et al (24)	9	6	
Ramsay et al (24)	10	12	
Jandoc et al (9)			Total series length range: 3 to 72
Hudson et al (17)	18 (IQR 12-32)	19 (IQR 12-34)	
Ewusie et al (19)			10% (141/1365) of studies included fewer than 16 points
Hategeka et al (18)	18 (range 3-120)	20 (range 4-90)	
Statistical Characteristics			
Segmented regression	Analyses used segmented regression		
Jandoc et al (9)	67% (134/220)		
Hudson et al (17)	78% (90/116)		
Ewusie et al (19)	65% (889/1365)		
Hategeka et al (18)	63% (75/120)		
Unspecified method	Analysis method was unspecified		
Jandoc et al (9)	43% (58/134)		
Ewusie et al (19)	20% (267/1365)		
Hategeka et al (18)	12% (14/120)		
Autocorrelation	Considered autocorrelation		
Jandoc et al (9)	33% (74/220)		
Hudson et al (17)	55% (63/116)		
Ewusie et al (19)	60% (812/1365)		
Hategeka et al (18)	55% (66/120)		
Reporting Characteristics			
Measure of precision	Reported a CI or SE with effect estimates		
Jandoc et al (9)	70% (153/220)		
Hudson et al (17)	76% (74/97)		
Hategeka et al (18)	80% (60/75)		
Graph	Included a graph		
Jandoc et al (9)	84% (184/220)		
Hudson et al (17)	95% (109/116)		
Ewusie et al (19)	89% (1218/1365)		
Hategeka et al (18)	93% (111/120)		

Statistical methods were examined in all reviews. Ramsay et al (24) concluded that the ITS studies included in their review were often incorrectly analysed. When re-analysed using appropriate methods, almost half of the studies that had reported statistically significant effect estimates were shown to have no statistically significant differences in slope or level. Segmented linear regression was the most common statistical method; across the reviews, its use ranged from 63% to 78% (9, 17-19, 24) (Table 1). In three reviews, it was noted that in many studies, the statistical method was not specified (ranging from 12% to 43%) (9, 18, 19). Autocorrelation was commonly not considered (9, 17-19). However, although these reviews made some assessment of how autocorrelation was handled in the studies, they (except Hudson et al (17)) did not examine how autocorrelation was accounted for (e.g. including an autoregressive error term, using an appropriate statistical method) and none assessed whether estimates of autocorrelation were reported.

Aspects of reporting that were consistently examined across the reviews were completeness of reporting of effect estimates, and whether graphical displays were used (Table 1). Measures of precision, such as CIs or SEs were not always reported with the effect measure (included in 70% to 80% of studies) (9, 17, 18). Graphical displays were frequently used, with the percentage of studies including a graphical display ranging from 84% to 95% across the reviews (9, 17-19). None of the reviews examined the characteristics of the graphs included in the studies.

A common theme of these reviews was that the reporting of ITS studies was generally incomplete, with limited information provided about details of the statistical methods and considerations such as autocorrelation. This thesis extended this existing body of knowledge by investigating design characteristics, statistical methods and completeness of reporting of ITS studies of public health interventions and exposures.

1.5 Graphing ITS data

The ITS design inherently lends itself to a visual display, and when well presented, allows readers to easily assess the impacts of the interruption on the outcome (3, 8, 10, 18). Recognition of the usefulness of graphing ITS data is borne out by the majority of ITS study publications presenting a graph (9, 17-19). A further benefit of visually displaying data from an ITS study is that it allows systematic reviewers to extract the data (e.g. using digitising software) and undertake a re-analysis (9, 28). This is particularly important for reviewers wishing to perform a meta-analysis, where consistency in effect measures across studies and effect estimates with SEs are required, but not always provided (9, 17, 18, 28). While general visualisation guidance exists (29-36), no specific guidance was available for how to graph ITS data to best facilitate interpretation and data extraction; this forms a part of the research presented in this thesis.

1.6 Statistical methods used to analyse ITS studies

Many statistical methods can be employed in the analysis of ITS studies. In this thesis, statistical methods that are commonly used (ordinary least squares (OLS), generalised least squares (GLS) and autoregressive integrated moving average (ARIMA)) (9, 17-19), as well as a method which has been suggested for its potential to reduce bias in estimating the magnitude of autocorrelation (restricted maximum likelihood (REML) (37)), were examined. These methods are now described in further detail.

Using OLS regression to fit a segmented linear regression model yields unbiased estimates of the regression parameters (2). However, this method does not account for autocorrelation and so, in the presence of (likely) positive autocorrelation, the SEs of the effect estimate are underestimated (38). The Newey-West (NW) estimator of the variance of the regression parameters (39), estimated using OLS, is used to try and account for autocorrelation (for any number of lag times). However, it has been shown to still underestimate the SEs in time series (40). The extent to which this underestimation occurs in estimating regression parameters in ITS studies is currently unknown.

Other regression methods are available that account for autocorrelation, including GLS methods such as Prais-Winsten (PW) and Cochrane-Orcutt (CO) (2, 41). For both methods, the residuals from an OLS regression model are used to estimate the lag-1 autocorrelation. This estimate is then used to transform the data and remove the autocorrelation from the errors, upon which the regression parameters are then estimated from the transformed data. The CO method applies the transformation from the second observation onwards (t = 2, 3, …, n). The PW method is a modification to the CO method in which a transformed value is used for the first observation. The PW method is therefore likely to be more efficient in small series since it does not discard the first observation. The sampling properties of the estimators of the regression parameters are likely to be adversely affected when the series length is small due to poor estimation of the autocorrelation. The performance of GLS methods when analysing ITS studies, which are often short, has not been previously investigated.

In an ARIMA model, information from past values, including lagged values of the dependent variable and errors, are explicitly modelled (42, 43). This is achieved by including regression coefficients for these variables in the ARIMA model. The lagged values can be from a range of previous time points, extending beyond lag-1 models. By explicitly modelling the influence of data from previous time points, their impact at subsequent times is quantified and estimates of the magnitude of autocorrelation can be obtained along with regression parameter estimates. ARIMA models are frequently recommended for time series data and for ITS studies (5, 44). The performance of ARIMA methods compared to other statistical methods when analysing ITS studies has not previously been investigated.

It is well known that maximum likelihood estimators of variance components are biased in small samples because they do not account for the degrees of freedom (d.f.) used when estimating the fixed effect regression parameters (45). Restricted maximum likelihood (REML) is a variant of maximum likelihood (ML) estimation and attempts to address the bias by separating the log-likelihood into two terms; one that involves the mean and variance parameters, and one which is only dependent on the variance parameters. By maximising the latter term first with the appropriate number of d.f., an estimate of the variance parameter can be obtained which can be used when maximising the former, thus correctly accounting for the d.f. (37, 46). Both ML and REML methods can estimate and account for autocorrelation extending beyond lag-1 models. Although REML methods have been investigated for time series data, it is unknown how they compare with other methods in ITS studies.

There are few resources available to guide researchers analysing ITS studies as to the choice of statistical method, and which method may be preferable for series with particular characteristics (e.g. short ITS). The Cochrane Effective Practice and Organisation of Care (EPOC) guide recommends using a regression analysis with adjustment for autocorrelation or an ARIMA method (47). Penfold and Zhang (8) and Wagner et al (10) advise that a method that can control for autocorrelation should be used, and suggest the PROC AUTOREG procedure in the SAS program, which implements GLS methods (48). If the outcome is count data, Lopez-Bernal et al (3) and Gebski et al (25) suggest using Poisson regression, though if autocorrelation is present Lopez-Bernal add that PW or ARIMA models should be used, and Gebski et al (25) note that in this circumstance different modelling may be required. While these resources suggest a variety of methods for the analysis of ITS studies, there have been no statistical simulation studies that have compared the performance of different statistical estimation methods for ITS studies, therefore, this was an avenue of research explored in this thesis.

1.7 Sample size methods for interrupted time series studies

There has been little investigation or development of sample size methods for ITS. McLeod and Vingilis (49) derived a sample size formula to detect a level change (difference in means between the pre- and post-interruption periods) using ARIMA models and either an autoregressive term of 1 or an integrated moving average term of 1. They noted that the sample size was strongly dependent on the magnitude of autocorrelation. Zhang et al (50) conducted a simulation study in which time series were generated based on modifications of existing SAS algorithms, including autocorrelation (51). They estimated the power to detect various target differences in standardised level and slope changes for different sample sizes, under a range of scenarios. They investigated the impact of autocorrelations ranging from -0.9 to 0.9, and the impact of an unequal series length between periods. They found that power increased with increasing sample size or effect size, but decreased as autocorrelation increased. They also noted that unequal series length between periods had a detrimental effect on power in some situations. Liu et al (27) conducted a simulation study for count outcomes in which they estimated the power for various sample sizes, using Poisson and negative binomial methods including both level and slope changes. They examined autocorrelation ranging from -0.9 to 0.9, and concluded that power was affected by autocorrelation, decreasing as autocorrelation increased.

Estimates of effect size and autocorrelation are key parameters for sample size calculation of ITS series, yet there is little guidance available to inform the typical size of these effects that might be seen in ITS of public health interruptions; this forms a part of the research presented in this thesis.

1.8 Research aims and objectives

The design characteristics, models and statistical methods being used to analyse ITS with public health impacts are currently unknown. There is limited information available to inform the design of ITS studies, with key parameters required for sample size calculation, such as the likely magnitude of autocorrelation, generally being unavailable. ITS studies are particularly amenable to visual display; however, no guidelines are available to assist researchers in the creation of ITS graphs that accurately depict the data. There are also few studies evaluating and comparing the performance of different statistical methods for analysing ITS data, and how their performance is influenced by characteristics of the time series (e.g. magnitude of autocorrelation, series length). Further guidance for how to best design, analyse and report the results of ITS studies is particularly important in public health, where other study designs are often not feasible.

Therefore, the aims of this thesis were to assess the design, reporting and statistical methods used in recent ITS studies that have evaluated the impact of interventions or exposures on public health outcomes; examine the properties of graphs presented in ITS studies and provide recommendations for visualising ITS data; evaluate the performance of a range of statistical methods that can be used to analyse ITS data; and provide tools and guidance for the design, conduct and reporting of ITS studies. The specific objectives include:

- Objective 1: Assess the design, statistical methods, and reporting used in recent ITS studies that evaluate the impact of interruptions on public health-related outcomes by reviewing the:
    - study and design characteristics;
    - types of health-related outcomes being investigated;
    - models used;
    - statistical methods employed;
    - effect measures reported; and,
    - graphs included.

- Objective 2: Examine the properties of graphs presented in ITS studies and provide recommendations for graphing ITS data by:
    - proposing recommendations for graphing ITS data adapted from graphing recommendations from the seminal data visualisation literature;
    - assessing whether graphs from recently published ITS studies met these recommendations;
    - demonstrating the use of the recommendations by applying them in two examples; and,
    - providing computer code to implement the recommendations.
- Objective 3: Investigate the performance of a set of statistical methods used in analysing ITS data by:
    - simulating continuous outcome data under a range of realistic scenarios which included a single interruption at the mid-point of the series, different level and slope changes, varying lengths of series, constant variance and, varying magnitudes of lag-1 autocorrelation.
    - assessing the performance of statistical methods using a range of criteria (e.g. bias, empirical SE, model based SE, 95% CI coverage and power);
    - assessing autocorrelation estimates from the statistical methods as well as the commonly used Durbin-Watson (DW) test for detecting autocorrelation.
    - investigating the impact of scenario parameters (level changes, slope changes, series length and autocorrelation) on statistical method performance.
- Objective 4: Investigate how the results of a set of statistical methods compare in practice by:
    - fitting segmented linear regression models with a continuous outcome, a single interruption, and allowing for lag-1 autocorrelation, to each dataset of a large sample of real-world ITS studies;
    - applying each of the statistical methods to each of the datasets;
    - comparing level change and slope change estimates, SEs, CIs and p-values; and,
    - comparing estimates of autocorrelation.
- Objective 5: Create a large repository of ITS data by collating published and digitally extracted time series data used in ITS studies.

1.9 Outline of the thesis

This thesis includes three published manuscripts and two manuscripts that have been submitted for publication.

- Chapter 2 details a protocol for a review of ITS studies. The review aimed to examine the design characteristics, statistical methods and completeness of reporting in a random sample of recent ITS studies examining interruptions with public health impacts. This protocol has been published in *BMJ Open* (52).
- Chapter 3 presents the results of the review of 200 ITS studies published between 2013 and 2017, examining characteristics of the study, design, outcome, model, statistical methods and reported effect measures and their components. The review has been published in the *Journal of Clinical Epidemiology* (53).
- Chapter 4 examines the properties of plots presented in ITS studies and provides recommendations for graphing ITS data along with two examples and computer code to aid implementation of the recommendations. This study has been published in *Research Synthesis Methods* (54).
- Chapter 5 describes the methods and results of a statistical simulation study comparing a set of statistical methods used to analyse ITS studies. A manuscript has been submitted to the *Biometrical Journal* and is available as a pre-print (55).
- Chapter 6 describes the methods and presents the results of an empirical evaluation that compares the results from a set of statistical methods applied to a large sample of real-world ITS studies. A manuscript has been submitted to *BMC Medical Research Methodology*.
- Chapter 7 presents a summary of the findings and suggestions for further research.

Chapter 2. Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: protocol for a review

This chapter presents a protocol for a review of the design characteristics and statistical methods used in a random sample of contemporary ITS studies examining public health interventions or exposures that impact on health-related outcomes. This review was necessary to understand how the ITS design is being used in current practice and to inform the subsequent chapters of this thesis.

The aim of the review was to examine the design characteristics, statistical methods and completeness of reporting in a random sample of ITS studies examining interruptions with public health impacts.

The protocol presented in this chapter provides a description of the review methods. This includes details of the eligibility criteria, the search strategy, study and outcome selection processes, methods for data extraction and management, as well as data extraction items.

Chapter 2 is presented as a manuscript, which was published in *BMJ Open* in October, 2018 (52). The complete list of data extraction items developed for this review, and referred to as `Supplementary Additional file 1' in the manuscript, is appended to this thesis in Appendix A.

Open access Protocol

BMJ Open

Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: protocol for a review

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

To cite: 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**:e024096. doi:10.1136/bmjopen-2018-024096

- Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-024096).

Received 11 May 2018
Revised 5 September 2018
Accepted 18 October 2018

For numbered affiliations see end of article.

Correspondence to
Associate Professor Joanne E McKenzie;
joanne.mckenzie@monash.edu

ABSTRACT

Introduction An interrupted time series (ITS) design is an important observational design used to examine the effects of an intervention or exposure. This design has particular utility in public health where it may be impracticable or infeasible to use a randomised trial to evaluate health system-wide policies, or examine the impact of exposures (such as earthquakes). There have been relatively few studies examining the design characteristics and statistical methods used to analyse ITS designs. Further, there is a lack of guidance to inform the design and analysis of ITS studies. This is the first study in a larger project that aims to provide tools and guidance for researchers in the design and analysis of ITS studies. The objectives of this study are to (1) examine and report the design characteristics and statistical methods used in a random sample of contemporary ITS studies examining public health interventions or exposures that impact on health-related outcomes, and (2) create a repository of time series data extracted from ITS studies. Results from this study will inform the remainder of the project which will investigate the performance of a range of commonly used statistical methods, and create a repository of input parameters required for sample size calculation.

Methods and analysis We will collate 200 ITS studies evaluating public health interventions or the impact of exposures. ITS studies will be identified from a search of the bibliometric database PubMed between the years 2013 and 2017, combined with stratified random sampling. From eligible studies, we will extract study characteristics, details of the statistical models and estimation methods, effect metrics and parameter estimates. Further, we will extract the time series data when available. We will use systematic review methods in the screening, application of inclusion and exclusion criteria, and extraction of data. Descriptive statistics will be used to summarise the data.

Ethics and dissemination Ethics approval is not required since information will only be extracted from published studies. Dissemination of the results will be through peer-reviewed publications and presentations at conferences. A repository of data extracted from the published ITS studies will be made publicly available.

Strengths and limitations of this study

- To our knowledge, this will be the first study specifically examining the use of the interrupted time series (ITS) design in a representative sample of studies in public health.
- A priori systematic review methods will be used in the screening, application of inclusion and exclusion criteria, and data extraction.
- A wide range of items capturing the design characteristics, statistical methods and parameter estimates will be extracted and summarised, and a repository of time series data will be created.
- For some items, the sample size may not be large enough to precisely estimate the percentage of ITS studies with a particular element.
- Our search strategy is unlikely to locate all published ITS studies in public health, since studies will use terminology other than our search terms. Further, we will only search a single database (ie, PubMed); however, this database has the broadest coverage of public health and health services research journals.

INTRODUCTION

Background

An interrupted time series (ITS) design is one in which data are measured at multiple time points before and after the introduction of an intervention (or an exposure) to examine the effect of the intervention (or exposure). This 'quasi-experimental' design is superior to many other observational study designs such as before and after designs in that it avoids threats to internal validity such as short-term fluctuations, secular trends and regression to the mean. ITS designs are used to examine the effects of public health system-wide policy interventions (eg, mass media campaigns[1 2]) where it is impracticable or infeasible to use a randomised trial. In addition, they can be used to evaluate the effects of policies/interventions retrospectively using administrative databases.[3 4] Or,

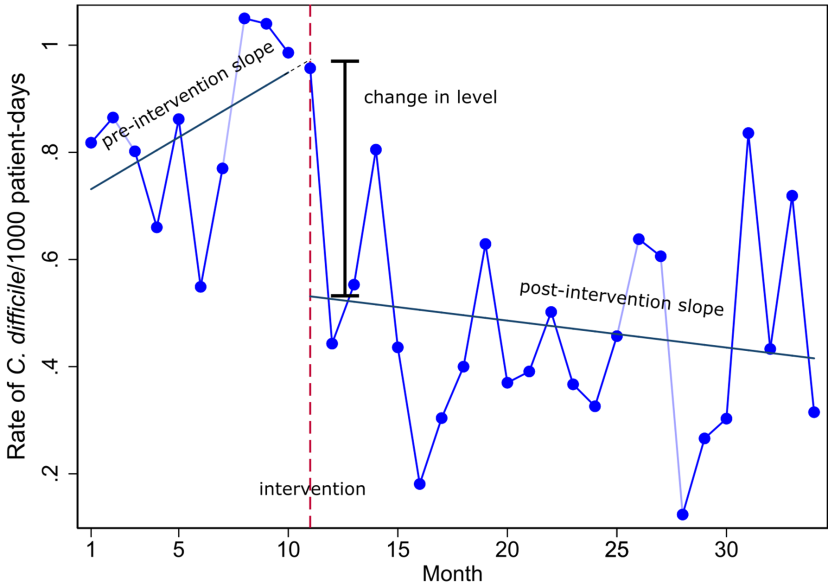


Figure 1 The rate of *Clostridium difficile* infections (per 1000 patient days) prebleach and postbleach disinfection per month.[28] Various effect estimates can be constructed from the preintervention and postintervention slopes, such as the change in level and change in slopes.

they can be used to examine the impact of exposures such as earthquakes (eg,[5 6]) or nuclear power station leaks (eg,[7]). An important benefit of an ITS design is that it can account for the preintervention trend in estimating the effect of the intervention.[8] Figure 1 provides a graphical representation of a simple ITS design, with data collected in preintervention and postintervention phases.

A feature of data collected over time is that the data points tend to be correlated. This is known as autocorrelation or serial correlation.[9 10] This association could be positive (whereby data points close together in time are more similar than data points further apart) or negative (whereby data points close together are more dissimilar than data points further apart). A specific type of autocorrelation that may be observed in ITS designs with data collected over a long period of time is seasonality, which refers to periodic, repetitive and predictable patterns in the levels of the time series. Influenza rates, for example, may demonstrate patterns of higher levels in winter and lower levels in summer months, recurring every calendar year. Some methods that are used to analyse ITS studies, such as classical linear regression, assume that the observations are independent. If positive autocorrelation is present and not accounted for, this may lead to standard errors that are too small, with resulting confidence intervals that are too narrow and p values that are too small.[10] Autocorrelation is also one of the key parameters required for sample size estimation.[11]

In addition to accounting for autocorrelation, other important aspects of analysing ITS data include specification of the structure of the model (ie, number of segments and their shape[12]), statistical estimation methods and choice of effect metrics. Regarding the latter, even in the circumstance where a simple segmented linear model is fitted (figure 1), a variety of effect metrics may be calculated. The effect of the intervention on the outcome can be calculated by estimating the change in level, or the change in slope of the preintervention and postintervention trends, or both (see figure 1). A combination of the two allows estimation of the intervention effect at a specified time post intervention based on the predicted counterfactual (ie, using the trend in the preintervention period to predict what would have occurred in the postintervention period, in the absence of the intervention). Further, these effect metrics can be expressed in absolute or relative terms with associated confidence intervals. For example, a level change corresponding to a drop in mortality could be expressed as an absolute effect such as number of deaths or as a relative effect such as percentage change in the number of deaths.[13] Choosing which metric is most helpful for understanding and communicating the impact of an intervention is not straightforward.[14]

There have been relatively few studies examining the design characteristics and statistical methods used to analyse ITS designs. Ramsay *et al*[15] examined ITS studies included in two systematic reviews and demonstrated that when reanalysed using more appropriate statistical methods, approximately half of the studies that had found statistically significant intervention effects, had their results overturned (ie, statistically significant results became non-significant). Jandoc *et al*[16] examined a cohort of 220 ITS studies, published from 1984 to 2013, in drug

utilisation research and found that important elements of an ITS design were not being accounted for in the analyses, such as taking into account autocorrelation and seasonality. Ewusie *et al*[17] are undertaking a scoping review of ITS methods used to analyse health research and the way in which the results are reported, focusing on the strengths and limitations of each method. None of these reviews has focused on ITS studies evaluating public health interventions, or exposures, that impact on health-related outcomes.

There is a lack of available information and guidance to inform the design and analysis of ITS studies. First, there are no databases providing empirical estimates of key parameters (such as autocorrelation coefficients) required for sample size calculation. Second, there is an absence of clear guidance to inform the choice of statistical analysis. Third, the choice of intervention effect estimates and their standard errors that are most robust to misspecification of the model and estimation method is unclear.

This is the first study in a larger project that aims to address these issues by providing tools and guidance for researchers on the design and analysis of ITS studies. The project includes a series of studies that will examine the design features and statistical methods used in practice, investigate the performance of a range of commonly used statistical methods through numerical simulation and empirical evaluation, and create a repository of input parameters required for sample size calculation. Here, we report the planned design of the first study, the objectives of which are to (1) examine and report the design characteristics and statistical methods used in a random sample of contemporary ITS studies examining public health interventions or exposures that impact on health-related outcomes, and (2) create a repository of time series data extracted from ITS studies.

METHODS AND ANALYSIS

Overview

We will identify and describe ITS studies evaluating public health interventions or the impact of exposures. ITS studies will be identified from a search of the bibliometric database PubMed, combined with stratified random sampling. Study selection and data extraction will be undertaken by one author, and for a fraction of studies, two authors. From eligible studies, we will extract study characteristics, details of the statistical models and estimation methods, effect metrics and parameter estimates. Further, we will extract the time series data, when available.

Eligibility criteria

Following are the inclusion and exclusion criteria applied to determine which ITS studies will be included.

Inclusion criteria

Our definition of an ITS study is informed by the design features taxonomy for quasi-experimental designs proposed by the Cochrane Non-Randomized Studies Methods Group[18] and further developed by Reeves *et al*.[19] ITS studies meeting the following criteria will be included (with rationale and further explanations for each criterion provided below):

1. There are at least two segments separated by a clearly defined intervention or exposure with at least three points in each segment (eg, preintervention and postintervention time series, each with at least three points).
2. Observations are collected on a group of individuals (eg, community, hospital) at each time point.
3. The study is investigating the impact of a public health intervention or exposure that has public health implications (eg, patient health outcome, resource use).

Criterion 1 is that used by the Cochrane Effective Practice and Organisation of Care Group as minimum criteria for an ITS study to be considered eligible for inclusion in systematic reviews undertaken within this group.[20] The rationale for this criterion is that three time points pre interruption and post interruption allow the possibility of using segmented time series regression (although using as few as three points would not be recommended[11]).

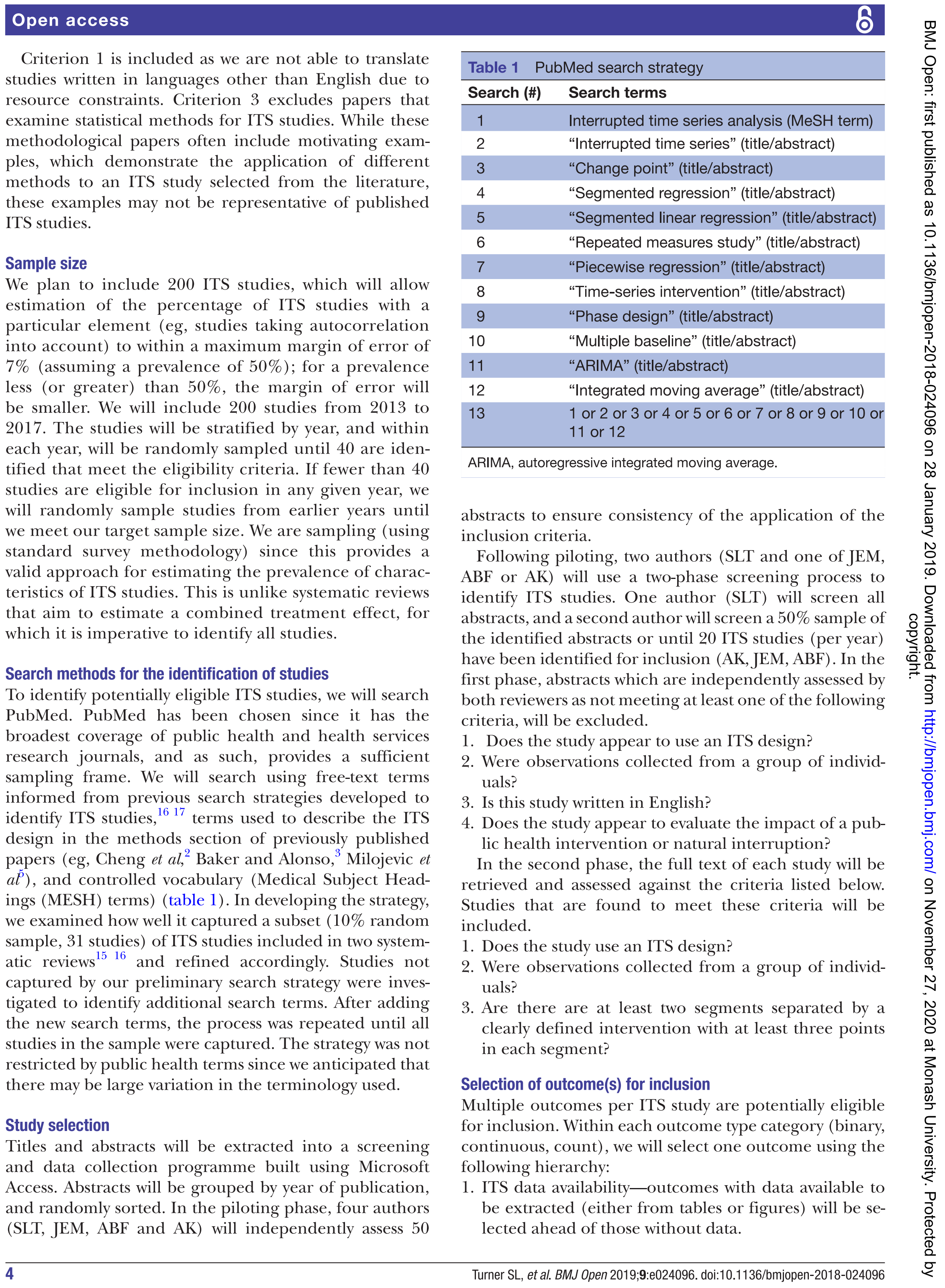
Criterion 2 restricts the inclusion to ITS studies focused on examining the effects of public health interventions (or exposures) on populations, and importantly, excludes ITS studies that examine the effects of an intervention on individuals (also known as single-case designs[21] or single-case experimental designs[22]). The group of individuals studied at each time point may or may not include the same individuals. The study may also include multiple series (multiple intervention and control series measured in aggregate or groups (eg, hospitals, communities)).

Our definition of public health interventions (criterion 3) is informed by that used by the Cochrane Public Health Review Group,[23] and includes interventions that aim to prevent or promote population health for communicable and non-communicable diseases (eg, vaccination and screening programmes; programmes aimed to reduce the use of tobacco or alcohol; public information/awareness campaigns for stroke recognition). The interventions may fall outside of the health service (such as education, work environment, housing and the built environment, natural environment interventions), but will be included if they aim to improve population health-related outcomes. Further, there will be no restriction on the level that the intervention is targeted at, which may be, for example, individuals, communities or health systems. Our definition of exposures are events that are not under investigators' control (eg, earthquakes, financial crises, tsunamis, environmental chemicals).

Exclusion criteria

ITS studies meeting the following criteria will be excluded:

1. Studies written in a language other than English.
2. Single-case designs.
3. Methodological papers examining ITS studies.



Criterion 1 is included as we are not able to translate studies written in languages other than English due to resource constraints. Criterion 3 excludes papers that examine statistical methods for ITS studies. While these methodological papers often include motivating examples, which demonstrate the application of different methods to an ITS study selected from the literature, these examples may not be representative of published ITS studies.

Sample size

We plan to include 200 ITS studies, which will allow estimation of the percentage of ITS studies with a particular element (eg, studies taking autocorrelation into account) to within a maximum margin of error of 7% (assuming a prevalence of 50%); for a prevalence less (or greater) than 50%, the margin of error will be smaller. We will include 200 studies from 2013 to 2017. The studies will be stratified by year, and within each year, will be randomly sampled until 40 are identified that meet the eligibility criteria. If fewer than 40 studies are eligible for inclusion in any given year, we will randomly sample studies from earlier years until we meet our target sample size. We are sampling (using standard survey methodology) since this provides a valid approach for estimating the prevalence of characteristics of ITS studies. This is unlike systematic reviews that aim to estimate a combined treatment effect, for which it is imperative to identify all studies.

Search methods for the identification of studies

To identify potentially eligible ITS studies, we will search PubMed. PubMed has been chosen since it has the broadest coverage of public health and health services research journals, and as such, provides a sufficient sampling frame. We will search using free-text terms informed from previous search strategies developed to identify ITS studies,[16 17] terms used to describe the ITS design in the methods section of previously published papers (eg, Cheng *et al*,[2] Baker and Alonso,[3] Milojevic *et al*[5]), and controlled vocabulary (Medical Subject Headings (MESH) terms) (table 1). In developing the strategy, we examined how well it captured a subset (10% random sample, 31 studies) of ITS studies included in two systematic reviews[15 16] and refined accordingly. Studies not captured by our preliminary search strategy were investigated to identify additional search terms. After adding the new search terms, the process was repeated until all studies in the sample were captured. The strategy was not restricted by public health terms since we anticipated that there may be large variation in the terminology used.

Table 1 PubMed search strategy

Search (#)	Search terms
1	Interrupted time series analysis (MeSH term)
2	“Interrupted time series” (title/abstract)
3	“Change point” (title/abstract)
4	“Segmented regression” (title/abstract)
5	“Segmented linear regression” (title/abstract)
6	“Repeated measures study” (title/abstract)
7	“Piecewise regression” (title/abstract)
8	“Time-series intervention” (title/abstract)
9	“Phase design” (title/abstract)
10	“Multiple baseline” (title/abstract)
11	“ARIMA” (title/abstract)
12	“Integrated moving average” (title/abstract)
13	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12

ARIMA, autoregressive integrated moving average.

Study selection

Titles and abstracts will be extracted into a screening and data collection programme built using Microsoft Access. Abstracts will be grouped by year of publication, and randomly sorted. In the piloting phase, four authors (SLT, JEM, ABF and AK) will independently assess 50 abstracts to ensure consistency of the application of the inclusion criteria.

Following piloting, two authors (SLT and one of JEM, ABF or AK) will use a two-phase screening process to identify ITS studies. One author (SLT) will screen all abstracts, and a second author will screen a 50% sample of the identified abstracts or until 20 ITS studies (per year) have been identified for inclusion (AK, JEM, ABF). In the first phase, abstracts which are independently assessed by both reviewers as not meeting at least one of the following criteria, will be excluded.

1. Does the study appear to use an ITS design?
2. Were observations collected from a group of individuals?
3. Is this study written in English?
4. Does the study appear to evaluate the impact of a public health intervention or natural interruption?

In the second phase, the full text of each study will be retrieved and assessed against the criteria listed below. Studies that are found to meet these criteria will be included.

1. Does the study use an ITS design?
2. Were observations collected from a group of individuals?
3. Are there are at least two segments separated by a clearly defined intervention with at least three points in each segment?

Selection of outcome(s) for inclusion

Multiple outcomes per ITS study are potentially eligible for inclusion. Within each outcome type category (binary, continuous, count), we will select one outcome using the following hierarchy:

1. ITS data availability—outcomes with data available to be extracted (either from tables or figures) will be selected ahead of those without data.

Table 2 Example data extraction items

	Examples of data extraction items
Study characteristics	Author name; year of publication; rationale for using an ITS design; type and description of the intervention.
Design	Time interval (eg, monthly); total number of observations; total number of time intervals and number of segments; number of time intervals per segment; average number of observations per time interval; and whether there is a comparison group.
Outcome	Description (eg, vehicle occupant injury) and classification (eg, count) of the outcome at the individual observation level; description of the aggregate level outcome (eg, rate per population of motor vehicle occupant injuries).
Model	Model shape (eg, level change or slope change, or both, and whether this shape is prespecified or not); number of segments; model type (eg, autoregressive integrated moving average (ARIMA), segmented regression, other regression, pre–post); modelling approach for any transition period; and, if there was a comparison group, how it was incorporated in the analysis.
Statistical methods	Statistical estimation method (eg, logistic, Poisson, overdispersed Poisson, generalised estimating equation (GEE); whether autocorrelation, seasonality and outliers were investigated; and, how they were handled in the analysis; whether and how non-stationarity was tested for.
Effect measures	Reported effect measures (eg, change in level, change in slope); whether an absolute or relative measure; effect estimates and statistics associated with the effect measure (eg, p values, CIs); details on any forecasting (eg, projecting from one segment to a specified time point in another segment) and whether there was mention of any ceiling or floor effects.

ITS, interrupted time series.

2. Stated primary outcome (or reported in the title or objectives).
3. First reported result outcome in the abstract.
4. First reported outcome in the results.

Uncertainty in the selection of the review outcomes will be discussed among the review team.

Data extraction and management

The data extraction process will initially be piloted by four reviewers (SLT, JEM, ABF and AK) on a sample of 10 ITS studies to ensure consistency of data extraction and will be adjusted as necessary. Following piloting, we will use double data extraction for all items (except extraction of the time series data) on a randomly selected 20% of studies. Discrepancies and uncertainty in coding will be discussed in meetings with three reviewers. For any items where we observe a high degree of inconsistency, we will undertake double data extraction for these items on a further randomly selected sample of studies.

We will extract data pertaining to the study characteristics, design, outcome, model, statistical methods and effect measures. Further details are provided in table 2, with the complete list of items available in online supplementary additional file 1.

The time series data for each included outcome will be extracted when possible. For studies that present their data in graphical form only, we will extract from graphs using WebPlotDigitizer software[24]. This software has been shown to give reliable and valid results.[25] The data will be kept in a Microsoft Access 2016 database[26] on a secure server. Any free text responses will be input as prespecified options or categories wherever possible.

Analysis

We will calculate descriptive summary statistics. For categorical data, such as the type of model used, we will present percentages and frequencies. For counts, such as the number of time intervals, we will present means (with standard deviations) and medians (with interquartile range). Statistical analyses will be undertaken in Stata Release 15[27]. The data sets arising from this study will be made available on figshare.

Patient and public involvement

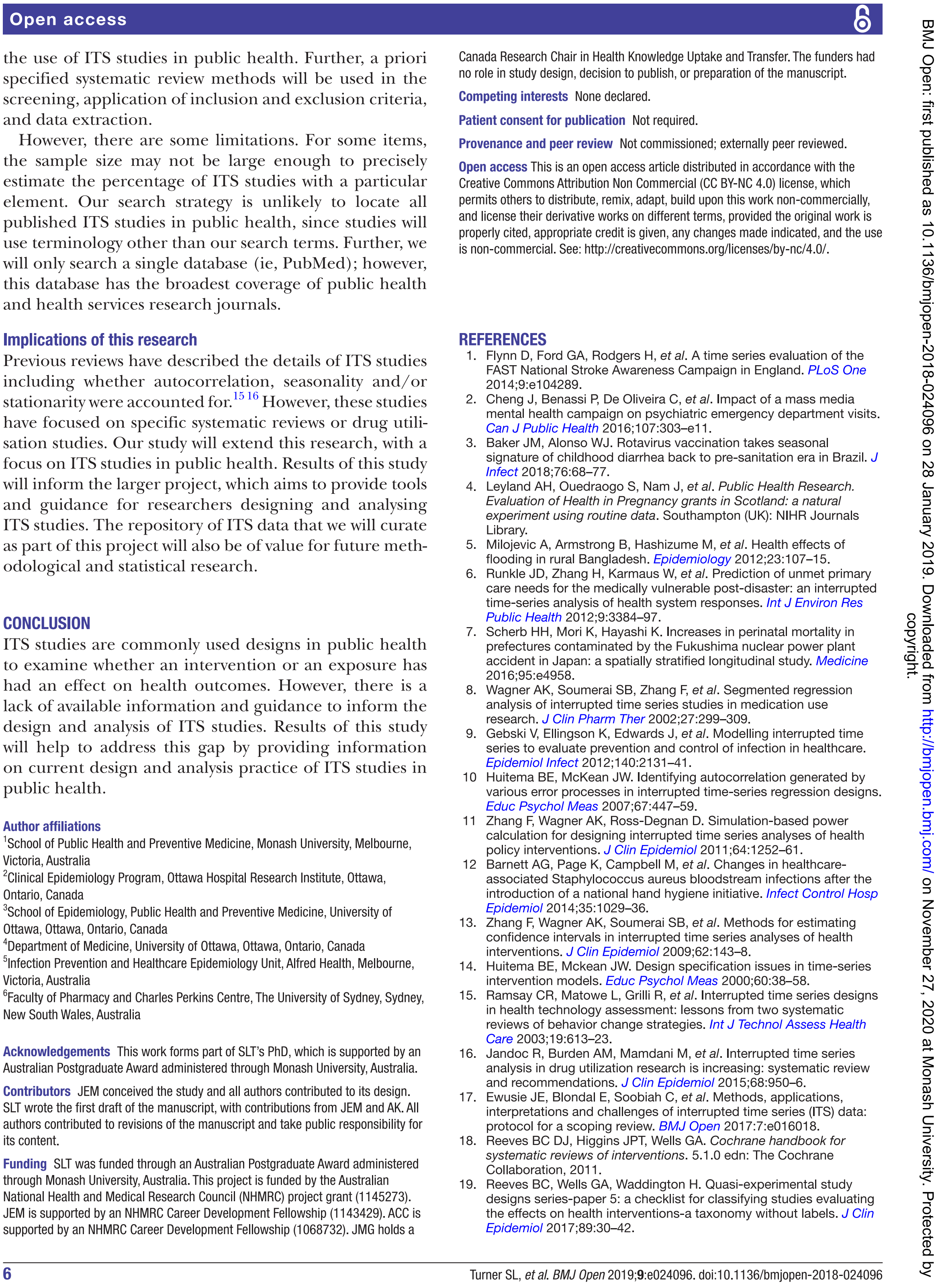
No patients will be involved in this project and information will only be extracted from published studies.

DISCUSSION

This is a first study in a larger project that aims to provide tools and guidance for researchers in the design and analysis of ITS studies. This study will provide information on design characteristics of a contemporary sample of ITS studies in public health, statistical methods used in practice, and provide a repository of ITS data. Results from this study will underpin the remainder of the project by informing a numerical simulation study to investigate the performance of commonly used statistical methods; an empirical study to investigate the impact of using different statistical methods for analysing ITS on real data; and a study to create a repository of parameter values for sample size calculation, along with generalisable 'rules of thumb' on the selection of values.

Strengths and limitations

There are several strengths to our study. To our knowledge, this will be the first study specifically examining



the use of ITS studies in public health. Further, a priori specified systematic review methods will be used in the screening, application of inclusion and exclusion criteria, and data extraction.

However, there are some limitations. For some items, the sample size may not be large enough to precisely estimate the percentage of ITS studies with a particular element. Our search strategy is unlikely to locate all published ITS studies in public health, since studies will use terminology other than our search terms. Further, we will only search a single database (ie, PubMed); however, this database has the broadest coverage of public health and health services research journals.

Implications of this research

Previous reviews have described the details of ITS studies including whether autocorrelation, seasonality and/or stationarity were accounted for.[15 16] However, these studies have focused on specific systematic reviews or drug utilisation studies. Our study will extend this research, with a focus on ITS studies in public health. Results of this study will inform the larger project, which aims to provide tools and guidance for researchers designing and analysing ITS studies. The repository of ITS data that we will curate as part of this project will also be of value for future methodological and statistical research.

CONCLUSION

ITS studies are commonly used designs in public health to examine whether an intervention or an exposure has had an effect on health outcomes. However, there is a lack of available information and guidance to inform the design and analysis of ITS studies. Results of this study will help to address this gap by providing information on current design and analysis practice of ITS studies in public health.

Author affiliations
[1]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
[3]School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Ontario, Canada
[4]Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada
[5]Infection Prevention and Healthcare Epidemiology Unit, Alfred Health, Melbourne, Victoria, Australia
[6]Faculty of Pharmacy and Charles Perkins Centre, The University of Sydney, Sydney, New South Wales, Australia

Acknowledgements This work forms part of SLT's PhD, which is supported by an Australian Postgraduate Award administered through Monash University, Australia.

Contributors JEM conceived the study and all authors contributed to its design. SLT wrote the first draft of the manuscript, with contributions from JEM and AK. All authors contributed to revisions of the manuscript and take public responsibility for its content.

Funding SLT was funded through an Australian Postgraduate Award administered through Monash University, Australia. This project is funded by the Australian National Health and Medical Research Council (NHMRC) project grant (1145273). JEM is supported by an NHMRC Career Development Fellowship (1143429). ACC is supported by an NHMRC Career Development Fellowship (1068732). JMG holds a Canada Research Chair in Health Knowledge Uptake and Transfer. The funders had no role in study design, decision to publish, or preparation of the manuscript.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Flynn D, Ford GA, Rodgers H, *et al*. A time series evaluation of the FAST National Stroke Awareness Campaign in England. *PLoS One* 2014;9:e104289.
2. Cheng J, Benassi P, De Oliveira C, *et al*. Impact of a mass media mental health campaign on psychiatric emergency department visits. *Can J Public Health* 2016;107:303–e11.
3. Baker JM, Alonso WJ. Rotavirus vaccination takes seasonal signature of childhood diarrhea back to pre-sanitation era in Brazil. *J Infect* 2018;76:68–77.
4. Leyland AH, Ouedraogo S, Nam J, *et al*. *Public Health Research. Evaluation of Health in Pregnancy grants in Scotland: a natural experiment using routine data*. Southampton (UK): NIHR Journals Library.
5. Milojevic A, Armstrong B, Hashizume M, *et al*. Health effects of flooding in rural Bangladesh. *Epidemiology* 2012;23:107–15.
6. Runkle JD, Zhang H, Karmaus W, *et al*. Prediction of unmet primary care needs for the medically vulnerable post-disaster: an interrupted time-series analysis of health system responses. *Int J Environ Res Public Health* 2012;9:3384–97.
7. Scherb HH, Mori K, Hayashi K. Increases in perinatal mortality in prefectures contaminated by the Fukushima nuclear power plant accident in Japan: a spatially stratified longitudinal study. *Medicine* 2016;95:e4958.
8. 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;27:299–309.
9. Gebski V, Ellingson K, Edwards J, *et al*. Modelling interrupted time series to evaluate prevention and control of infection in healthcare. *Epidemiol Infect* 2012;140:2131–41.
10 Huitema BE, McKean JW. Identifying autocorrelation generated by various error processes in interrupted time-series regression designs. *Educ Psychol Meas* 2007;67:447–59.
11 Zhang F, Wagner AK, Ross-Degnan D. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. *J Clin Epidemiol* 2011;64:1252–61.
12 Barnett AG, Page K, Campbell M, *et al*. Changes in healthcare-associated Staphylococcus aureus bloodstream infections after the introduction of a national hand hygiene initiative. *Infect Control Hosp Epidemiol* 2014;35:1029–36.
13. Zhang F, Wagner AK, Soumerai SB, *et al*. Methods for estimating confidence intervals in interrupted time series analyses of health interventions. *J Clin Epidemiol* 2009;62:143–8.
14. Huitema BE, Mckean JW. Design specification issues in time-series intervention models. *Educ Psychol Meas* 2000;60:38–58.
15. 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;19:613–23.
16. 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;68:950–6.
17. Ewusie JE, Blondal E, Soobiah C, *et al*. Methods, applications, interpretations and challenges of interrupted time series (ITS) data: protocol for a scoping review. *BMJ Open* 2017;7:e016018.
18. Reeves BC DJ, Higgins JPT, Wells GA. *Cochrane handbook for systematic reviews of interventions*. 5.1.0 edn: The Cochrane Collaboration, 2011.
19. Reeves BC, Wells GA, Waddington H. Quasi-experimental study designs series-paper 5: a checklist for classifying studies evaluating the effects on health interventions-a taxonomy without labels. *J Clin Epidemiol* 2017;89:30–42.

20. EPOC. Cochrane Effective Practice and Organisation of Care (EPOC). EPOC Resources for review authors. 2017 epoc.cochrane.org/epoc-resources-review-authors.
21. Huitema BE. *The analysis of covariance and alternatives: statistical methods for experiments, quasi-experiments, and single-case studies*. 2nd ed. Hoboken, N.J: Wiley, 2011.
22. Shamseer L, Sampson M, Bukutu C, *et al*. CENT group. CONSORT extension for reporting N-of-1 trials (CENT) 2015: explanation and elaboration. *J Clin Epidemiol* 2016;76(Supplement C):18–46.
23. Cochrane. About Cochrane Public Health (CPH). 2018 http://ph.cochrane.org/about-cochrane-public-health-cph (accessed May 2018).
24. *WebPlotDigitizer [program]. 4.1 version*. Austin, Texas, USA, 2018.
25. Moeyaert M, Maggin D, Verkuilen J. Reliability, validity, and usability of data extraction programs for single-case research designs. *Behav Modif* 2016;40:874–900.
26. *Access [program]. 2016 version*: Microsoft, 2016.
27. *Stata Statistical Software: Release 15 [program]. 15 version*. Texas: StataCorp LLC, 2017.
28. Hacek DM, Ogle AM, Fisher A, *et al*. Significant impact of terminal room cleaning with bleach on reducing nosocomial Clostridium difficile. *Am J Infect Control* 2010;38:350–3.

Chapter 3. Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: a review

This chapter presents the results of the review of ITS studies evaluating public health interruptions. The protocol for the review is provided in Chapter 2. At the time of undertaking this research, two previous reviews examining design and analysis characteristics of ITS studies had been undertaken; the focus of these reviews was on ITS studies examining the impact of health services interventions (24) and health-care interventions on drug utilisation (9). This review extended the previous research, with a focus on ITS studies in public health and with more in-depth examination of certain design, analysis and reporting aspects. Results from this study informed subsequent chapters of the thesis.

In Chapter 3 the following aspects of the review are presented: results of the search; characteristics of the included studies; characteristics of the designs; outcome types; characteristics of the models; characteristics of the statistical methods; and, reported effect measures and their components. In the discussion, issues surrounding the design, methods and analysis, and reporting are addressed. Implications of the research are also discussed.

Chapter 3 is presented as a manuscript, published in the *Journal of Clinical Epidemiology* in February 2020 (53).

Additional files referred to in the manuscript are appended to this thesis as follows:

Location in thesis	**Referred to in manuscript**	**Content of appendix**
Appendix B	Additional File 1	Deviations, additions and amendments to the protocol
Appendix C	Additional File 2	Review search terms
Appendix D	Additional File 3	Citation details of the 200 studies from which data were extracted

Journal of Clinical Epidemiology 122 (2020) 1−11

REVIEW ARTICLE

Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: a review

Simon L. Turner[a], Amalia Karahalios[a], Andrew B. Forbes[a], Monica Taljaard[b,c], Jeremy M. Grimshaw[b,c,d], Allen C. Cheng[a,e], Lisa Bero[f], Joanne E. McKenzie[a,*]

[a]School of Public Health and Preventive Medicine, Monash University, 533 St. Kilda Road, Melbourne, Victoria 3004, Australia
[b]Clinical Epidemiology Program, Ottawa Hospital Research Institute, 1053 Carling Avenue, Ottawa, Ontario, Canada
[c]School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, 75 Laurier Avenue Eeast, Ottawa, Ontario, Canada
[d]Department of Medicine, University of Ottawa, Roger Guindon Hall, 451 Smyth Roadd, Ottawa, Ontario, Canada
[e]Infection Prevention and Healthcare Epidemiology Unit, Alfred Health, 55 Commercial Road, Melbourne, Victoria 3004, Australia
[f]Faculty of Pharmacy and Charles Perkins Centre, The University of Sydney, John Hopkins Dr, Camperdown NSW, Sydney, New South Wales 2006, Australia

Accepted 20 February 2020; Published online 25 February 2020

Abstract

Objectives: Interrupted time series (ITS) designs are frequently used in public health to examine whether an intervention or exposure has influenced health outcomes. Few reviews have been undertaken to examine the design characteristics, statistical methods, and completeness of reporting of published ITS studies.

Study Design and Setting: We used stratified random sampling to identify 200 ITS studies that evaluated public health interventions or exposures from PubMed (2013−2017). Study characteristics, details of statistical models and estimation methods used, effect metrics, and parameter estimates were extracted. From the 200 studies, 230 time series were examined.

Results: Common statistical methods used were linear regression (31%, 72/230) and autoregressive integrated moving average (19%, 43/230). In 17% (40/230) of the series, we could not determine the statistical method used. Autocorrelation was acknowledged in 63% (145/230) of the series. An estimate of the autocorrelation coefficient was given for only 1% of the series (3/230). Measures of precision were reported for 63% of effect measures (541/852).

Conclusion: Many aspects of the design, methods, analysis, and reporting of ITS studies can be improved, particularly description of the statistical methods and approaches to adjust for and estimate autocorrelation. More guidance on the conduct and reporting of ITS studies is needed to improve this study design.

Keywords: Interrupted time series; Segmented regression; Public health; Statistical methods; Review; Reporting quality; Quasi-experimental

1. Introduction

Randomized trials are the gold standard study design for investigating the impact of an intervention; however, it may not be ethically, practically, or economically possible to use this design [1,2]. In these circumstances, an interrupted time series (ITS) design, which is considered one of the stronger nonrandomized experimental designs, may be considered [3]. In a two-period ITS design, groups are separated in time either side of the interruption (intervention or exposure), rather than into treatment groups as is done in randomized trials. Modeling data from the preinterruption period and using this to predict into the postinterruption period provides a counterfactual for what would

Funding: This work was supported by the Australian National Health and Medical Research Council (NHMRC) project grant (1145273). S.L.T. is funded through an Australian Postgraduate Award administered through Monash University, Australia. J.E.M. is supported by an NHMRC Career Development Fellowship (1143429). A.C.C. is supported by an NHMRC Career Development Fellowship (1068732). J.M.G. holds a Canada Research Chair in Health Knowledge Uptake and Transfer and a Canadian Institute of Health Research (CIHR) Foundation grant (FDN 143269). The funders had no role in study design, decision to publish, or preparation of the manuscript.

Conflict of interests: All authors have no competing interests to declare.

Authors' contributions: J.E.M. conceived the study, and all authors contributed to its design. S.L.T., J.E.M., A.K., and A.B.F. contributed to the data collection. S.L.T. wrote the first draft of the manuscript, with contributions from J.E.M. S.L.T., J.E.M., A.K., A.B.F., M.T., J.M.G., A.C.C., and L.B. contributed to revisions of the manuscript and take public responsibility for its content.

* Corresponding author. Tel.: +61 3 9903 0380; fax: +61 3 9903 0556.
E-mail address: joanne.mckenzie@monash.edu (J.E. McKenzie).

https://doi.org/10.1016/j.jclinepi.2020.02.006
0895-4356/© 2020 Elsevier Inc. All rights reserved.

What is new?

Key findings

- In our review of 200 interrupted time series (ITS) studies, the series length in most studies was likely to be too small to be able to adequately account for any secular or seasonal trends and autocorrelation.
- Reporting of statistical methods and the approaches to adjust for and estimate autocorrelation was often incomplete. Effect estimates were often reported without measures of precision, and estimates of autocorrelation were almost never reported.
- Drawing conclusions from ITS studies can be strengthened using a current control series, but only one-fourth of studies included some form of control.

What this adds to what was known?

- Three previous reviews have examined the design characteristics and statistical methods used in ITS studies. Our review provides a recent assessment of the analysis and reporting practices of ITS studies, extends the design features and statistical methods examined, and has a focus on ITS studies in public health.

What is the implication and what should change now?

- Where possible, longer series should be used more frequently. The authors should detail the statistical method used as well as explicitly describing how autocorrelation was estimated and adjusted for and provide estimates of autocorrelation. Estimates of precision for effect measures should be provided. When feasible, a control series should be considered.

have occurred in the absence of the interruption, analogous to a comparator group in a randomized trial.

There are a number of factors to consider when designing, analyzing, and reporting the results from ITS studies. First, there is a tendency for data collected over time to be correlated (known as serial correlation or autocorrelation [4]), so the selected sample size and analysis methods need to appropriately account for the autocorrelation. Failing to account for (likely) positive autocorrelation can lead to an underestimation of the required sample size or standard errors that are too small, with resulting confidence intervals (CIs) that are too narrow and P-values that are too small, leading to inflated type I error rates [5]. Second, decisions about the structure of the model (e.g., the number of segments, timing of the impact of the interruption, and shape) need to be made. Third, a statistical method needs to be chosen from the wide range of available methods. Finally, effect metric(s) that characterize the impact of the interruption need to be selected. Even in the case of a simple two-period design where a segmented linear model is fitted, a variety of effect metrics are available (e.g., change in level at the point of the interruption and change in slope after the interruption).

Although a number of tutorial papers are available that provide general advice on the design and analysis of ITS designs (e.g., Wagner et al. [6], Lopez Bernal et al. [2], and Penfold and Zhang [7]), there is currently limited guidance to inform the sample size and selection of statistical methods. Sample size calculations require knowledge of parameters, such as autocorrelation coefficients [8], of which there are currently no databases providing empirical estimates. Furthermore, although there are many methods available for statistical analysis of these studies, there is limited guidance identifying those methods to avoid or those that are acceptable and in what circumstances. It is also not known which effect metrics are the most robust to misspecification of the model and estimation methods.

This review is the first stage of a larger project that aims to provide tools and guidance for public health researchers on the design and analysis of ITS studies [9]. This stage of the project aimed to examine the design characteristics, statistical methods, and completeness of reporting in a random sample of contemporary ITS studies examining public health interruptions (interventions or exposures) on health-related outcomes.

2. Review methods

A detailed protocol for this study has been published [9]. Here, we provide a brief overview of the methods along with deviations from the protocol (Additional File 1).

2.1. Eligibility criteria

Our definition of an ITS design was informed by the design features taxonomy for quasi-experimental designs [10,11]. ITS studies meeting the following criteria were included: (1) there were at least two segments separated by a clearly defined intervention or exposure with at least three points in each segment; (2) observations were collected on a group of individuals (e.g., community and hospital) at each time point; and (3) the study investigated the impact of a public health intervention or exposure that has public health implications. ITS studies meeting any of the following criteria were excluded: (1) written in a language other than English, (2) single-case design, or (3) were reported in a methodological paper examining ITS studies.

2.2. Literature search

The bibliographic database PubMed was searched (to December 12, 2017) using terms informed from previous

search strategies developed to locate ITS studies [12,13], terms used to describe the design in the methods section of published ITS studies (e.g., [14—16]), and controlled vocabulary (Additional File 2).

2.3. Sample size and study selection

We included a sample of 200 ITS studies, which allowed estimation of the percentage of ITS studies with a particular element within a maximum margin of error of 7% (assuming a prevalence of 50%). We included studies published between 2013 and 2017. Studies were stratified by year and, within each year, were randomly sampled until 40 were identified that met the inclusion criteria. We had planned to sample earlier years if we did not meet the target sample for any particular year, but this was not required.

Four authors (S.L.T., A.B.F., A.K., and J.E.M.) piloted the eligibility criteria on 50 abstracts. A further 450 abstracts were screened by two authors (S.L.T. and one of A.B.F., A.K., or J.E.M.). The remaining abstracts (717) were screened by one author (S.L.T.) and checked with at least two authors (A.B.F., A.K., J.E.M.) when there was uncertainty about any of the inclusion criteria and in all cases where inclusion criterion 3 was met (i.e., whether the study investigated the impact of a public health intervention or exposure that had public health implications). This latter check was undertaken because deciding on whether the study investigated a public health interruption was a subjective criterion to apply.

Four authors (S.L.T., A.B.F., A.K., and J.E.M.) piloted the eligibility criteria on 10 full-text articles. Two authors (S.L.T. and A.K.) screened 44 full-text articles, and the remaining articles were screened by one author (S.L.T.) until 200 ITS studies were identified.

2.4. Outcome(s) selection

For each included ITS study, multiple outcomes were potentially eligible for inclusion. We developed the following hierarchy, a priori, to select an outcome within each category of outcome type (binary, continuous, count, and proportion): (1) ITS data availability—outcomes with data available to be extracted (either from tables or figures) were selected ahead of those without data, (2) stated primary outcome (or reported in the title or objectives), (3) first reported result outcome in the abstract, or (4) first reported outcome in the results.

Other forms of multiplicity of time series arose that we had not considered a priori, and for which we developed selection rules *post hoc*. These included (1) selection of the time series with the interruption in studies with a control group and (2) selection of the first reported subgroup time series in studies with multiple subgroups.

Uncertainty in the selection of the review time series and outcomes were discussed by four authors (S.L.T., A.B.F., A.K., and J.E.M.).

2.5. Data extraction and management

The data extraction process was initially piloted by four authors (S.L.T., A.B.F., A.K., and J.E.M.) on a sample of 10 studies. Data extraction was performed for 46 (23%) papers (and their supplementary files, when available) by two or more authors (S.L.T. and one or more of A.B.F., A.K., and J.E.M.). A single author (S.L.T.) extracted data from a further 154 papers. We did not observe a high degree of inconsistency in the double data extraction for any of the items. Any discrepancies and uncertainty in the coding were discussed by the review team (S.L.T. and two or more of A.B.F., A.K., and J.E.M.). These discussions resulted in the addition of some new items and clarifications to the coding of some items. When clarifications were made, extracted data for those items were checked in all articles.

We extracted data on the study characteristics, design, outcome, model, statistical methods, and effect metrics (Table 1). A complete list is available in the protocol [9], with amendments and new items outlined in Additional File 1.

2.6. Analysis

We present descriptive summary statistics. For some outcomes, we calculated a difference in percentages between groups (defined by model type) with 95% CIs. Statistical analyses were undertaken using Stata Release 15 [17].

3. Results

3.1. Results of the search

Of the 6,548 abstracts identified through the search, 2,904 were published within the 5-year scope of the review (2013—2017). These 2,904 abstracts were randomly sorted and screened until we had likely identified the required number of studies that would meet our inclusion criteria after full-text screening. This involved screening 1,218 abstracts, of which 279 were eligible for inclusion. Full text screening was then performed (in random order) on 227 of the 279 potentially eligible studies, until the target sample size of 200 was reached. Data were extracted from these 200 studies, which included 230 time series (Fig. 1; Additional File 3).

3.2. Study characteristics

Nearly all the studies evaluated the impact of interventions (94%, 188/200), and the rest evaluated the impact of exposures (e.g., effect of 2008 financial crisis on suicide rates [18] and the impact of the 2011 Fukushima nuclear power plant accident on perinatal mortality [19]; Table 2). Over half of the studies (52%, 104/200) investigated policy changes. Interventions targeting practice change (e.g., the impact of antibiotic stewardship program introduction of daily chlorhexidine bathing on hospital

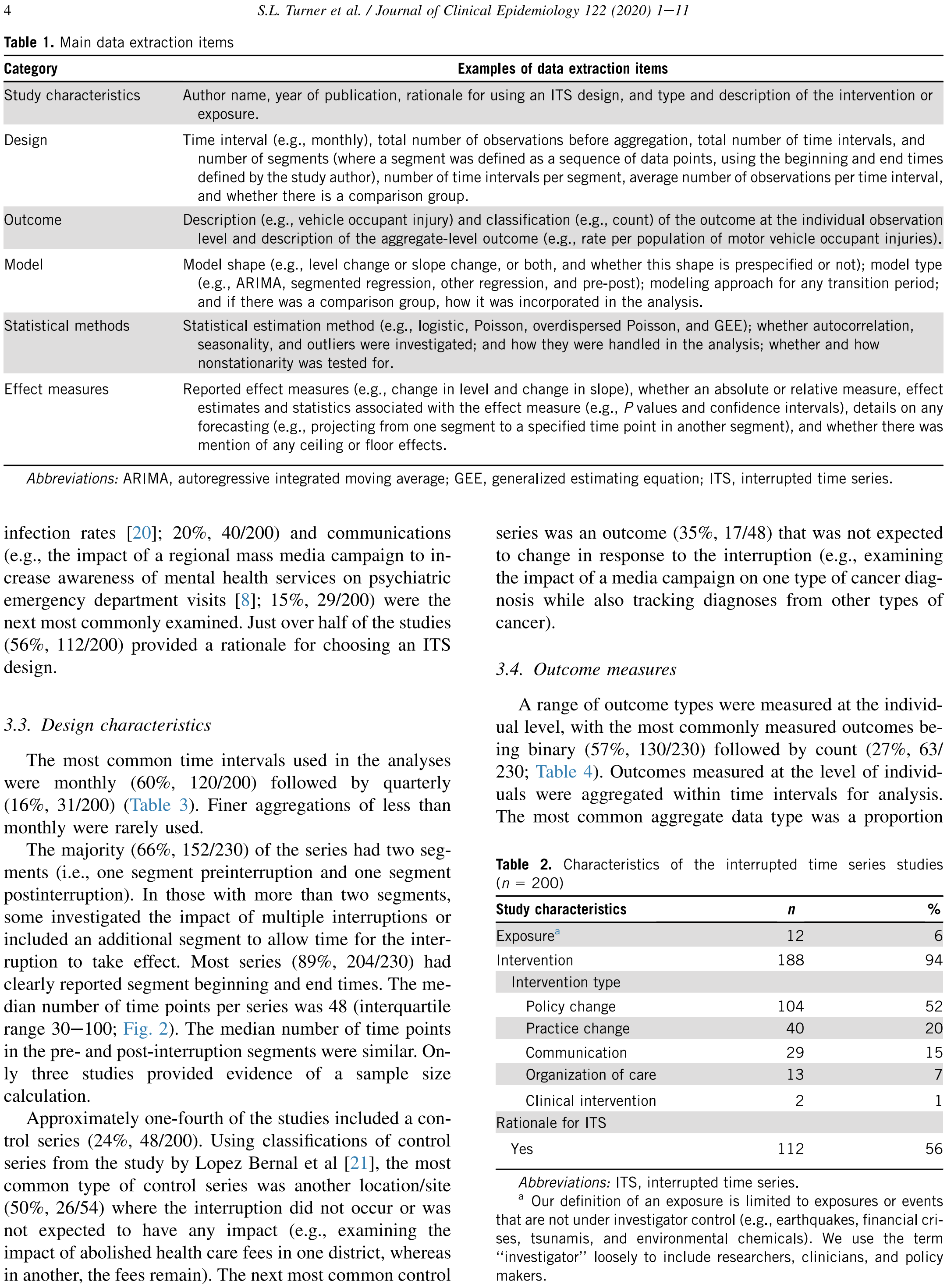


Table 1. Main data extraction items

Category	Examples of data extraction items
Study characteristics	Author name, year of publication, rationale for using an ITS design, and type and description of the intervention or exposure.
Design	Time interval (e.g., monthly), total number of observations before aggregation, total number of time intervals, and number of segments (where a segment was defined as a sequence of data points, using the beginning and end times defined by the study author), number of time intervals per segment, average number of observations per time interval, and whether there is a comparison group.
Outcome	Description (e.g., vehicle occupant injury) and classification (e.g., count) of the outcome at the individual observation level and description of the aggregate-level outcome (e.g., rate per population of motor vehicle occupant injuries).
Model	Model shape (e.g., level change or slope change, or both, and whether this shape is prespecified or not); model type (e.g., ARIMA, segmented regression, other regression, and pre-post); modeling approach for any transition period; and if there was a comparison group, how it was incorporated in the analysis.
Statistical methods	Statistical estimation method (e.g., logistic, Poisson, overdispersed Poisson, and GEE); whether autocorrelation, seasonality, and outliers were investigated; and how they were handled in the analysis; whether and how nonstationarity was tested for.
Effect measures	Reported effect measures (e.g., change in level and change in slope), whether an absolute or relative measure, effect estimates and statistics associated with the effect measure (e.g., *P* values and confidence intervals), details on any forecasting (e.g., projecting from one segment to a specified time point in another segment), and whether there was mention of any ceiling or floor effects.

Abbreviations: ARIMA, autoregressive integrated moving average; GEE, generalized estimating equation; ITS, interrupted time series.

infection rates [20]; 20%, 40/200) and communications (e.g., the impact of a regional mass media campaign to increase awareness of mental health services on psychiatric emergency department visits [8]; 15%, 29/200) were the next most commonly examined. Just over half of the studies (56%, 112/200) provided a rationale for choosing an ITS design.

3.3. Design characteristics

The most common time intervals used in the analyses were monthly (60%, 120/200) followed by quarterly (16%, 31/200) (Table 3). Finer aggregations of less than monthly were rarely used.

The majority (66%, 152/230) of the series had two segments (i.e., one segment preinterruption and one segment postinterruption). In those with more than two segments, some investigated the impact of multiple interruptions or included an additional segment to allow time for the interruption to take effect. Most series (89%, 204/230) had clearly reported segment beginning and end times. The median number of time points per series was 48 (interquartile range 30–100; Fig. 2). The median number of time points in the pre- and post-interruption segments were similar. Only three studies provided evidence of a sample size calculation.

Approximately one-fourth of the studies included a control series (24%, 48/200). Using classifications of control series from the study by Lopez Bernal et al [21], the most common type of control series was another location/site (50%, 26/54) where the interruption did not occur or was not expected to have any impact (e.g., examining the impact of abolished health care fees in one district, whereas in another, the fees remain). The next most common control series was an outcome (35%, 17/48) that was not expected to change in response to the interruption (e.g., examining the impact of a media campaign on one type of cancer diagnosis while also tracking diagnoses from other types of cancer).

3.4. Outcome measures

A range of outcome types were measured at the individual level, with the most commonly measured outcomes being binary (57%, 130/230) followed by count (27%, 63/230; Table 4). Outcomes measured at the level of individuals were aggregated within time intervals for analysis. The most common aggregate data type was a proportion

Table 2. Characteristics of the interrupted time series studies (n = 200)

Study characteristics	*n*	%
Exposure[a]	12	6
Intervention	188	94
Intervention type		
Policy change	104	52
Practice change	40	20
Communication	29	15
Organization of care	13	7
Clinical intervention	2	1
Rationale for ITS		
Yes	112	56

Abbreviations: ITS, interrupted time series.

[a] Our definition of an exposure is limited to exposures or events that are not under investigator control (e.g., earthquakes, financial crises, tsunamis, and environmental chemicals). We use the term ''investigator'' loosely to include researchers, clinicians, and policy makers.

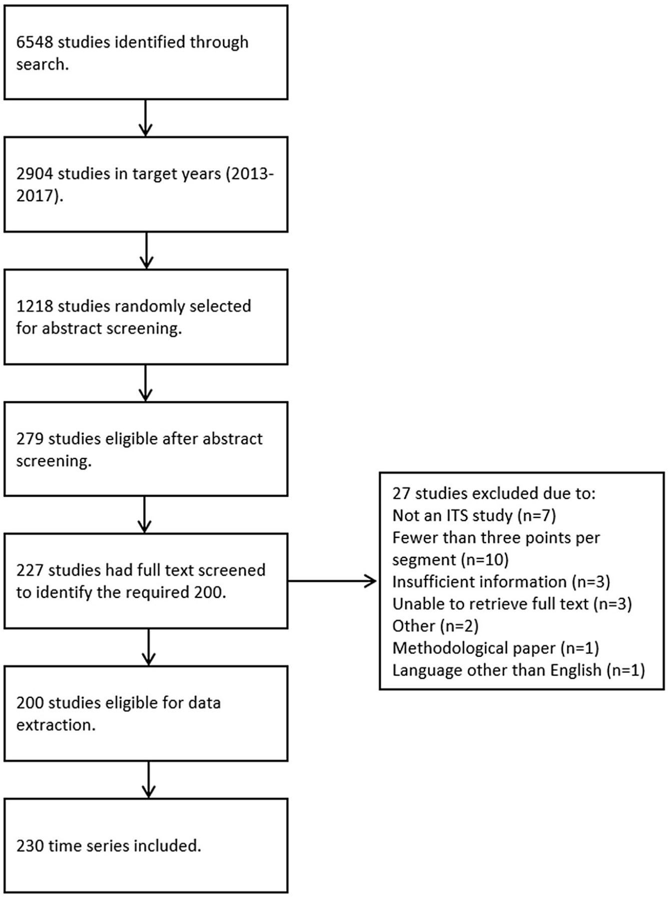

6548 studies identified through search.
2904 studies in target years (2013-2017).
1218 studies randomly selected for abstract screening.
279 studies eligible after abstract screening.
227 studies had full text screened to identify the required 200.
27 studies excluded due to:
Not an ITS study (n=7)
Fewer than three points per segment (n=10)
Insufficient information (n=3)
Unable to retrieve full text (n=3)
Other (n=2)
Methodological paper (n=1)
Language other than English (n=1)
200 studies eligible for data extraction.
230 time series included.

Fig. 1. Flow diagram of search results. Search terms can be found in Table 1. ITS, interrupted time series.

(50%, 115/230), with continuous (18%, 41/230), count (19%, 43/230), and rate (13%, 31/230) similarly used. Within individual-level outcome types, data were aggregated in different ways (Fig. 3). For example, individual binary-level outcomes were aggregated into proportions (72%, 94/130), counts (18%, 24/130), rates (7%, 9/130), and continuous (2%, 3/130) outcomes.

3.5. Model structure characteristics

The model structure used most frequently included both a level change and a slope change (70%, 161/230; Table 5). In some series, only a level change was modeled (10%, 23/230; akin to fitting a mean preinterruption and postinterruption), and in a smaller proportion, only a slope was modeled (3%, 7/230; i.e., examining a trend over time ignoring any interruption(s)). In 10% of series (24/230), it was not possible to determine the model structure.

Almost all studies (91%, 182/200) modeled the timing of the impact of the interruption in the analysis using a specified date, as opposed to searching for the date where the interruption had the largest impact. The impact of the interruption in the time series may not be expected to occur concurrently with the interruption itself, and 34% (78/230) allowed for such a delay. This delay was modeled using a variety of methods, with the most common (37%, 29/72) being to exclude the period between the interruption and hypothesized impact. It was also common (33%, 26/72) for the delay to be incorporated into one of the preinterruption segments. A separate “roll-in” or “implementation period” segment was modeled in 15% (12/72) of the series.

3.6. Statistical methods

A range of statistical methods were reported as being used to analyze the ITS studies. Most commonly linear

Table 3. Design characteristics of included studies

Design characteristics	*n* or median	% or IQR
Time interval type[a]		
Daily	3	2
Weekly	9	5
Two weekly	1	1
Monthly	120	60
Quarterly	31	16
Six monthly	3	2
Annually	20	10
Other	12	6
Cannot determine	1	1
Was the segment timing clear[b]		
Yes	204	89
Number of segments per series[b]	2	2–3
Number of series with only two segments[b]	152	66
Number of time points per series[b]		
All series	48	30–100
ARIMA model used (n = 43)	108	49–144
Non-ARIMA model used (n = 187)	45	27–72
Number of time points per segment[c]		
All segments (n = 476)	20	12–35
First segment (n = 204)	21	12–36
Second segment (n = 198)	20	12–36
Third or higher segment (n = 74)	15	8–26
Was there a sample size calculation? [a]		
Yes	3	1.5
Was there a control series?[a]		
Yes	48	24
Type of control[d]		
Location[e]	26	54
Outcome[e]	17	35
Behavior	3	6
Characteristic	2	4
Historical	2	4

Abbreviations: ARIMA, autoregressive integrated moving average; IQR, interquartile range.

[a] Denominator is the number of studies, n = 200.

[b] Denominator is the number of series because multiple outcomes per study were eligible for inclusion, n = 230. Each time series was partitioned into two or more discrete segments where a segment was defined as a sequence of data points, using the beginning and end times defined by the study author.

[c] Denominator is the number of segments used in analysis, where the segment timing could be determined from the report.

[d] Denominator is the number of series with a control series, n = 48. Classifications from the study by Lopez Bernal [21]. Location (e.g., a different area), outcome (e.g., an outcome not affected by the intervention), behavior (e.g., a group of individuals who never performed the behavior being investigated), characteristic (e.g., a group not targeted by an intervention designed to target a group holding a certain characteristic), historical (e.g., comparing a previous age group to a current age group) [21].

[e] Two studies used both a location and an outcome-based control series, thus percentages sum to greater than 100%.

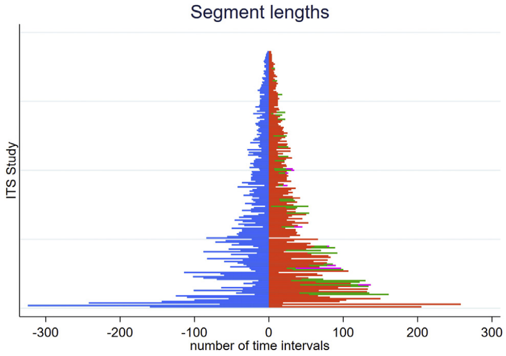


Fig. 2. Number of time points in each series separated by segment. Each horizontal bar represents the length of a time series (in terms of number of data points). Colors indicate segment (first segment = blue, second segment = orange, third segment = green, and fourth segment = pink). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

regression (31%, 72/230) and ARIMA (19%, 43/230) methods were used (Table 6). Linear regression methods were used both with and without adjustment for autocorrelation (14%, 33/230 and 11%, 25/230, respectively). Some of the autocorrelation adjustment methods included Prais-Winsten (2.6%, 6/230), Cochrane-Orcutt (1.7%, 4/230), and the use of Newey-West standard errors (0.8%,2/230). In 17% (40/230) of the series, we could not determine the statistical method used.

In 30% (68/230) of the series, the authors referred to or provided details of testing for model fit. This differed by the statistical method used, with testing of model fit more frequently reported when ARIMA was used (81%, 35/43) compared with non-ARIMA methods (18%, 33/187; difference 64% [95% CI 51%–77%]).

In approximately two-thirds (63%, 145/230) of the series, autocorrelation was acknowledged. In studies using ARIMA methods, autocorrelation was acknowledged more frequently (91%, 39/43) compared with studies using other statistical methods (57%, 106/187; difference 34% [95% CI 23%–45%]). Of the 187 series in which a non-ARIMA method was used, 40% (74/187) stated a method for handling autocorrelation. Specifically, 29% (55/187) reported using a statistical method that adjusted for autocorrelation, and 10% (19/187) determined that adjustment for autocorrelation was unnecessary (either from a statistical test for autocorrelation that was nonsignificant or a sensitivity analysis that yielded similar results for methods that did and did not adjust for autocorrelation). A range of detection methods for identifying autocorrelation were used, with the Durbin–Watson test being the most common (38%, 55/145). Of these, the observed Durbin–Watson statistic was reported in only 29% (16/55) of the series. There was insufficient information to determine which detection

Table 4. Characteristics of the outcomes, n = 230

Outcome characteristics	n	%
Data type—individual		
Binary	130	57
Count	63	27
Continuous	34	15
Proportion	3	1
Data type—aggregate		
Proportion	115	50
Count	43	19
Continuous	41	18
Rate	31	13

method had been used in 24% (35/145) of the series. An estimate of the autocorrelation coefficient was reported in only three of the 230 series (1%).

Longer term periodic (seasonal) trends in the data were not commonly described (37%, 84/230). A large percentage of ARIMA models reported adjusting for seasonality (72%, 31/43); however, when non-ARIMA methods were used, investigation of seasonality was less common (28%, 53/187; difference 44% [95% CI 29%—59%]). Nonstationarity was acknowledged in 22% (50/230) of the series. When ARIMA was used, nonstationarity was reported in 81% (35/43) of the series, whereas when non-ARIMA methods were used, investigation of nonstationarity was reported only in 8% (15/187) of the series (difference 73% [95% CI 61%—86%]).

Of the 48 studies that included a control/comparison group, 44% (21/48) used a unified model that included the control/comparison and intervention series. A narrative approach was used in 50% (24/48) of these series, where differences in the series were described or effect measures were stated separately for each group, or both, but no formal statistical comparison was undertaken.

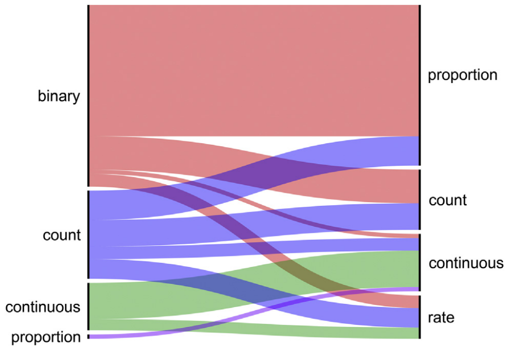


Fig. 3. Mapping of individual-level outcome types to aggregate outcome types. The thickness of each line is proportional to the number of outcomes mapping between individual- and aggregate-level outcome types.

Table 5. Characteristics of the model structure

Model characteristics	n	%
Shape[a]		
Level change only	23	10
Slope change only	7	3
Level and slope change	161	70
Other	15	7
Cannot determine	24	10
Impact of interruption modeled at a specified point[b]		
Yes	182	91
Allowed for a delayed impact[a]		
Yes	78	34
Method of dealing with delay[c]		
Excluded times between interruption and change point	29	37
Change point modeled at a later time to interruption	26	33
Modeled separate segment between change point and interruption	12	15
Sensitivity analysis	4	5
Unclear	5	6
Other	2	3

[a] Denominator is the number of series, n = 230.
[b] Denominator is the number of studies, n = 200.
[c] Denominator is the number of series with a delayed impact, n = 78.

A variety of statistical packages were used, with SAS (28%, 56/200) and Stata (24%, 48/200) noted most commonly. Graphs depicting the time series data were reported for nearly all the eligible series (93%, 214/230).

3.7. Effect measures and components

Commonly, multiple effect measures (e.g., change in level) and components of effect measures (e.g., preintervention slope) were reported from the analysis of each series (median 4, IQR 2—5; Table 7). These effect measures characterize the impact of the interruption in different ways (e.g., short-term or long-term impact; Fig. 4).

The most frequently reported effect measures were the level change at the point of interruption (70%, 160/230; a measure of short-term impact) and slope change (54%, 125/230; a measure of long-term impact). The pre- and post-interruption means were reported in 28% (65/230) and 33% (76/230) of the series, respectively, and a difference in these means was reported in 17% of series (38/230). In 25% (58/230) of the series, the authors used the preinterruption trend to forecast into the postinterruption period.

For 27% of the series (63/230), all three of the major coefficients (preinterruption slope, level change at the point of interruption, and slope change) were reported. Similarly, for 27% (62/230) of the series, none of these measures were

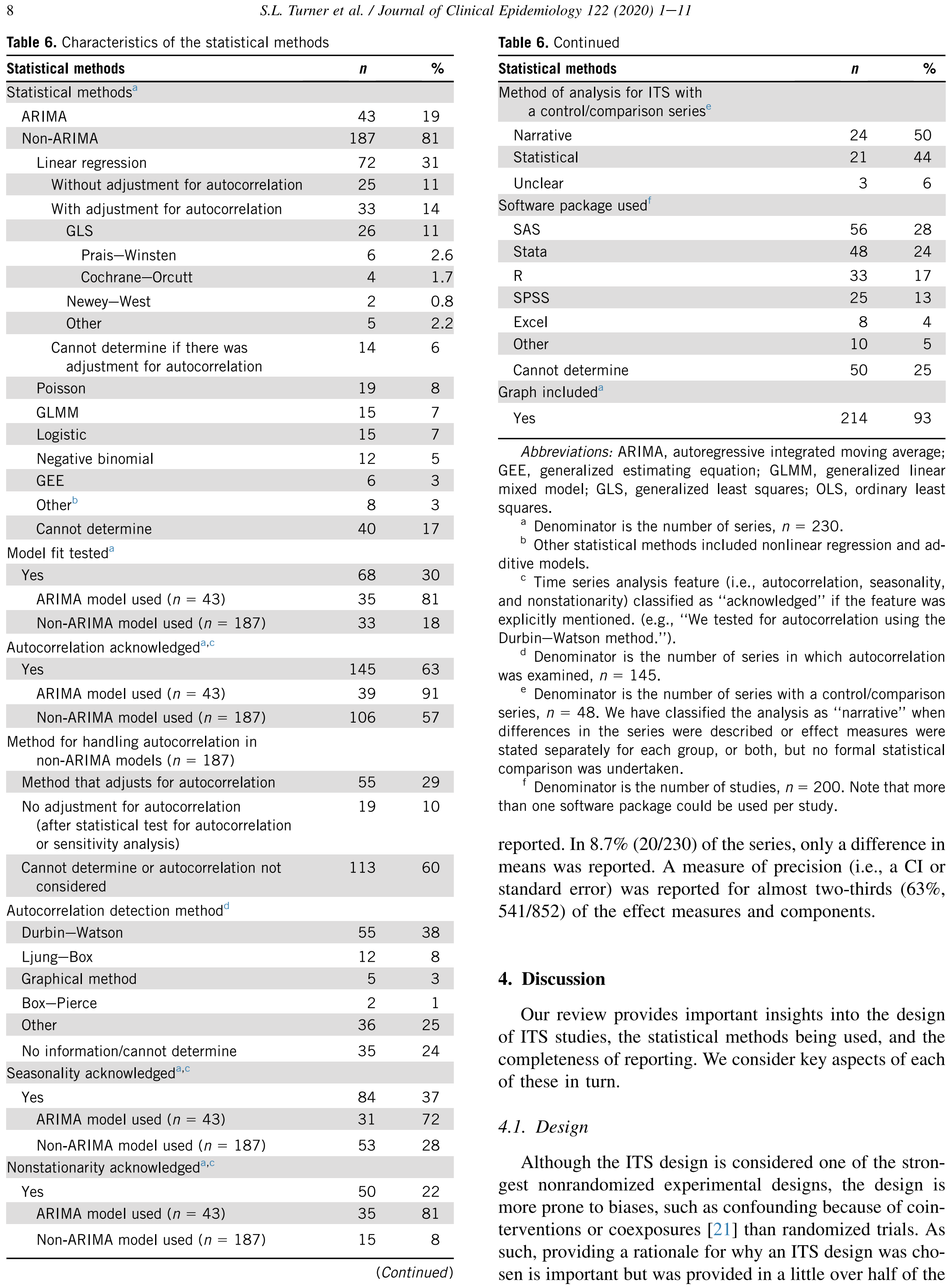


Table 6. Characteristics of the statistical methods

Statistical methods	n	%
Statistical methods[a]		
ARIMA	43	19
Non-ARIMA	187	81
Linear regression	72	31
Without adjustment for autocorrelation	25	11
With adjustment for autocorrelation	33	14
GLS	26	11
Prais–Winsten	6	2.6
Cochrane–Orcutt	4	1.7
Newey–West	2	0.8
Other	5	2.2
Cannot determine if there was adjustment for autocorrelation	14	6
Poisson	19	8
GLMM	15	7
Logistic	15	7
Negative binomial	12	5
GEE	6	3
Other[b]	8	3
Cannot determine	40	17
Model fit tested[a]		
Yes	68	30
ARIMA model used ($n = 43$)	35	81
Non-ARIMA model used ($n = 187$)	33	18
Autocorrelation acknowledged[a,c]		
Yes	145	63
ARIMA model used ($n = 43$)	39	91
Non-ARIMA model used ($n = 187$)	106	57
Method for handling autocorrelation in non-ARIMA models ($n = 187$)		
Method that adjusts for autocorrelation	55	29
No adjustment for autocorrelation (after statistical test for autocorrelation or sensitivity analysis)	19	10
Cannot determine or autocorrelation not considered	113	60
Autocorrelation detection method[d]		
Durbin–Watson	55	38
Ljung–Box	12	8
Graphical method	5	3
Box–Pierce	2	1
Other	36	25
No information/cannot determine	35	24
Seasonality acknowledged[a,c]		
Yes	84	37
ARIMA model used ($n = 43$)	31	72
Non-ARIMA model used ($n = 187$)	53	28
Nonstationarity acknowledged[a,c]		
Yes	50	22
ARIMA model used ($n = 43$)	35	81
Non-ARIMA model used ($n = 187$)	15	8
Method of analysis for ITS with a control/comparison series[e]		
Narrative	24	50
Statistical	21	44
Unclear	3	6
Software package used[f]		
SAS	56	28
Stata	48	24
R	33	17
SPSS	25	13
Excel	8	4
Other	10	5
Cannot determine	50	25
Graph included[a]		
Yes	214	93

(*Continued*)

Table 6. Continued

Abbreviations: ARIMA, autoregressive integrated moving average; GEE, generalized estimating equation; GLMM, generalized linear mixed model; GLS, generalized least squares; OLS, ordinary least squares.

[a] Denominator is the number of series, $n = 230$.

[b] Other statistical methods included nonlinear regression and additive models.

[c] Time series analysis feature (i.e., autocorrelation, seasonality, and nonstationarity) classified as "acknowledged" if the feature was explicitly mentioned. (e.g., "We tested for autocorrelation using the Durbin–Watson method.").

[d] Denominator is the number of series in which autocorrelation was examined, $n = 145$.

[e] Denominator is the number of series with a control/comparison series, $n = 48$. We have classified the analysis as "narrative" when differences in the series were described or effect measures were stated separately for each group, or both, but no formal statistical comparison was undertaken.

[f] Denominator is the number of studies, $n = 200$. Note that more than one software package could be used per study.

reported. In 8.7% (20/230) of the series, only a difference in means was reported. A measure of precision (i.e., a CI or standard error) was reported for almost two-thirds (63%, 541/852) of the effect measures and components.

4. Discussion

Our review provides important insights into the design of ITS studies, the statistical methods being used, and the completeness of reporting. We consider key aspects of each of these in turn.

4.1. Design

Although the ITS design is considered one of the strongest nonrandomized experimental designs, the design is more prone to biases, such as confounding because of cointerventions or coexposures [21] than randomized trials. As such, providing a rationale for why an ITS design was chosen is important but was provided in a little over half of the

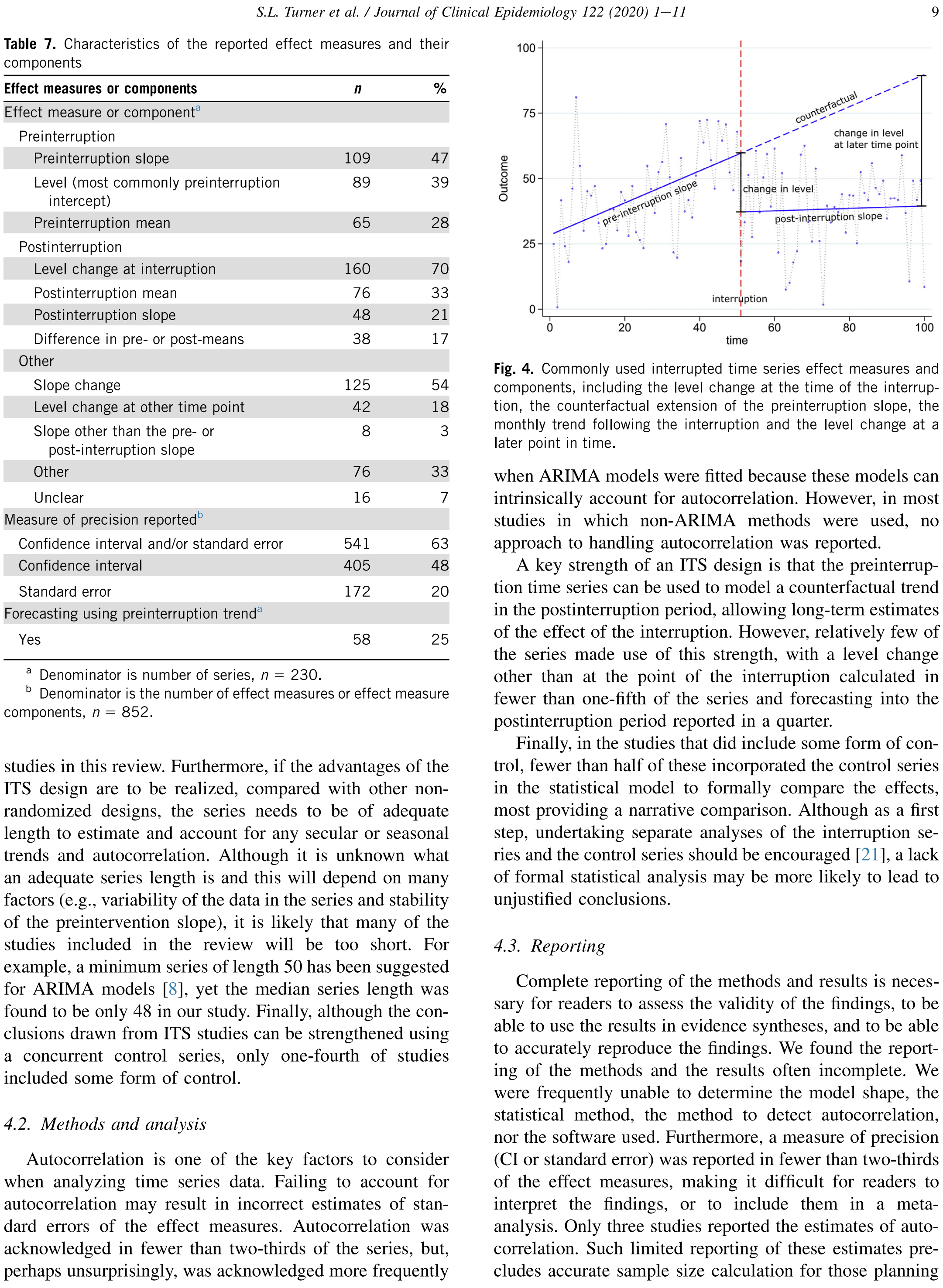


Table 7. Characteristics of the reported effect measures and their components

Effect measures or components	*n*	%
Effect measure or component[a]		
Preinterruption		
Preinterruption slope	109	47
Level (most commonly preinterruption intercept)	89	39
Preinterruption mean	65	28
Postinterruption		
Level change at interruption	160	70
Postinterruption mean	76	33
Postinterruption slope	48	21
Difference in pre- or post-means	38	17
Other		
Slope change	125	54
Level change at other time point	42	18
Slope other than the pre- or post-interruption slope	8	3
Other	76	33
Unclear	16	7
Measure of precision reported[b]		
Confidence interval and/or standard error	541	63
Confidence interval	405	48
Standard error	172	20
Forecasting using preinterruption trend[a]		
Yes	58	25

[a] Denominator is number of series, $n = 230$.

[b] Denominator is the number of effect measures or effect measure components, $n = 852$.

studies in this review. Furthermore, if the advantages of the ITS design are to be realized, compared with other nonrandomized designs, the series needs to be of adequate length to estimate and account for any secular or seasonal trends and autocorrelation. Although it is unknown what an adequate series length is and this will depend on many factors (e.g., variability of the data in the series and stability of the preintervention slope), it is likely that many of the studies included in the review will be too short. For example, a minimum series of length 50 has been suggested for ARIMA models [8], yet the median series length was found to be only 48 in our study. Finally, although the conclusions drawn from ITS studies can be strengthened using a concurrent control series, only one-fourth of studies included some form of control.

4.2. Methods and analysis

Autocorrelation is one of the key factors to consider when analyzing time series data. Failing to account for autocorrelation may result in incorrect estimates of standard errors of the effect measures. Autocorrelation was acknowledged in fewer than two-thirds of the series, but, perhaps unsurprisingly, was acknowledged more frequently when ARIMA models were fitted because these models can intrinsically account for autocorrelation. However, in most studies in which non-ARIMA methods were used, no approach to handling autocorrelation was reported.

Fig. 4. Commonly used interrupted time series effect measures and components, including the level change at the time of the interruption, the counterfactual extension of the preinterruption slope, the monthly trend following the interruption and the level change at a later point in time.

A key strength of an ITS design is that the preinterruption time series can be used to model a counterfactual trend in the postinterruption period, allowing long-term estimates of the effect of the interruption. However, relatively few of the series made use of this strength, with a level change other than at the point of the interruption calculated in fewer than one-fifth of the series and forecasting into the postinterruption period reported in a quarter.

Finally, in the studies that did include some form of control, fewer than half of these incorporated the control series in the statistical model to formally compare the effects, most providing a narrative comparison. Although as a first step, undertaking separate analyses of the interruption series and the control series should be encouraged [21], a lack of formal statistical analysis may be more likely to lead to unjustified conclusions.

4.3. Reporting

Complete reporting of the methods and results is necessary for readers to assess the validity of the findings, to be able to use the results in evidence syntheses, and to be able to accurately reproduce the findings. We found the reporting of the methods and the results often incomplete. We were frequently unable to determine the model shape, the statistical method, the method to detect autocorrelation, nor the software used. Furthermore, a measure of precision (CI or standard error) was reported in fewer than two-thirds of the effect measures, making it difficult for readers to interpret the findings, or to include them in a meta-analysis. Only three studies reported the estimates of autocorrelation. Such limited reporting of these estimates precludes accurate sample size calculation for those planning future studies. The importance of reporting correlation coefficients has been recognized for other designs (e.g., cluster randomized trials [22]), but as our review identifies, has not yet gained recognition for ITS designs. Recommendations for reporting (Jandoc et al. [3]) and the development of a reporting guideline for ITS designs [23] may facilitate improved reporting.

5. What this study adds to what is already known

We identified three previous studies that have examined the design characteristics and statistical methods used in ITS studies. Ramsay et al. [12] examined ITS studies included in two systematic reviews (published from 1976 to 1998). The first review included 20 ITS studies examining the effects of mass media interventions targeted at improving the utilization of health services, whereas the second review included 38 ITS studies examining the effects of different clinical guideline dissemination and implementation strategies. Jandoc et al. [3] examined a cohort of 220 ITS studies in drug utilization research (published from 1994 to 2013). Hudson et al. [24] included 116 ITS studies that assessed health care interventions (published in 2015). Our review differs in that it provides a recent assessment of the analysis and reporting practices of ITS studies, extends the design features and statistical methods examined, and has a focus on ITS studies in public health. Nevertheless, we found many similarities with Ramsay et al. [12], Jandoc et al. [3], and Hudson et al. [24]. They also identified poor reporting of the study design, methodology, and autocorrelation. In addition, effect estimates were often reported without estimates of precision, and few studies forecasted using preinterruption data (Jandoc et al. [3] and Hudson et al. [24]). We also noted some differences. The median number of included time points was greater in our sample compared with the ITS studies included in Ramsey et al. [12] (median 48 points vs. median 20 points); we found a greater proportion of studies using graphical figures (93% vs. 84%), and fewer studies using a control series (24% vs. 35%) compared with Jandoc et al. [3]; and we found a greater proportion of studies reporting nonstationarity (22% vs. 8%) and seasonality (37% vs. 24%) compared with Hudson et al. [24].

6. Strengths and limitations

We used systematic review methods when addressing selection, inclusion, and exclusion criteria as well as for data extraction. We outlined and followed our methods in a published protocol [9].

One limitation of this review was the use of a single database (PubMed) to select studies. However, we feel that PubMed provides a sufficient sampling frame, given its extensive coverage of public health and health service research journals. As well, our search strategy may not have located all ITS studies. Although we used free-text terms derived from a range of sources and the MeSH term for ITS studies, the sensitivity of the strategy is unknown.

Finally, as with all studies examining the design and methods in reported articles, assessing the conduct of studies is difficult, and may be inaccurate, when the reporting of these aspects is incomplete. An example of this is the reporting of statistical methods, where it can be difficult to determine the estimation method used because this is often not reported.

7. Implications of this research

Our research has implications for those designing and conducting ITS studies. For the former group, our results suggest that longer series should be used more frequently, that autocorrelation should be handled explicitly and that an appropriate statistical method should be used and fully reported. Estimates of the longer term effects of the interruption should be considered more frequently. When feasible, a control series might be considered and, when used, methods for handling the control group in the analysis should be prespecified [21]. Furthermore, investigators should collect data on possible external interventions, programs, or events that may confound the interruption. Effect estimates should be reported with the inclusion of measures of precision (such as CIs and standard errors) and estimates of autocorrelation coefficients should be provided.

8. Conclusion

ITS studies are commonly used designs in public health to examine whether an intervention or an exposure has influenced health outcomes. Our review found that there were many aspects of the reporting of methods and results that could be improved, especially the reporting of the statistical methods and approaches used to adjust for and estimate autocorrelation. Many ITS designs could be strengthened using longer series and more frequent use of control groups. More guidance on the conduct and reporting of ITS studies is needed to improve this important study design.

Acknowledgments

This work forms part of S.L.T.'s PhD, which is supported by an Australian Postgraduate Award administered through Monash University, Australia.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.02.006.

References

[1] Lagarde M. How to do (or not to do) assessing the impact of a policy change with routine longitudinal data. Health Policy Plan 2011;27(1): 76—83.

[2] Lopez-Bernal J, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 2017;46:348—55.

[3] Jandoc R, Burden A, Mamdani M, Lèvesque L, Cadarette S. Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. J Clin Epidemiol 2015;68:950—6.

[4] Gebski V, Ellingson K, Edwards J, Jernigan J, Kleinbaum D. Modelling interrupted time series to evaluate prevention and control of infection in healthcare. Epidemiol Infect 2012;140(12):2131—41.

[5] 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.

[6] Wagner AK, Soumeria SB, Zhang F, Ross-Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. J Clin Pharm Ther 2002;27(4):299—309.

[7] Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. Acad Pediatr 2013;13(6 Suppl):S38—44.

[8] Zhang F, Wagner AK, Ross-Degnan D. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. J Clin Epidemiol 2011;64:1252—61.

[9] Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Cheng AC, 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.

[10] Reeves BC, Deeks JJ, Higgins JPT, Wells GA. Chapter 13: Including non-randomized studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration; 2011. Available at: www.handbook.cochrane.org.

[11] Reeves BC, Wells GA, Waddington H. Quasi-experimental study designs series—paper 5: a checklist for classifying studies evaluating the effects on health interventions—a taxonomy without labels. J Clin Epidemiol 2017;89:30—42.

[12] Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. 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.

[13] Ewusie JE, Blondal E, Soobiah C, Beyene J, Thabane L, Straus SE, et al. Methods, applications, interpretations and challenges of interrupted time series (ITS) data: protocol for a scoping review. BMJ Open 2017;7(6):e016018.

[14] Cheng J, Benassi P, de Oliveira C, Zaheer J, Collins M, Kurdyak P. Impact of a mass media mental health campaign on psychiatric emergency department visits. Can J Public Health 2016;107(3): e303—11.

[15] Baker JM, Alonso WJ. Rotavirus vaccination takes seasonal signature of childhood diarrhea back to pre-sanitation era in Brazil. J Infect 2018;76:68—77.

[16] Milojevic A, Armstrong B, Hashizume M, McAllister K, Faruque A, Yunus M, et al. Health effects of flooding in rural Bangladesh. Epidemiology 2012;23(1):107—15.

[17] StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC; 2017.

[18] Hawton K, Bergen H, Geulayov G, Waters K, Ness J, Cooper J, et al. Impact of the recent recession on self-harm: longitudinal ecological and patient-level investigation from the Multicentre Study of Self-harm in England. J Affect Disord 2016;191:132—8.

[19] Scherb HH, Dipl-Math RN, Mori K, Hayashi K. Increases in perinatal mortality in prefectures contaminated by the Fukushima nuclear power plant accident in Japan: a spatially stratified longitudinal study. Medicine 2016;95(38):e4958.

[20] Kim JS, Chung YK, Lee SS, Kim HS, Park EY, Shin KS, et al. Effect of daily chlorhexidine bathing on the acquisition of methicillin-resistant Staphylococcus aureus in a medical intensive care unit with methicillin-resistant S aureus endemicity. Am J Infect Control 2016; 44(12):1520—5.

[21] Lopez Bernal J, Cummins S, Gasparrini A. The use of controls in interrupted time series studies of public health interventions. Int J Epidemiol 2018;47:2082—93.

[22] Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012;345: e5661.

[23] Lopez-Bernal J. Framework for Enhanced Reporting of Interrupted Time Series (FERITS) 2018: [updated 22 May 2018. Reporting guidelines under development for observational studies]. Available at: http://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-observational-studies/#92. Accessed February 13, 2020.

[24] Hudson J, Fielding S, Ramsay CR. Methodology and reporting characteristics of studies using interrupted time series design in healthcare. BMC Med Res Methodol 2019;19:137.

Chapter 4. Creating effective interrupted time series graphs: review and recommendations

In this chapter the properties of plots presented in ITS studies were examined and recommendations for graphing ITS data were provided. Data from ITS studies are particularly amenable to visual display and, when clearly depicted, can readily show the short- and long- term impact of the interruption. In addition, the importance of the construction of the graphs extends beyond the individual study, since well-constructed graphs can facilitate the inclusion of the study in systematic reviews and meta-analyses.

This research stemmed from the observation that there was a wide variety of graphical displays being used in the ITS studies included in the review (Chapter 3) and that many did not clearly depict the results (e.g. level change). Further, digital extraction of the data from the graphs for the research presented in Chapter 6, highlighted issues in the plotting of the data that compromised the digital extraction for some graphs.

In Chapter 4, graphing recommendations for ITS graphs are suggested, based on seminal visualisation resources. The quality of the graphs from the ITS studies included in the review (Chapter 3) are formally evaluated by comparing their features against the recommendations. Applications of the recommendations are applied to two published graphs, to demonstrate how improvements can be made. Computer code is provided (using the statistical package Stata 15.0 (56)), to allow users to easily implement the recommendations in their own work.

Chapter 4 is presented as a manuscript, published in *Research Synthesis Methods* in July 2020 (54). The computer code for generating the demonstration data sets and creating the figures within the manuscript, referred to as `Supporting File 1' in the manuscript, is attached as Appendix E in this thesis.

Received: 28 January 2020 | Revised: 18 May 2020 | Accepted: 9 July 2020

DOI: 10.1002/jrsm.1435

SPECIAL ISSUE PAPER

Research Synthesis Methods WILEY

Creating effective interrupted time series graphs: Review and recommendations

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

[1]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia

[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[3]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada

[4]Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada

[5]Infection Prevention and Healthcare Epidemiology Unit, Alfred Health, Melbourne, Victoria, Australia

[6]Faculty of Medicine and Health, School of Pharmacy and Charles Perkins Centre, The University of Sydney, Sydney, New South Wales, Australia

Correspondence
Joanne E. McKenzie, School of Public Health and Preventive Medicine, Monash University, Level 4, 553 St. Kilda Road, Melbourne, VIC 3004, Australia.
Email: joanne.mckenzie@monash.edu

Funding information
Australian Government Research Training Program, Grant/Award Numbers: PhD Scholarship (Elizabeth Korevaar), PhD Scholarship (Simon Turner); Canadian Institute of Health Research Foundation Grant, Grant/Award Number: 143269; National Health and Medical Research Council, Grant/Award Numbers: 1068732, 1143429, 1145273

Introduction: Interrupted Time Series (ITS) studies may be used to assess the impact of an interruption, such as an intervention or exposure. The data from such studies are particularly amenable to visual display and, when clearly depicted, can readily show the short- and long-term impact of an interruption. Further, well-constructed graphs allow data to be extracted using digitizing software, which can facilitate their inclusion in systematic reviews and meta-analyses.

Aim: We provide recommendations for graphing ITS data, examine the properties of plots presented in ITS studies, and provide examples employing our recommendations.

Methods and results: Graphing recommendations from seminal data visualization resources were adapted for use with ITS studies. The adapted recommendations cover plotting of data points, trend lines, interruptions, additional lines and general graph components. We assessed whether 217 graphs from recently published (2013-2017) ITS studies met our recommendations and found that 130 graphs (60%) had clearly distinct data points, 100 (46%) had trend lines, and 161 (74%) had a clearly defined interruption. Accurate data extraction (requiring distinct points that align with axis tick marks and labels that allow the points to be interpreted) was possible in only 72 (33%) graphs.

Conclusion: We found that many ITS graphs did not meet our recommendations and could be improved with simple changes. Our proposed recommendations aim to achieve greater standardization and improvement in the display of ITS data, and facilitate re-use of the data in systematic reviews and meta-analyses.

KEYWORDS

data visualization, display of data, graph, interrupted time series, meta-analysis, systematic review

1 | INTRODUCTION

Interrupted time series (ITS) studies are a common design used in areas such as public health, health policy and health services research to examine the effects of an interruption on an outcome. The interruption could be planned, such as the roll out of a new health policy, or unplanned, such as an unintended environmental exposure. This makes ITS studies a valuable design for inclusion in systematic reviews intended to inform policy decisions. However, effective and accurate presentation of the data from ITS studies is needed to enable their inclusion in systematic reviews (including meta-analysis) and to aid interpretation of the results from the review.

The ITS design inherently lends itself to a visual display. In an ITS study, data on a group of individuals (eg, hospital, country) are collected at multiple time points both before and after the interruption. By modeling data from the pre-interruption period, the underlying secular trend can be established and extrapolated to the post-interruption period, creating a counterfactual for what would have occurred in the absence of the interruption. Statistical comparisons between the counterfactual and observed data at different points post interruption can be used to estimate the short- and long-term effects of the interruption. These features can be visually displayed, and in well-designed graphs, the impacts of the interruption on the outcome will likely be evident. For these reasons, visual displays of ITS data in both primary studies, and systematic reviews of ITS studies, are a valuable part of reporting.

A further benefit of visually displaying data from an ITS study is that it allows systematic reviewers to extract the data (eg, using digitizing software) and undertake a re-analysis. A re-analysis may be required in the circumstance where effect estimates have been incompletely reported (ie, when an effect estimate is reported without a measure of precision[1]); where the data has been incorrectly analyzed (eg, when there has been no adjustment for autocorrelation—a common complication with time series data[2-4]); or, when the reviewer wishes to estimate the effect of the interruption using a different effect measure to that reported in the paper (eg, an estimate of level change may be reported but a reviewer may be interested in a difference in slopes or the combined effect of level and slope change). Such re-analyses are particularly important when reviewers wish to meta-analyze the estimated effects, where consistency in the effect measures across studies, and estimates and their standard errors from correct analyses of time series data are required.

The aim of this research is to provide recommendations for the accurate display of ITS data in primary ITS studies and systematic reviews of ITS studies. We first provide principles and recommendations for graphing ITS data (Section 2). We then examine how often graphs are used to display data in reports of ITS studies and the extent to which those graphs meet our recommendations (Section 3). Finally, we present examples of ITS graphs before and after incorporating our recommendations (Section 4).

2 | RECOMMENDATIONS FOR GRAPHING INTERRUPTED TIME SERIES DATA

In this section we suggest recommendations along with the rationale for the visual display of results from ITS studies (Table 1). These have been informed from seminal data visualization resources, including Boers,[5] Cleveland,[6] Few,[7] Lane and Sándor,[9] Tufte[8,10,11] and Yau.[12] In forming these recommendations, we first articulated the primary purpose of an ITS graph, which we consider to be the presentation of an accurate visual depiction of the ITS data, including display of essential details for accurate extraction of the data points. Considering our purpose, the following components should be displayed:

1. the data points;
2. the interruption time;
3. pre-interruption and post-interruption trend lines; and
4. the counterfactual trend line.

Plotting other features of ITS analyses (eg, seasonality, confidence intervals) and features common to many graphs (eg, axis labels, additional text, legends) also need to be considered. The recommendations have been divided into those which are "core" for readers to be able to correctly interpret and extract data from the graphs; and those which are "additional," where applying the recommendation is not required, but may aid in interpretation and enhance visual appeal of the graph. Our terminology is described in Figure A1 of Appendix. Code for creating the example graphs using Stata[13] is available in Supporting Information S1.

TABLE 1 Recommendations for graphing interrupted time series data

Characteristic	Recommendation	Importance	References
Data Points	Plot data points. Each point should be clearly visible	Core	Boers,[5] Cleveland,[6] Few,[7] Tufte[8]
	Use the same data points as were used in the analysis	Core	Lane,[9] Tufte[10,11]
	Line up data points with x-axis tick marks	Core	Few,[7] Tufte,[8] Yau[12]
	Do not join data points with lines	Additional	Few,[7] Tufte[8]
Interruption	Indicate interruption time with vertical line or shading	Core	Cleveland,[6] Yau[12]
	Show any transition or roll-out period with vertical lines or shading	Core	Cleveland,[6] Lane,[9] Tufte[11]
	Label the interruption line	Core	Tufte[8,11]
Trend lines	Plot the fitted pre- and post-interruption trends	Core	Cleveland,[6] Few,[7] Lane,[9] Yau[12]
	Use bold and solid lines for fitted trends	Additional	Few,[7] Yau[12]
	Match trend line with data point color	Additional	Boers,[5] Cleveland,[6] Few,[7] Tufte,[10,11] Yau[12]
Counterfactual	Indicate counterfactual with a trend line	Core	Cleveland,[6] Tufte,[10,11] Yau[12]
	Use a different line pattern for the counterfactual trend as compared with the fitted pre- and post-interruption trend lines	Additional	Tufte,[11] Yau[12]
	Match counterfactual trend line with data point color	Additional	Boers,[5] Cleveland,[6] Few,[7] Tufte,[10,11] Yau[12]
Additional lines	Use a different color and symbol for each additional series	Core	Cleveland,[6] Few,[7] Tufte[11]
	Use neutral colors for additional lines (eg, confidence intervals)	Additional	Boers,[5] Cleveland,[6] Few[7]
General graph components	Show axis tick marks	Core	Tufte,[10,11] Yau[12]
	Label axes	Core	Tufte,[10,11] Yau[12]
	Align axis labels with axis tick marks	Core	Yau[12]
	Include axis titles in which the variables and units of measurement are clear	Core	Tufte,[11] Yau[12]
	Explain all elements of the graph in the figure and/or caption	Core	Boers,[5] Cleveland,[6] Few,[7] Lane,[9] Tufte[8,11]
	If used, grid lines should be faint	Additional	Boers,[5] Few,[7] Tufte[8]
	Use the smallest scale that incorporates all data	Additional	Boers,[5] Cleveland,[6] Lane,[9] Tufte[8]
	Minimize the visual impact of additional text (eg, legends and keys)	Additional	Boers,[5] Few,[7] Tufte[10,11]
	Use horizontal text whenever possible	Additional	Yau[12]
	Use color blind friendly colors	Additional	Boers,[5] Tufte,[8] Yau[12]

2.1 | Data points

A core recommendation for data points is to plot all the raw data points, since they form the basis of the analysis and allow examination of variation in the data along with other distributional information[5-8](Table 1). Further, plotting of points allows readers to extract the data, which is particularly important for systematic reviewers. The symbols chosen for plotting the data points should be clearly visible in the presence of other lines (such as trend lines) and we recommend including a center point (eg, x or +), since this facilitates accurate data extraction. Consideration should also be given to select colors from a color-blind-friendly palette.[5,8,12]

A further core recommendation is to reflect the same data as included in the analysis (eg, if data were analyzed at monthly intervals, monthly data points should be shown, rather than aggregating to a different time period for the graph).[9-11] For data extraction to be accurate, the data points should align with the x-axis tick-marks, hence this is a core recommendation.[7,8,12] An additional recommendation is to *not* join consecutive data points by lines. The addition of such lines can obfuscate data points, does not add relevant information, and introduces visual clutter that can obscure more important elements such as the trend lines.[7,8]

2.2 | Interruption

We suggest three core recommendations for presenting the interruption (Table 1). We recommend that each interruption of the time series be shown with either a vertical line at the point of the interruption (Figure 1A) or by a light shading of a time period (Figure 1B). Plotting this line allows a reader to readily visualize the change in level at the time of the interruption.[6,12] A single line is more appropriate when the interruption is a singular event that occurs at a given time (eg, an earthquake). Shading can be useful to indicate that the interruption is taking place over the indicated time period (eg, a policy implementation) or to indicate sections that are used (or not used) in the statistical analysis (eg, a transition period). Note however, shading can add visual clutter so using light, neutral colors is recommended to avoid this.[6,9,11] As interruptions are key components of an ITS design, labels should also be added to clearly indicate what each line (or area) represents, though care should be taken so that they do not obscure data points.[8,11] Further, it is important that the labels used in the graph match the terminology used in the manuscript.

2.3 | Trend lines

As a core recommendation, we suggest including lines to show the fitted trends, since trend lines allow the reader to readily visualize changes (or lack thereof) in the level and slope[6,7,9,12] (Figures 1 and 2A). Presentation of such lines allows readers to assess goodness of fit and whether an appropriate model has been fitted (eg, when straight lines are fitted but the data indicates a curvilinear line may be preferable).[6,7] Additional recommendations are to use bold and solid lines for the fitted trends, and to match the color of the pre- and post-interruption trend lines with the matching data points. Use of bold and solid lines focuses attention on the trend lines[6,9,11] (Figure 1). Use of the same colors for trend lines and data points visually indicates that they represent trends of the same data series, and the reader is likely to subconsciously make this association quickly.[5-7,10-12]

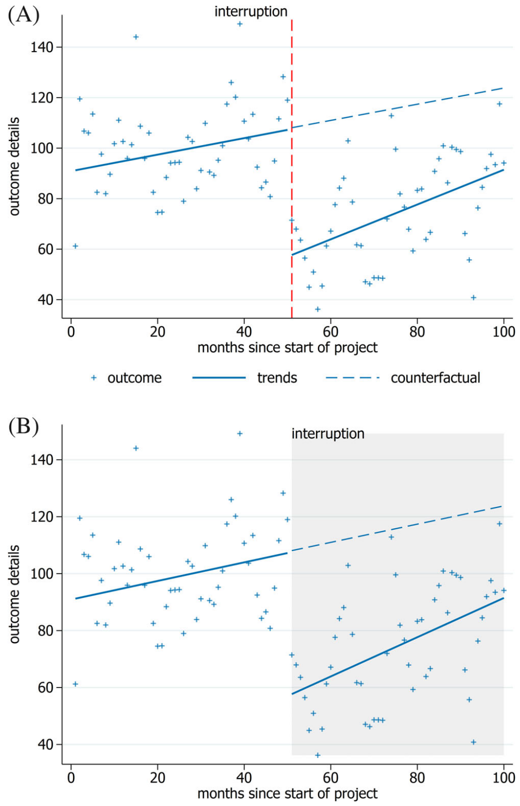


FIGURE 1 Interrupted time series graphs showing key components. The data points are plotted, and the combination of trend lines and counterfactual allow visualization of the change in level from counterfactual to post-intervention trend as well as the change in slope. The timing of the interruption is indicated by a vertical line in, A, and a shaded area in, B. The legend is included in the first example (A) but could easily be removed to increase the size of the graph if appropriate text was used in the caption of the graph (B) [Colour figure can be viewed at wileyonlinelibrary.com]

2.4 | Counterfactual trend line

As a core recommendation, we suggest plotting the counterfactual trend line to allow a reader to easily make a visual assessment of any change between it and the post-interruption trend line[6,10-12] (Figure 1). The counterfactual line is formed by continuing the trend estimated from the pre-interruption time period into the post-

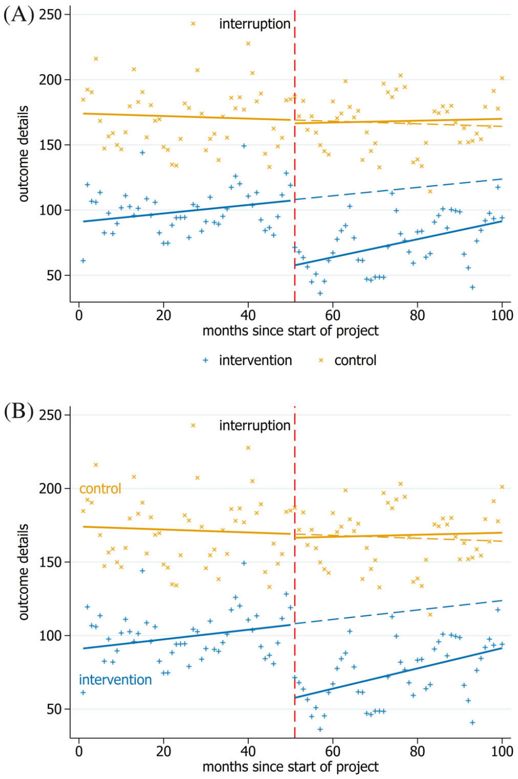


FIGURE 2 Multiple interrupted time series in a single graph. The different series can be distinguished by different colors, symbols, or ideally both, A. Care must be given to not overly clutter the graph. The legend could be omitted with extra labels on the graph, as in, B, or with text in the figure caption [Colour figure can be viewed at wileyonlinelibrary.com]

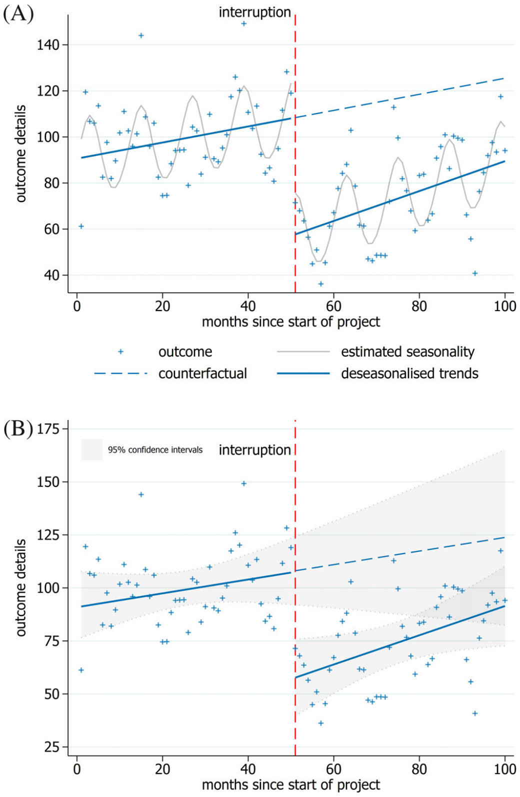


FIGURE 3 Interrupted time series graphs with additional lines. Neutral colors should be used, such as gray for the seasonality curves, A, and confidence intervals for the modeled trends, B. Note that with appropriate text in the figure caption, the legend can be removed, placing greater emphasis on the data points, B [Colour figure can be viewed at wileyonlinelibrary.com]

interruption period. The difference in the lines indicates changes that may be due to the interruption and can be seen over the entire post-interruption period, allowing assessment of both short- and long-term effects. Additional recommendations are to plot the counterfactual trend line using the same color as the data point color, and to use a different line pattern (eg, dashed) as compared with the pre-interruption trend. The latter makes it intuitively clear that the counterfactual is an extension of the pre-interruption trend and is not based on observed data in that time period[5-7,10-12] (Figure 1).

2.5 | Additional lines

Additional lines may be considered in some circumstances to reflect aspects of the analysis or provide additional information. For example, in ITS studies with multiple series (eg, those with a control series), it can be helpful to plot all the series in one graph (Figure 2). Additionally, periodic trends in the data (such as seasonality) may be modeled in the analysis and these can be shown on the graph (Figure 3A). Finally, another potentially useful addition is the inclusion of confidence limits for the modeled trends (Figure 3B).

When plotting multiple series, it is important to distinguish the series. Therefore, as a core recommendation, we suggest that this be done by using a different symbol and color for each data series[6,7,11](Figure 2). Symbols should be chosen so that they are easy to distinguish, this is helpful particularly when the two series overlap and for readers wishing to extract data.[6] If multiple series are difficult to distinguish when plotted on the same graph, a side-by-side presentation could be useful. Plotting both graphs on the same y-axis scale will aid comparison.

Inclusion of additional information on a graph needs to be undertaken carefully so as not to obscure the essential components of the graph or add clutter.[5-7] For example, a seasonal trend could be added using a neutral color that does not inhibit view of the trend lines (Figure 3A). Light gray is recommended, since it does not immediately attract a reader's attention, thus allowing the more important elements to dominate.[5-7] Similarly, confidence limits could be presented using dashed lines or with light shading in a neutral color; however, care should be taken so that shading does not obscure the data points or the interruption period shading (when used) (Figure 3B).

2.6 | General graph components

There are many other components of a graph (eg, axes, grid lines, legends and other additional text) that can enable the reader to more easily interpret the data. Here we focus on the components that are likely to be most helpful for aiding interpretation and those which facilitate accurate extraction of data from the graph. Core recommendations are that axis tick marks are shown and labeled, with labels aligned to the tick marks, since these are essential for accurate extraction of data points from the graph[10-12] (Figure 4). Additionally, grid lines can be used to show elements of the data distribution or analysis (eg, minimum and maximum values, points of the level change; Figure 4A). When included, it is helpful for grid lines to be faint so that they do not detract from the rest of the graph.[5,7,8]

An additional recommendation is to use the smallest scale that still incorporates all of the data, since this maximizes the space for plotting the points, which adds clarity and facilitates data extraction[5,6,8,9] (see the comparison between Figure 4A and Figure 4B). For time series data there is no requirement to include zero on the vertical scale.[5,6,9]

Text can be added to graphs to provide additional information. As a core recommendation, we suggest that detailed titles should be included with the axes to ensure that the variables plotted and their units of measurement are known. Additionally, axis label text is generally easier to read if horizontal[12] (eg, compare the horizontal axis labels in Figure 4A to the vertical axis labels in Figure 4B). Additional text in the form of a legend can be helpful but is not required when the information is included in the figure caption or on the graph itself[7,9] (eg, labeling the intervention and control series in Figure 2B).

Another example of additional text is the inclusion of effect estimates on the graph (Figure 4A). This enables readers to both visualize the level and slope change as well as quantify the values. If not included on the graph, this information could be placed in the caption ensuring that the results from the analysis are readily available.[5-9,11] Reporting of effect estimates should be accompanied by a measure of precision, such as a confidence interval, in order to show the uncertainty.[11]

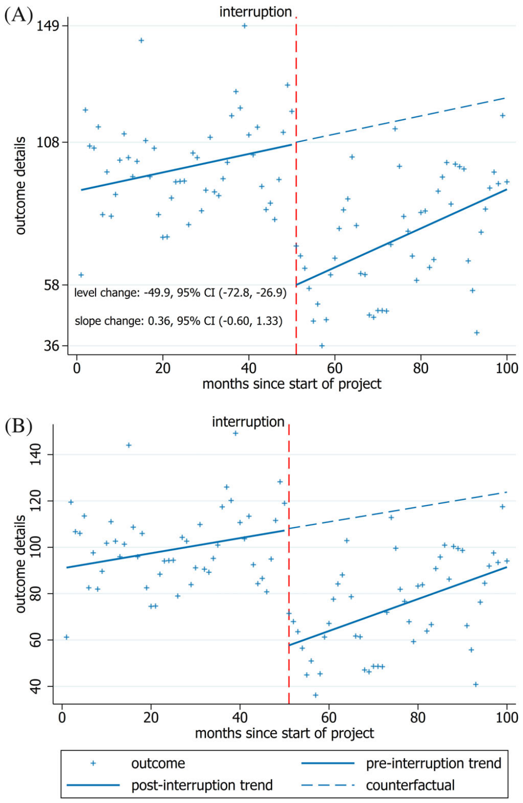


FIGURE 4 Interrupted time series graphs with additional text. By removing the legend and restricting the scale of the y-axis to the range of data, the space available to plot the data is maximized, A. Including text showing the level and slope change with confidence intervals in (A) enables the reader to quickly access the analysis results. The horizontal grid lines and y-axis labels in (A) have been shown at the minimum, maximum and level change points. Leaving the box around the legend in, B, unnecessarily clutters the image. Rotating the text for the y-axis labels from vertical (B) to horizontal (A) also aids in interpretation without unduly reducing the space available to plot the data [Colour figure can be viewed at wileyonlinelibrary.com]

A risk of additional text is that it can clutter the graph, obscure data points and remove attention from the plotted data. Removing extraneous components such as boxes surrounding additional text (eg, legend boxes), using a minimal number of words and using a small font can reduce this risk.[5,6,11] For example, in Figure 4B the legend takes up too much space and the box around it draws attention away from the plotted data.

3 | REVIEW OF INTERRUPTED TIME SERIES GRAPHS

3.1 | Methods

In a previous study we undertook a review examining the design characteristics and statistical methods used in ITS studies evaluating public health interventions. Details of our methods are outlined in our protocol.[14] In brief, we identified 200 ITS studies examining the impact of a public health intervention or exposure that has public health implications published between 2013 and 2017 and indexed on PubMed. For each ITS study, multiple outcomes were potentially eligible for inclusion. Within each outcome type category (binary, continuous, count) a prespecified hierarchy of decision rules was used to select one outcome (see Turner et al[14] for further details). We extracted data on the study characteristics, statistical models, estimation methods, effect measures and parameter estimates, and whether a graph was included.

For the present paper, we assessed the quality of the graphs included in our review by examining whether the graphs met the core recommendations outlined in Table 1. Two reviewers undertook this assessment independently for all graphs (SLT and one of AK, EK or JEM). Frequencies and percentages of graphs meeting each of the core recommendations are tabulated.

3.2 | Results

Our review of 200 ITS studies included 230 outcomes, for which there were 217 associated ITS graphs (Table 2). Data points were plotted in 60% (130/217) of graphs. Points were joined by lines in 64% (138/217) of the graphs, against our recommendation. Tick marks were provided on the x-axis in most graphs (90%, 195/217). However, when tick marks were provided, the data points were aligned in only 56% (109/195) of the graphs. Joint criteria for accurate data extraction require distinct points that align with axis tick marks and labels that allow the points to be interpreted; these criteria were jointly met in only 33% (72/217) of graphs.

A distinct line was used to indicate the interruption in 73% (158/217) of the graphs. Lines were used to indicate the trends in each segment in under half of graphs (48%, 103/217), and the counterfactual was rarely plotted (17%, 37/217). The segments or interruption line were labeled in 60% of graphs (130/217). When additional series were plotted, they met our recommendation of

TABLE 2 Prevalence of adherence to core graphing recommendations, N = 217

Graph feature	n	%
Data points		
Core:		
Distinct individual points plotted	130	60
Tick marks on the x-axis	195	90
Data points align with tick marks on the x-axis (n = 195)	109	56
x-axis labels aligned with tick marks (n = 195)	140	72
x-axis units of measurement clear[a]	211	97
Tick marks on the y-axis	215	99
y-axis labels aligned with tick marks (n = 215)	215	100
y-axis units of measurement clear[a]	148	68
Accurate data extraction possible[b]	72	33
Additional:		
Data points not joined by lines	79	36
Lines		
Core:		
There was a line representing the time of the interruption	158	73
There were labels for the segments or interruption line on the graph	130	60
There were lines for the trends in each segment	103	48
There was a line for the counterfactual	37	17
Additional:		
Multiple series were plotted	91	42
How multiple series were plotted (N = 91)		
Different color and/or pattern and different symbol	32	35
Different color and/or pattern same symbol	44	48
Same color and/or pattern different symbol	14	15
Same color and/or pattern same symbol	1	1
Other model-based lines were plotted, for example, Seasonality curves	21	10

[a] A graph was assessed as having a clear unit of measurement if the axis title made the unit of measurement clear, or the unit of measurement was clear from the axis labels (eg, dates).

[b] Joint criteria for accurate data extraction requires distinct points that align with axis tick marks and labels that allow the points to be interpreted.

using both a different color and symbol in 35% of the graphs (32/91). Other model-based lines such as seasonality curves or confidence intervals were presented in 10% of the graphs (21/217).

4 | APPLICATION OF GRAPHING RECOMMENDATIONS

In the following examples we apply our graphing recommendations to two graphs from ITS studies. These examples were identified from our review of ITS graphs (Section 3), and were selected for their potential to demonstrate many of the graphing recommendations.[15,16] Datasets were extracted from published graphs using WebPlotDigitizer which has been found to be a valid tool for data extraction.[17,18] Note that while our analytical methods may differ to those applied in the original publications, this is inconsequential for our purpose of demonstrating the impact of applying the graphing recommendations.

4.1 | Example 1

An ITS study examined the impact of an amendment to self-defense laws on homicide rates in Florida, United States of America.[15] The amendments provided legal immunity to individuals that used lethal force in self-defense situations. Four states (New York, New Jersey, Ohio, and Virginia), where the amendment to the law did not occur, were used to form a single control series. Data on the homicide rate per 100 000 people were collected from 1999 to 2015 and aggregated at monthly time intervals. A main finding of the analysis was that there was an increase in homicide rates in Florida due to the amendment.

The graph in the published paper depicts the data points, trends and seasonality for the intervention and control series (Figure 5A). The graph includes a legend to denote the series. Tick marks are included with labeled axes and the intervention time period is represented with shading.

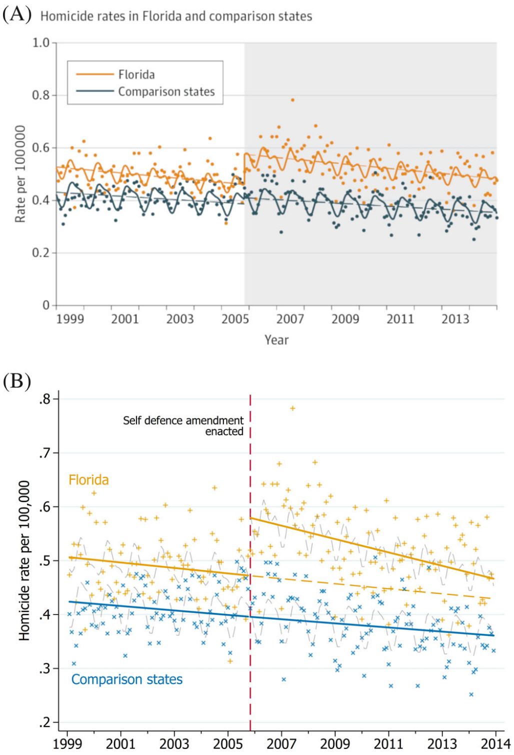


FIGURE 5 A, Published version of the graph. Reproduced with permission from JAMA Internal Medicine 2017;177(1):44-50. doi: 10.1001/jamainternmed.2016.6811. Copyright©(2017) American Medical Association. All rights reserved. B, Revised graph using the proposed graphing recommendations [Colour figure can be viewed at wileyonlinelibrary.com]

Using our recommended guidelines would lead to a number of changes (Figure 5B). Data points have been plotted using different colors and symbols for the two series as this aids data extraction. The intervention period has been labeled and the shading has been removed to lessen the visual clutter; however, we recognize that the approach taken to identifying the interruption is one of choice and that some authors or journals may have preference for shading. Bold and solid lines have been used to prominently display the trends in the pre and post interruption periods. The counterfactual lines have been added as dashed lines to allow visual assessment of the level and slope changes over time. Seasonal effects are still shown, though they are now plotted in light gray dashed lines so as not to obscure the more important data points and trend lines. The x-axis has been adjusted so that the points are more clearly aligned with the tick marks to facilitate data extraction. To free up more space in the graph, the x-axis title has been removed, since it is clear from the x-axis label what the unit of measurement is (ie, years). For a similar reason, the range of the y-axis has been changed from 0 to 1 to 0.2 to 0.8. An increase in the amount of space available for the data will, again, facilitate data extraction. The word “homicide” has been included in the y-axis title to provide a more detailed description of the outcome. Finally, the legend has been removed, but information from the legend has been incorporated into the graph by labeling the series with matching colored text.

4.2 | Example 2

An ITS study examined the impact of a physician call system change on hospital readmissions in Sunnybrook Health Sciences Centre in Toronto, Canada.[16] The intervention involved changing how hospital admissions were distributed to inpatient physician teams. Before the system change admissions were concentrated to a single team during a given 24-hour period, after the change, admissions were distributed over all teams each day. Data on the proportion of 4-week readmissions were collected from 2004 to 2013 and aggregated to monthly time intervals. A main finding of the analysis showed that there was an increase in the proportion of readmissions following the call system change.

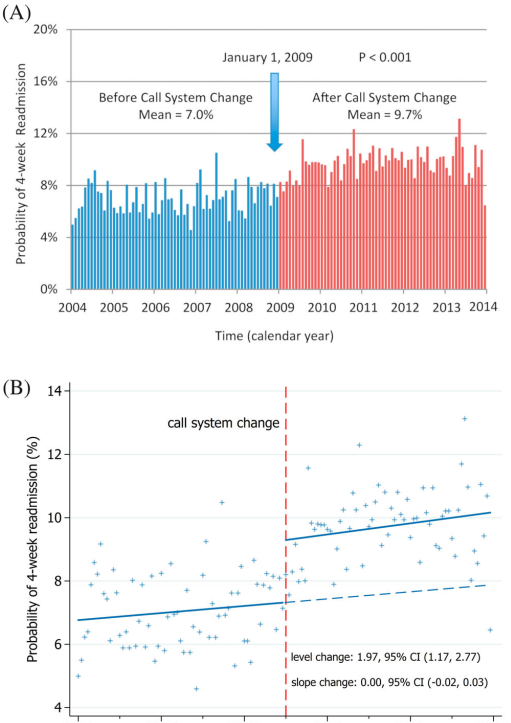


FIGURE 6 A, Published version of the graph. Reprinted from The American Journal of Medicine, 2016;129(7):706-14.e2, Hospital Readmissions Following Physician Call System Change: A Comparison of Concentrated and Distributed Schedules, Copyright©(2016) with permission from Elsevier. B, Revised graph using the proposed graphing recommendations [Colour figure can be viewed at wileyonlinelibrary.com]

The graph in the published paper uses vertical bars to indicate the monthly proportions (Figure 6A). Colored bars are used to represent the period before (blue) and after (red) the call system change. Text is used to denote the time of the call system change as well as the mean values of the monthly proportions in each segment (pre- and post-interruption) and a *P*-value for the difference in these mean pre- and post-interruption proportions. Tick marks are included with labeled axes.

Using our recommended guidelines would lead to a number of changes (Figure 6B). The data points have been plotted, which allows the spread of the data to be more easily seen, allows the data to be extracted and reduces the visual clutter. The interruption has been represented by a vertical line which is labeled. The counterfactual and trend lines have been plotted, allowing the reader to more easily see the impact of the intervention. The x-axis has been adjusted so that the points are more clearly aligned with the tick marks to facilitate data extraction. The range of the y-axis has been decreased to allow the data to fill the available space. Using additional text, the level and slope changes are given, along with 95% confidence intervals.

5 | DISCUSSION

In this study we have proposed a set of recommendations for graphing ITS data. Using a sample of graphs identified from a review of ITS studies,[1] we assessed whether the graphs met our recommended reporting items. We found that, in general, ITS graphs did not meet our core recommendations and could be improved. We demonstrated the impact of applying the recommendations to two examples.

These recommendations are of relevance to authors of ITS studies and systematic reviewers. Improvement in visual representation of ITS data in primary studies will not only facilitate better understanding of the data for readers but will allow a greater number of ITS studies to contribute to meta-analyses. Further, even in systematic reviews where meta-analysis of ITS data may not be possible, if data can be extracted from the graphs in the primary studies, and re-graphed in a standardized format (through use of our recommendations) in the systematic review, this may aid interpretation. Re-graphing the data in a standardized way overcomes the many varied visual displays of ITS data that are found in the literature.

Our recommendations were underpinned by a twofold purpose: (a) to provide an accurate visual description of the ITS data and, (b) to display essential details for accurate extraction of the data points. For many study designs, the connection between these purposes may not seem apparent. For example, in most randomized trials (except perhaps for small crossover trials), the individual observations are not plotted, and nor would it be a sensible suggestion to extract data from such graphs. In the case of ITS studies, however, most publications include a graph,[1] the series typically do not include a large number of data points to extract,[1] and extraction enables re-analysis of the data that may overcome some of the shortcomings that are commonly observed in the analysis of ITS data (eg, reporting of incomplete effect estimates, no adjustment for autocorrelation[1,2,19,20]). Further, while the purposes differ, they lead to consistent recommendations. For example, the recommendation to plot data points, while necessary for data extraction, is also a recommendation for good data visualisation.[5-8]

General recommendations for good graphing proposed in the visual display literature informed our recommendations (Boers,[5] Cleveland,[6] Few,[7] Lane,[9] Tufte[8,10,11] and Yau[12]). We considered how these recommendations should be modified for graphing data from ITS studies. While our approach to developing the recommendations was grounded in the visual display literature, other approaches (such as consensus-based methods) could equally have been adopted, and may have resulted in different recommendations or a different emphasis on which items were recommended as core or additional. Future research that examines readers' understanding of alternative displays of the graph components will be helpful for refining the recommendations (eg, whether the use of connecting lines between points facilitates visual detection of autocorrelation).

Finally, our recommendations have been developed primarily considering the graphing of ITS data where segmented linear regression models are fitted. While it is likely that most recommendations will apply when other models and statistical methods are adopted (eg, plot data points, indicate interruption time), some recommendations may not apply or need to be adapted. For example, if a forecasting method is used,[21] a post-interruption trend may not be estimated, and therefore the recommendation to plot this trend line does not apply.

6 | CONCLUSION

Graphs are useful for visually displaying the data and results from ITS studies. Well-designed graphs allow accurate data extraction, and therefore re-use of the data in systematic reviews and meta-analyses. The proposed set of recommendations for graphing ITS studies aims to achieve greater standardization and improvement in the visual presentation of ITS data.

ACKNOWLEDGEMENTS

This work forms part of S.L.T.'s PhD, which is supported by an Australian Postgraduate Award administered through Monash University, Australia. This work is supported by the Australian National Health and Medical Research Council (NHMRC) project grant (1145273). S.L.T. and E.K. are supported through an Australian Government Research Training Program (RTP) Scholarship administered by Monash University, Australia. J.E.M. is supported by an NHMRC Career Development Fellowship (1143429). A.C.C. is supported by an NHMRC Career Development Fellowship (1068732). J.M.G. holds a Canada Research Chair in Health Knowledge Uptake and Transfer and a Canadian Institute of Health Research (CIHR) Foundation grant (FDN 143269). The funders had no role in study design, decision to publish, or preparation of the manuscript.

CONFLICT OF INTEREST

The author reported no conflict of interest.

AUTHOR CONTRIBUTIONS

Simon L. Turner conceived the study, reviewed data visualization resources, proposed the first set of recommendations and generated the computer code. Amalia Karahalios, Elizabeth Korevaar, Joanne E. McKenzie and Simon L. Turner extracted the data used in the review. Simon L. Turner wrote the first draft of the manuscript, with contributions from Joanne E. McKenzie. Simon L. Turner, Amalia Karahalios, Andrew B. Forbes, Elizabeth Korevaar, Monica Taljaard, Jeremy M. Grimshaw, Allen C. Cheng, Lisa Bero and Joanne E. McKenzie refined the recommendations, contributed to revisions of the manuscript and take public responsibility for its content.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Simon L. Turner https://orcid.org/0000-0001-9163-4524

REFERENCES

1. 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. https://doi.org/10.1016/j.jclinepi.2020.02.006.
2. Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. 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):613-623. https://doi.org/10.1017/s0266462303000576.

3. 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):2131-2141. https://doi.org/10.1017/s0950268812000179.
4. Huitema BE, McKean JW. Identifying autocorrelation generated by various error processes in interrupted time-series regression designs. *Educ Psychol Meas*. 2007;67(3):447-459. https://doi.org/10.1177/0013164406294774.
5. Boers M. Designing effective graphs to get your message across. *Ann Rheum Dis*. 2018;77(6):833-839. https://doi.org/10.1136/annrheumdis-2018-213396.
6. Cleveland WS. *The Elements of Graphing Data*. Murray Hill, NJ: AT&T Bell Laboratories; 1994.
7. Few S. *Show me the Numbers: Designing Tables and Graphs to Enlighten*. 2nd ed. Burlingame, CA: Analytics Press; 2012.
8. Tufte ER. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press; 2001.
9. Lane DM, Sándor A. Designing better graphs by including distributional information and integrating words, numbers, and images. *Psychol Methods*. 2009;14(3):239-257. https://doi.org/10.1037/a0016620.
10. Tufte ER. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press; 1997.
11. Tufte ER. *Beautiful Evidence*. Cheshire, CT: Graphics Press; 2006.
12. Yau N. *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics*. Indianapolis, IN: Wiley; 2011.
13. StataCorp. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.; 2017.
14. 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. https://doi.org/10.1136/bmjopen-2018-024096.
15. Humphreys DK, Gasparrini A, Wiebe DJ. Evaluating the impact of Florida's "stand your ground" self-defense law on homicide and suicide by firearm: an interrupted time series study. *JAMA Intern Med*. 2017;177(1):44-50. https://doi.org/10.1001/jamainternmed.2016.6811.
16. Yarnell CJ, Shadowitz S, Redelmeier DA. Hospital readmissions following physician call system change: a comparison of concentrated and distributed schedules. *Am J Med*. 2016;129(7):706-14.e2. https://doi.org/10.1016/j.amjmed.2016.02.022.
17. Drevon D, Fursa SR, Malcolm AL. Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. *Behav Modif*. 2017;41(2):323-339. https://doi.org/10.1177/0145445516673998.
18. WebPlotDigitizer [program]. 4.2 version. San Francisco, CA; 2019.
19. 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. https://doi.org/10.1186/s12874-019-0777-x.
20. Jandoc R, Burden AM, Mamdani M, Lévesque LE, Cadarette SM. Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. *J Clin Epidemiol*. 2015;68(8):950-956. https://doi.org/10.1016/j.jclinepi.2014.12.018.
21. Linden A. Using forecast modelling to evaluate treatment effects in single-group interrupted time series analysis. *J Eval Clin Pract*. 2018;24(4):695-700. https://doi.org/10.1111/jep.12946.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Turner SL, Karahalios A, Forbes AB, et al. Creating effective interrupted time series graphs: Review and recommendations. *Res Syn Meth*. 2020;1–12. https://doi.org/10.1002/jrsm.1435

APPENDIX

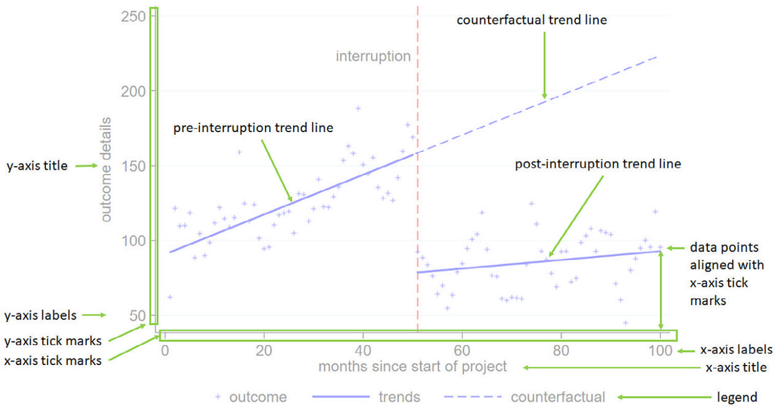


FIGURE A1 terminology used for some components of graphs [Colour figure can be viewed at wileyonlinelibrary.com]

Chapter 5. Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study

This chapter presents the results of a simulation study investigating the performance of a set of statistical methods used in analysing ITS data. Data were simulated under a range of scenarios which included different level and slope changes, varying lengths of time series and magnitudes of autocorrelation. The statistical methods investigated and values for the simulation parameters were informed by the review of ITS studies (Chapter 3).

In Chapter 5 details of the ITS model and estimation methods used in the simulation study are given, along with the simulation study methods. Results of the comparisons between statistical methods are summarised using the performance measures bias, empirical SE, model-based SE, CI coverage and power. Estimates of autocorrelation are also compared, along with the performance of the DW test. The findings led to recommendations for choice of statistical method, and how this choice may differ given the series length.

Chapter 5 is presented as a manuscript, which has been submitted to the *Biometrical Journal* and is available as a pre-print (55). Supporting computer code (using the statistical package Stata 15.0 (56)) required to create and analyse the simulated data sets and produce the graphs used in the manuscript can be found on the online repository figshare: https://doi.org/10.26180/13284329 (57).

Additional files referred to in the manuscript are appended to this thesis as follows:

Location in thesis	Referred to in manuscript	Content of appendix
Appendix F	Supplementary 1.1	Interrupted time series graphs showing different magnitudes of autocorrelation
	Supplementary 1.2	Graphs of slope change estimate distributions
	Supplementary 1.3	Graphs showing the variation of bias, empirical standard error, model based standard error, coverage and autocorrelation estimates for all parameter combinations
	Supplementary 1.4	Graph of standard error of level change estimates from OLS by autocorrelation
	Supplementary 1.5	Graphs of the ratio of empirical and model-based standard errors by series length for slope change
	Supplementary 1.6	Graphs of power for level and slope change
	Supplementary 1.7	Graphs of standard error of autocorrelation estimates
	Supplementary 1.8	Graphs of convergence of estimation methods
	Supplementary 1.9	Graphs of coverage by autocorrelation bias for slope change
Appendix G	Supplementary 2	Computer code to create and analyse simulated data sets

Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study

Simon L Turner[1], Andrew B Forbes[1], Amalia Karahalios[1], Monica Taljaard[2,3], Joanne E McKenzie[1*]

[1]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia.

[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada. 1053 Carling Ave, Ottawa.

[3]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada. 600 Peter Morand Crescent, Ottawa, Ontario K1G 5Z3.

*Corresponding Author:

Joanne McKenzie
mail: Level 4, 553 St. Kilda Road, Melbourne, 3004, Australia
email: joanne.mckenzie@monash.edu
ph: +61 3 9903 0380

Abstract

Interrupted time series (ITS) studies are frequently used to evaluate the effects of population-level interventions or exposures. To our knowledge, no studies have compared the performance of different statistical methods for this design. We simulated data to compare the performance of a set of statistical methods under a range of scenarios which included different level and slope changes, varying lengths of series and magnitudes of autocorrelation. We also examined the performance of the Durbin-Watson (DW) test for detecting autocorrelation. All methods yielded unbiased estimates of the level and slope changes over all scenarios. The magnitude of autocorrelation was underestimated by all methods, however, restricted maximum likelihood (REML) yielded the least biased estimates. Underestimation of autocorrelation led to standard errors that were too small and coverage less than the nominal 95%. All methods performed better with longer time series, except for ordinary least squares (OLS) in the presence of autocorrelation and Newey-West for high values of autocorrelation. The DW test for the presence of autocorrelation performed poorly except for long series and large autocorrelation. From the methods evaluated, OLS was the preferred method in series with fewer than 12 points, while in longer series, REML was preferred. The DW test should not be relied upon to detect autocorrelation, except when the series is long. Care is needed when interpreting results from all methods, given confidence intervals will generally be too narrow. Further research is required to develop better performing methods for ITS, especially for short series.

Keywords

Autocorrelation, Interrupted Time Series, Public Health, Segmented Regression, Statistical Methods, Statistical Simulation

1 Background

Interrupted time series (ITS) studies are frequently used to evaluate the impact of interventions or exposures that occur at a particular point in time (Ramsay, Matowe et al. 2003, Jandoc, Burden et al. 2015, Ewusie, Soobiah et al. 2020, Turner, Karahalios et al. 2020). Although randomised trials are the gold standard study design, randomisation may be infeasible or undesirable in the case of policy evaluation or interventions that are implemented at a population level. Randomization also is not an option for retrospective evaluation of interventions or exposures such as natural disasters or pandemics. The use of an ITS design may be considered in these situations, as they are one of the strongest non-randomised experimental designs (Wagner, Soumerai et al. 2002, Penfold and Zhang 2013, Jandoc, Burden et al. 2015, Kontopantelis, Doran et al. 2015, Lopez Bernal, Cummins et al. 2016, Hudson, Fielding et al. 2019).

In an ITS study, observations are collected at regular time points before and after an interruption, and often analysed in aggregate using a summary statistic (e.g. mean, proportion) within a time interval (e.g. weekly, monthly, or annually). A key feature of the design is that data from the pre-interruption interval can be used to estimate the underlying secular trend. When this trend is modelled correctly, it can be projected into the post-interruption interval, providing a counterfactual for what would have occurred in the absence of the interruption. From this counterfactual, a range of effect measures can be constructed that characterise the impact of the interruption. Two commonly used measures include the ‘change in level’ – which represents the change immediately after the interruption, and the ‘change in slope’ – which represents the difference in trends before and after the interruption.

A key feature of time series data is that there is the potential for non-independence of consecutive data points (serial autocorrelation) (Gebski, Ellingson et al. 2012). In the presence of positive autocorrelation, statistical methods that do not account for this correlation will give spuriously small standard errors (SEs) (Huitema and McKean 2007). Several statistical methods are available to account for autocorrelation, such as Prais-Winsten generalised least squares or the Newey-West correction to SEs, or to directly model the error, such as autoregressive integrated moving averages (ARIMA). Further, several methods are available for testing for the presence of autocorrelation, with the Durbin-Watson test being the most commonly used (Hudson, Fielding et al. 2019, Turner, Karahalios et al. 2020). While the performance of some of these methods has been examined for time series data (Smith and McAleer 1994, Alpargu and Dutilleul 2003), their performance in the context of ITS studies has received relatively little attention.

In this study, we therefore aimed to examine the performance of a range of statistical methods for analysing uncontrolled ITS studies using segmented linear models. We restrict our evaluation to ITS designs where there is a single interruption, with an equal number of time points pre and post interruption, and with first order autoregressive errors. The structure of the paper is as follows: In Section 2, we begin by introducing a motivating example for this research. In Section 3, we describe the statistical model and estimation methods used in our simulation study. In Sections 4 and 5, we present the methods and results from the statistical simulation study. In Section 6, we return to our motivating example and demonstrate the impact of applying the methods outlined in Section 3. Finally, in Section 7 we present key findings and implications for practice.

1.1 Motivating example

Healthcare-associated infections (HAIs) are a common complication affecting patients in hospitals. *Clostridium difficile (C difficile)* infection is an example of one such infection that can cause serious gastrointestinal disease. As such, many countries require mandatory surveillance of *C difficile* infection rates in hospitals. When outbreaks of *C difficile* occur, the cleaning and disinfection regimes in hospitals are often changed in an attempt to reduce the infection rate. The routine collection of data in this context has led to many retrospective investigations of the effects of different interventions (e.g. novel disinfectants) to reduce *C difficile* infection using ITS data (Brennan 2017). Hacek et al (Hacek, Ogle et al. 2010) provides an example of such a study, where they examined the effect of terminal room cleaning with dilute bleach (Figure 1) on the rate of patients (per 1000 patient days) with a positive test for *C difficile*. Data were aggregated at monthly intervals. The series was relatively short – a scenario which is not atypical of these studies – with 10 data points pre and 24 post the intervention (Brennan 2017). In the context of HAIs, there is a tendency for consecutive data points to be more similar to each other, manifesting as ‘clusters’ of data points in time (Figure 1). Fitting a segmented linear regression model to the data shows an apparent immediate decrease in the infection rate (level change), as well as a decrease in the trend (slope change). In the following section, we outline different statistical methods to estimate the model parameters and return to this example in Section 6, where we apply these methods and compare the results.

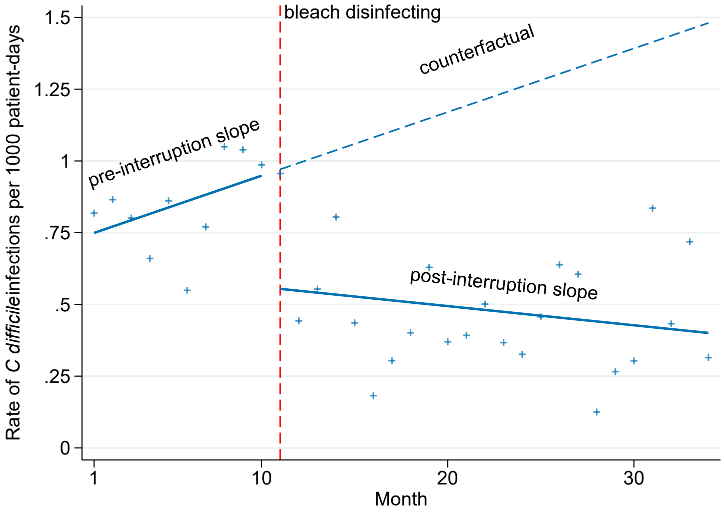


Figure 1: Rate of Clostridium difficile infections (per 1000 patient-days) pre and post bleach disinfection intervention per month.

2 Methods

2.1 Interrupted time series (ITS): model and estimation methods

We begin by describing the statistical model and parameters used in our simulation study followed by a brief description of some common statistical estimation methods and the Durbin-Watson test for autocorrelation.

2.1.1 Statistical model

We use a segmented linear regression model with a single interruption, which can be written using the parameterisation proposed by Huitema and McKean (Huitema 2011) as:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 D_t + \beta_3 [t - T_I] D_t + \varepsilon_t \qquad (1)$$

where Y_t represents the outcome at time point t of N time points. D_t is an indicator variable that represents the post-interruption interval (i.e. $D_t = 1 \ (t \geq T_I)$ where T_I represents the time of the interruption). The model parameters, β_0, β_1, β_2 and β_3 represent the intercept (e.g. baseline rate), slope in the pre-interruption interval, the change in level and the change in slope, respectively. The error term, ε_t, represents deviations from the fitted model, which are constructed as:

$$\varepsilon_t = \rho \varepsilon_{t-1} + w_t \qquad (2)$$

where w_t represents "white noise" that is normally distributed $w_t \sim N(0, \sigma^2)$, and ρ is the lag-1 autocorrelation of the errors which can range from -1 to +1. A lag-1 error means that the influence of errors on the current error is restricted to the value immediately prior. Longer lags are possible but in this paper we confine attention to lag-1 only (AR(1) errors).

2.1.2 Estimation methods

A range of statistical estimation methods are available for estimating the model parameters. These methods account for autocorrelation in different ways and are briefly described below. We focus on statistical methods that have been more commonly used (Ordinary Least Square (OLS), Generalised Least Squares (GLS), Newey-West (NW), Autoregressive Integrated Moving Average (ARIMA))(Jandoc, Burden et al. 2015, Hudson, Fielding et al. 2019, Ewusie, Soobiah et al. 2020, Turner, Karahalios et al. 2020). In addition, we have included Restricted Maximum Likelihood (REML) (with and without the Satterthwaite adjustment), which although is not a method in common use, is included because of its potential for reduced bias in the estimation of the autocorrelation parameter, as has been discussed for general (non-interrupted) time series (Cheang and Reinsel 2000). Further details and equations can be found in Appendix 1.

2.1.2.1 Ordinary Least Squares

Estimates of the regression parameters and their variances from model (1) can be obtained from fitting a segmented linear regression model using OLS (Appendix 1.1). In the presence of autocorrelation, the OLS estimators for the regression parameters are unbiased; however, the SEs will be incorrect (Kutner, Nachtscheim et al. 2008).

2.1.2.2 Newey-West

The NW estimator of the variance of the regression parameters estimated using OLS accommodates autocorrelation and heteroskedasticity of the error terms in the regression model (1) (Newey and West 1987) (Appendix 1.2).

2.1.2.3 Generalised least squares

Two common GLS methods for estimating the regression parameters and their variances are Cochrane-Orcutt (CO) and Prais-Winsten (PW). For both methods, a regression model is first fitted using OLS and an estimate of the autocorrelation is calculated from the residuals. This estimate is then used to transform the data and remove the autocorrelation from the errors, upon which the regression parameters are then estimated from the transformed data. If there is still some residual autocorrelation these steps are iterated until a criterion is met (e.g., the estimated value for autocorrelation has converged (StataCorp 2017)). The CO method applies the transformation from the second observation onwards (t=2, 3, … n). The PW method is a modification to the CO method in which a transformed value is used for the first observation (Appendix 1.3). The PW method is therefore likely to be more efficient in small series since it does not discard the first observation. The sampling properties of the estimators of the regression parameters are likely to be adversely affected when the series length is small due to poor estimation of the autocorrelation.

2.1.2.4 Restricted maximum likelihood

It is well known that maximum likelihood estimators of variance components are biased in small samples due to not accounting for the degrees of freedom (d.f.) used when estimating the fixed effect regression parameters (Singer and Willett 2003). Restricted maximum likelihood is a variant of maximum likelihood estimation and attempts to address the bias by separating the log-likelihood into two terms; one that involves the mean and variance parameters, and one which is only dependent on the variance parameters. By maximising the latter term first with the appropriate number of d.f., an estimate of the variance parameter can be obtained which can be used when maximising the former, thus correctly accounting for the d.f. (Thompson 1962, Cheang and Reinsel 2000).

For small samples, there is greater uncertainty in the estimation of the SE of the regression parameters. To account for this uncertainty in making inferences about the regression parameters, the Satterthwaite adjustment can be used to adjust the t-distribution d.f. used in hypothesis testing and calculation of confidence limits (Satterthwaite 1946).

2.1.2.5 *Autoregressive integrated moving average*

In an ARIMA model, information from past values, including lagged values of the dependent variable and errors, are explicitly modelled. This is achieved by including regression coefficients for these variables in the ARIMA model. The lagged values can be from a range of previous time points, extending beyond lag-1 models. By explicitly modelling the influence of data from previous time points, their impact at subsequent times is quantified and estimates of the magnitude of autocorrelation can be obtained along with regression parameter estimates. Here we consider ARIMA models with only a first order autoregressive term (an ARIMA(1,0,0) model) estimated via maximum likelihood. ARIMA models have been shown to not perform well with fewer than fifty points (Nelson 1998). Further details about the method can be found in Appendix 1.4, Nelson (Nelson 1998) and Box et al (Box 2016).

2.1.3 Durbin-Watson test for autocorrelation

The Durbin-Watson (DW) test is commonly used for detecting autocorrelation in time series. Often, the test is used as part of a two-stage analysis strategy to determine whether to use a method that adjusts for autocorrelation or use OLS (which does not adjust for autocorrelation). The null hypothesis is that there is no autocorrelation ($H_0: \rho = 0$) against the alternative that autocorrelation is present ($H_1: \rho \neq 0$). The DW-statistic can range between zero and four, with values close to two indicating no autocorrelation. The DW-statistic is compared to critical values to determine whether there is evidence of autocorrelation, no autocorrelation, or the test is inconclusive. The critical values differ by series length, significance level and the d.f. in the regression model. Further details are available in Appendix 1.5, Kutner et al (Kutner, Nachtscheim et al. 2008) and Durbin and Watson (Durbin and Watson 1950).

2.2 Simulation study methods

We undertook a numerical simulation study, examining the performance of a set of statistical methods under a range of scenarios which included different level and slope changes, varying lengths of series and magnitudes of autocorrelation. Design parameter values were combined using a fully factorial design with 10,000 data sets generated per combination. A range of criteria were used to evaluate the performance of the statistical methods. We now describe the methods of the simulation study using the ADEMP (defining aims, data-generating mechanisms, estimands, methods and performance measures) structure (Morris, White et al. 2019).

2.2.1 Data Generating Mechanisms

We simulated data from ITS studies by randomly sampling from a parametric model (equation 1), with a single interruption at the midpoint, and first order autoregressive errors (examples shown in Supplementary 1.1). We multiplied the first error term, ε_1, by $\sqrt{\frac{1}{1-\rho^2}}$ so that the variance of the error term was constant at all time points.

We created a range of simulation scenarios including different values of the model parameters and different numbers of data points per series (Table 1). These values were informed by our review of ITS studies (Turner, Karahalios et al. 2020), where we reanalysed available data sets to estimate level and slope changes (standardised by the residual standard deviation), and autocorrelation. We found a median standardised level change of 1.5 (inter-quartile range (IQR): 0.6 to 3.0), n=190), median standardised slope change of 0.13 (IQR: 0.06 to 0.27, n=190) and median autocorrelation 0.2 (IQR: 0 to 0.6, n=180). We therefore constructed models with level changes (β_2) of 0, 0.5, 1 and 2, and slope changes (β_3) of 0 and 0.1. We did not examine negative level or slope changes since we did not expect this to influence the performance metrics. Autocorrelation was varied between 0 and 0.8 in increments of 0.2 to cover the full range of autocorrelations observed in the ITS studies included in the review. The number of data points per series was varied from 6 to 100, equally divided before and after the interruption, informed by the number of data points observed in the ITS studies (median 48, IQR: 30 to 100, n=230). The increment size was varied; initially it was small (i.e. 2) so as to detect changes in the performance metrics that were expected to arise with smaller sample sizes and was increased to 4 and then 20. All combinations of the factors in Table 1 were simulated, leading to 800 different simulation scenarios (Table 1, Figure 2).

Table 1: Simulation parameters

Parameter	Symbol	Parameter Values
Intercept	β_0	0
Pre-interruption slope	β_1	0
Level change	β_2	0, 0.5, 1, 2
Change in slope post-interruption	β_3	0, 0.1
Autocorrelation coefficient	ρ	0, 0.2, 0.4, 0.6, 0.8
Variance of white noise error component	σ^2	1
Number of data points		6, 8, 10, 12, 14, 16, 18, 20 24, 28, 32, 36, 40, 44, 48, 52, 56 60, 80, 100

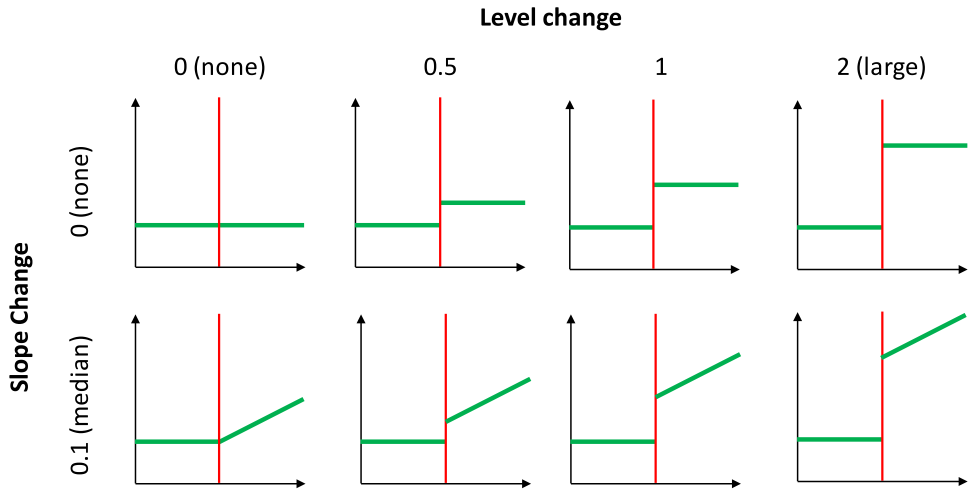


Figure 2: Structure of the eight models constructed from different combinations of the model input parameters (Table 1).

2.2.2 Estimands and other targets

The primary estimands of the simulation study are the parameters of the model, β_2 (level change) and β_3 (slope change) (Equation 1). These were chosen as they are commonly reported effect measures (Hudson, Fielding et al. 2019, Turner, Karahalios et al. 2020). We also examined the autocorrelation coefficient, ρ, and the value of the Durbin Watson statistic.

2.2.3 Statistical Methods to analyse ITS studies

Segmented linear regression models were fitted using the estimation methods described in Section 2.2. We evaluated estimation methods designed to estimate the model parameters under lag-1 autocorrelation (see Table 2 for details). For GLS, we restricted our investigation to the PW method, because it was expected to have better performance than the CO method (on which PW is based) given the PW method utilises all data points. For REML with the Satterthwaite adjustment, we substituted d.f. of 2 when the computed d.f. were less than 2, to avoid overly conservative confidence limits and hypothesis tests. We also investigated the commonly used Durbin-Watson method for detecting autocorrelation at a significance level of 0.05 (Durbin and Watson 1950).

Table 2 summarises the methods and model variations used to adjust for autocorrelation. Details of the `Stata` code used for generating the simulated data and the analysis code can be found in Supplementary File 2.

Table 2: Statistical methods and adjustments for autocorrelation.

Method	Autocorrelation adjustment
Ordinary Least Squares	None Newey-West SE adjustment (lag-1)
Generalised least squares	Prais-Winsten
Restricted maximum likelihood	Lag-1 autocorrelation Lag-1 autocorrelation with small sample Satterthwaite approximation
Autoregressive integrated moving average	Lag-1 autocorrelation (i.e. ARIMA(1,0,0))

2.2.4 Performance Measures

The performance of the methods was evaluated by examining bias, empirical SE, model-based SE, 95% confidence interval coverage and power (see Appendix 2 for formulae). Confidence intervals were calculated using the `simsum` package (White 2010) with t-distribution critical values. For each simulation scenario, we used 10,000 repetitions in order to keep the Monte Carlo Standard Error (MCSE) below 0.5% for all potential values of coverage and type I error rate. Model non-convergence was recorded and tabulated.

2.2.5 Coding and Execution

The statistical software `Stata` version 15 (Stata 2017) was used for the generation of the simulated data. A random seed was set at the beginning of the process and the individual random state was recorded for each repetition of the simulated data sets. Each dataset was independently simulated, using consecutive randomly generated numbers from the starting seed. We used a “burn in” period between each dataset of 300 random number generations so that any lag effects specific to the computer-generated series had time to dissipate (Huitema and McKean 2007).

Prior to running the simulations, we undertook initial checks to confirm that the data generation mechanism was working as expected. This involved fitting series of length 100,000 to check the estimated β parameters matched the input parameters. A larger sample of 1,000 datasets was then simulated and checked using summary statistics and graphs. When we were satisfied that the simulations were operating as expected, the full number of datasets were simulated.

2.2.6 Analysis of the simulated datasets

Analyses were performed using `Stata` version 15 (Stata 2017). A range of visual displays were constructed to compare the performance of the statistical methods. Frequency distributions were plotted to visualise the level- and slope-change estimates, autocorrelation coefficient estimates, and the results of the Durbin-Watson test for autocorrelation. Scatter plots were used to display the mean values for empirical and model-based SEs, coverage, power and autocorrelation coefficient estimates. Line plots were used to show confidence intervals for the level and slope change estimates. Results and summaries of the analyses were summarised (using the `simsum` package (White 2010)) and graphed using `Stata` version 15 (Stata 2017).

3 Results of the simulation study

3.1 Bias of level and slope change estimates

All methods yielded approximately unbiased estimates of level change and slope change across all simulation scenarios. Figure 3 presents level change estimates specific to the scenario of a level change of 2 and a slope change of 0.1 (Supplementary Figure S2 shows slope change estimates), but the other 7 combinations of level and slope changes were virtually identical (Supplementary 1.3.1 for level change, Supplementary 1.3.2 for slope change). Note that the Satterthwaite and NW adjustments do not impact the parameter estimates of level or slope change, hence distributions of these parameter estimates are not shown in Figures 3 and S2.

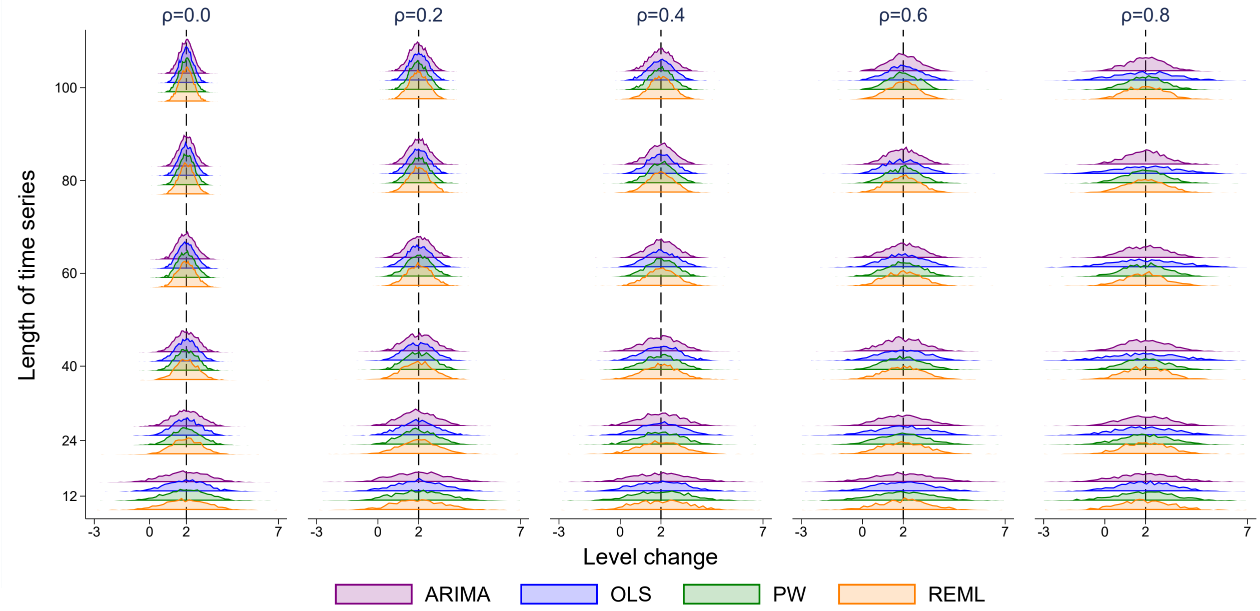


Figure 3: Distributions of level change estimates calculated from four statistical methods, from top to bottom: autoregressive integrated moving average (ARIMA) (purple), ordinary least squares regression (OLS) (blue), Prais-Winsten (PW) (green) and restricted maximum likelihood (REML) (orange). The vertical axis shows the length of the time series. The five vertical columns display the results for different values of autocorrelation. The vertical black line represents the true parameter value (β2). Each subset of four curves shows the distribution from a different analysis method for a given combination of time series length and autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other structures give similar results. The Satterthwaite adjustment to the REML method and the Newey-West adjustment to the OLS method do not impact the estimate of level or slope change, hence these parameter estimates are not shown.

3.2 Standard errors of level and slope change estimates

3.2.1 Empirical standard errors

Figure 3 and Supplementary Figure S2 visually indicate the precision of the estimators in terms of the spread of the distributions therein. To enable a direct quantitative assessment, we plotted the empirical SE of the level and slope changes for each method against selected series lengths and autocorrelation parameter sizes for a level change of 2 and slope change of 0.1 (Figure 4 and Figure 5). The size of the empirical SE of the level change was dependent on the underlying autocorrelation, length of the series and statistical method (Figure 4). Of note, the estimates obtained from the ARIMA and PW models yield almost identical empirical SEs. For each magnitude of autocorrelation, the empirical SE decreased as the length of the time series increased, as would be expected. An exception to this occurred for the OLS estimator (and to a lesser extent ARIMA) which exhibited unusual behaviour for an autocorrelation of 0.8, with the SE initially increasing with an increasing number of points in the series, and then decreasing. Supplementary simulations were undertaken to examine the behaviour of the OLS estimator for surrounding correlations (0.7 and 0.9), which showed a similar pattern of increasing SEs with an increasing number of points (Supplementary 1.4). The relationship between autocorrelation and the empirical SE was modified by the length of series. For small series (< 10 data points), the empirical SE decreased with increasing autocorrelation, while for longer series (≥ 10 data points) this relationship was reversed, with SEs increasing with increasing autocorrelation.

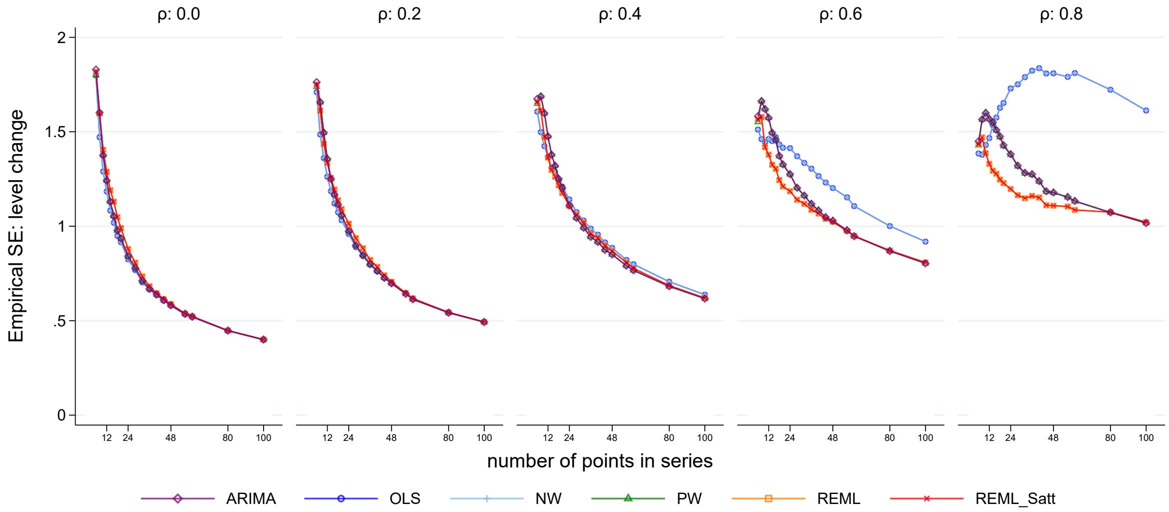


Figure 4: Empirical standard error (SE) of the level change. The horizontal axis shows the length of the time series, the vertical axis shows the empirical SE. The five vertical columns display the results for different values of autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; PW, Prais-Winsten; REML, restricted maximum likelihood.

The size of the empirical SE for slope change was dependent on the underlying autocorrelation and length of the series (Supplementary Figure S2 and Figure 5). The empirical SE decreased with increasing series length, but increased with increasing autocorrelation, as would be expected. In contrast to the level change, there were no important differences in the empirical SEs across the statistical methods, even when the autocorrelation was large. The observed patterns did not differ for any of the eight level and slope change combinations (Supplementary 1.3.3 for level change, Supplementary 1.3.4 for slope change).

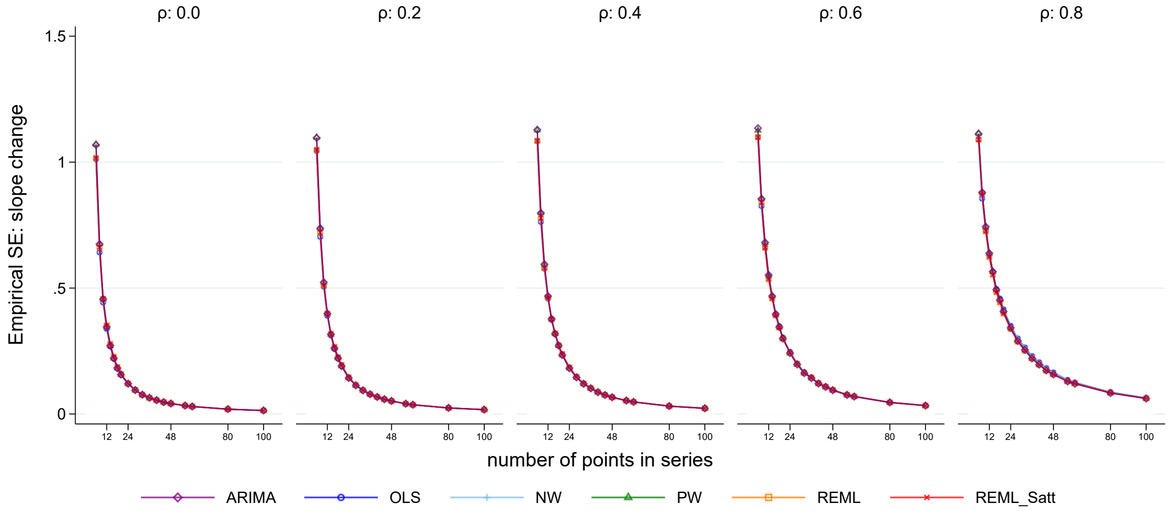


Figure 5: Empirical standard error (SE) of the slope change. The horizontal axis shows the length of the time series, the vertical axis shows the empirical SE. The five vertical columns display the results for different values of autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; PW, Prais-Winsten; REML, restricted maximum likelihood.

3.2.2 Comparison between empirical and model-based standard errors

To enable appropriate confidence interval coverage and size of significance tests, the model-based SE needs to be similar to the empirical SE (Morris, White et al. 2019). In this section we present the comparison between the empirical and model-based SEs; results for the model-based SEs alone can be found in S1.3.5 for level change and S1.3.6 for slope change.

For the level change parameter (β_2) estimated by OLS, the ratio of model-based to empirical SEs were close to one (indicating the empirical and model-based SEs were similar) for all series lengths when there was no underlying autocorrelation (Figure 6). However, as autocorrelation increased, as expected, the OLS model-based SEs became increasingly smaller relative to the empirical SEs, indicating the model-based SEs is are downwardly biased. The NW method performed only slightly better than the OLS (except when the autocorrelation was zero); however, the NW model-based SEs were still downwardly biased across all scenarios, were worse than OLS for small series lengths, and only marginally better than OLS for large series lengths. Although the empirical SEs of the ARIMA and PW methods were similar, they had quite different model-based SEs. The PW model-based SEs were smaller than the empirical SEs for all magnitudes of autocorrelation, though the model-based SEs approached the empirical SEs with increasing series length. The ARIMA model-based SEs were larger than the empirical SEs for small series (fewer than 24 points) at small underlying values of autocorrelation ($\rho < 0.4$) and also for larger series (more than 24 points) at higher magnitudes of autocorrelation ($\rho > 0.4$). Aside from these scenarios, the ARIMA model-based SEs were approximately equal to the empirical SEs. The REML method behaved similarly to the PW method but, relatively, did not underestimate the SEs to the same extent. For small values of underlying autocorrelation ($\rho < 0.4$) and series greater than 30 points, the model-based SEs were similar to the empirical SEs.

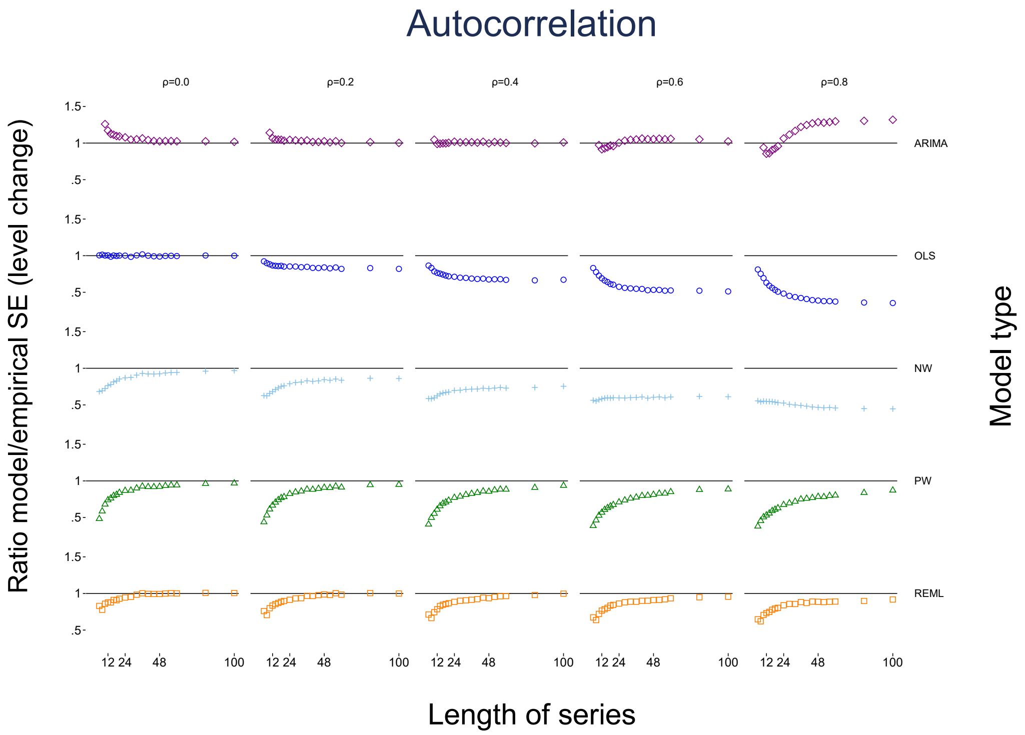


Figure 6: Scatter plots of the ratio of model-based standard error (SE) to the empirical SE for the level change parameter with different levels of autocorrelation and series length. The horizontal axis represents the number of points in the time series, the vertical axis shows the ratio of model-based to empirical SE. The five vertical columns display the results for different values of autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. The first two series lengths are not shown for the ARIMA method due to extreme values. The Satterthwaite adjustment to the REML does not impact the estimate of SE, hence details of this method are not shown. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood.

For the slope change parameter (β_3), the ratios of model-based to empirical SEs followed similar patterns as for the level change parameter (β_2). For any given series length, as the magnitude of autocorrelation increased, model-based SEs became increasingly smaller compared with the empirical SEs for most statistical methods (Supplementary 1.5). Model-based and empirical SEs tended towards equivalence as series lengths increased, with the exception of OLS and NW at high values of autocorrelation ($\rho > 0.6$). For REML and ARIMA, the pattern of ratios of model-based to empirical SEs for β_3 slightly differed compared with β_2. Specifically, the REML model-based SEs were smaller than the empirical SEs for small series, and then increased to be slightly larger as the number of points increased. For ARIMA, the model-based SEs were smaller than the empirical SEs for large underlying values of autocorrelation ($\rho \geq 0.6$) for small to moderate length series. The observed patterns did not differ for any of the eight level and slope change combinations (S 1.3.5 for level change, S 1.3.6 for slope change).

3.3 Confidence interval coverage

For all combinations of level change, slope change, number of time points and autocorrelation, most methods had coverage (percentage of 95% confidence intervals including the true parameter) that was less than the nominal 95% level for both level and slope change (Figure 7 for level change and Figure 8 for slope change, both with a level change of 2 and slope change of 0.1, Supplementary 1.3.7 for level change and Supplementary 1.3.8 for slope change for other parameter combinations). The exceptions were OLS when there was no underlying autocorrelation, and REML with the Satterthwaite adjustment for moderate to large length series. In general, mean values of coverage decreased with increasing autocorrelation and increased with increasing series length. However, coverage of the OLS method decreased with increasing autocorrelation as well as with increasing series length (with the exception of the zero autocorrelation scenario). The NW method exhibited a similar pattern to OLS, but generally had better coverage (except for small autocorrelations), although coverage was often poor (under 90% for all but the longest series with low autocorrelation, $\rho < 0.4$). REML with the Satterthwaite small sample adjustment yielded coverage greater than the nominal 95% level when the number of data points was greater than 30 in the presence of autocorrelation. Confidence interval coverage patterns generally reflected those observed with the comparisons between the model-based and empirical SE.

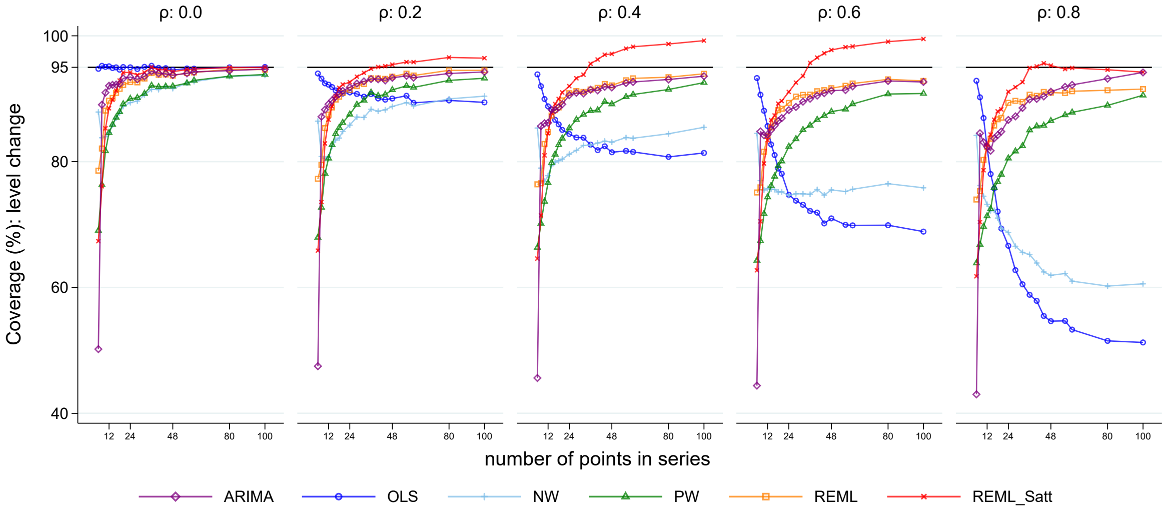


Figure 7: Coverage for the level change parameter. Each point is the proportion of the 10,000 simulations in which the 95% confidence interval included the true value of the parameter. The solid black line depicts the nominal 95% coverage level. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

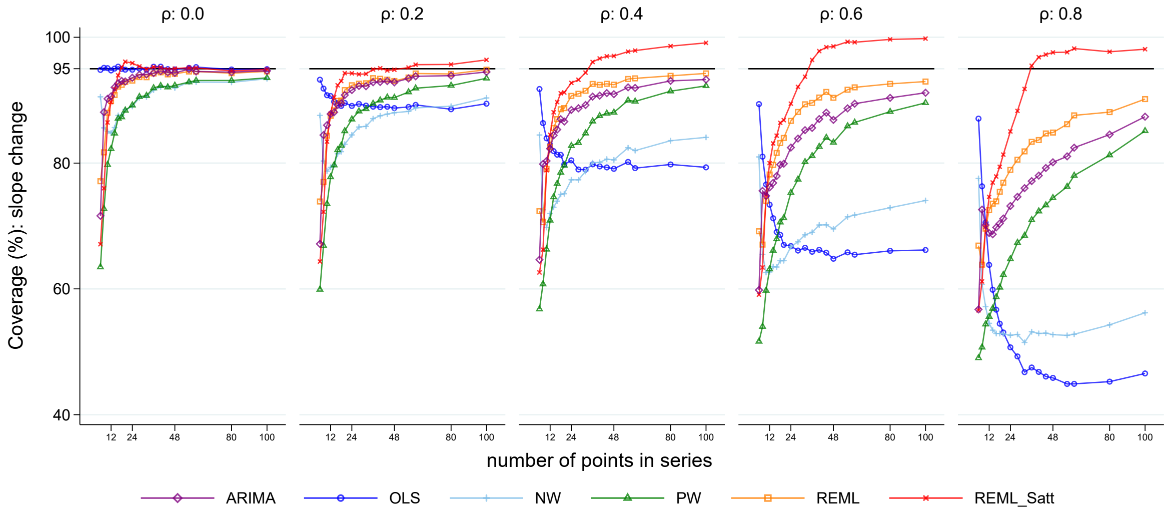
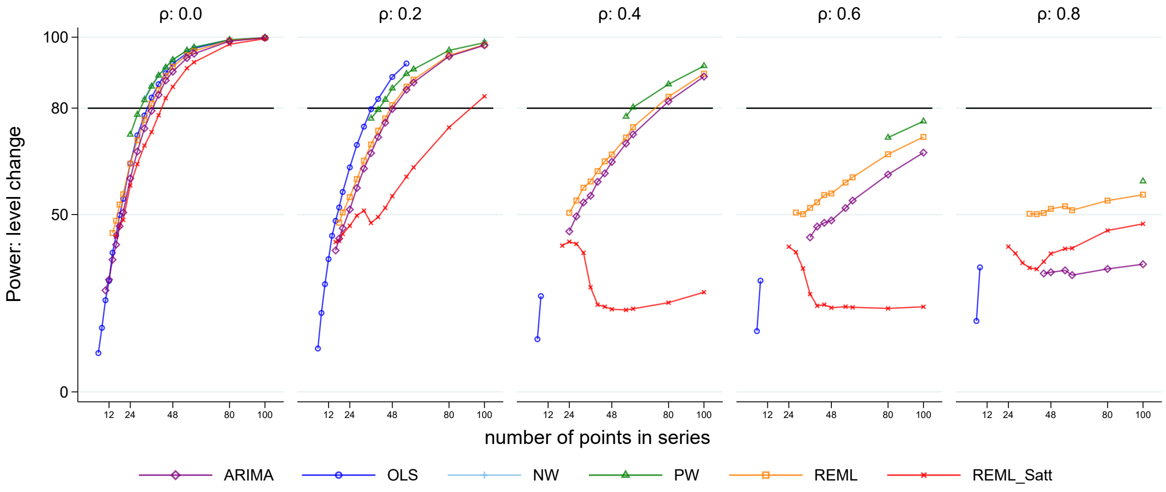


Figure 8: Coverage for the slope change parameter. Each point is the proportion of the 10,000 simulations in which the 95% confidence interval included the true value of the parameter. The solid black line depicts the nominal 95% coverage level. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

3.4 Power

Coverage was less than the nominal 95% level in the majority of scenarios (except for the OLS model in the absence of autocorrelation and some scenarios involving the REML method with Satterthwaite adjustment). In scenarios where coverage is less than 95%, examining power is misleading. Due to there being only a very small number of configurations in Figure 7 and Supplementary 1.6 in which 95% coverage was achieved, we adopt a more liberal approach and consider configurations in which the coverage was at least 90%. As such, the results presented below should be viewed as approximate power only and will generally be lower than the value observed if coverage was at least 95%.

For scenarios with a level change of two, power was low for series with a small number of points, but predictably, increased as the number of points increased for all methods, except the REML method with Satterthwaite adjustment (Figure 9). As the magnitude of autocorrelation increased its power decreased, to a point where it became lower than for other methods. This was due to the REML method with Satterthwaite adjustment having greater than 95% coverage in these situations and hence substantially lower than 5% Type I error rates. For smaller values of the level change parameter, predictably, power decreased (Supplementary 1.6.1). Similar patterns were observed for slope change (Supplementary 1.6.2).



True level change parameter value = 2. Only scenarios with coverage > 90% are used here.

Figure 9: Power for level change. Each point is the mean number of times the 95% confidence interval of the estimate did not include zero from 10,000 simulations. The simulation combination presented is for a level change of 2 and slope change of 0.1. Power for other model combinations is available in Supplementary 1.8.1. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite; NW, Newey-West.

3.5 Autocorrelation Coefficient

Most of the statistical methods yield an estimate of the autocorrelation coefficient. All methods underestimated the autocorrelation for series with a small number of points (Figure 10 and Figure 11 show parameter values of 2 for level change and 0.1 for slope change). However, underestimation was most pronounced for scenarios with small series and large underlying autocorrelation. The REML method always yielded estimated autocorrelations closer to the true underlying autocorrelation compared with the other methods. The empirical SEs for autocorrelation generally decreased as the series length increased for all methods (except for small series with fewer than 20 points) (Supplementary 1.7). The observed patterns did not differ for any of the eight level and slope change combinations (Supplementary 1.3.9).

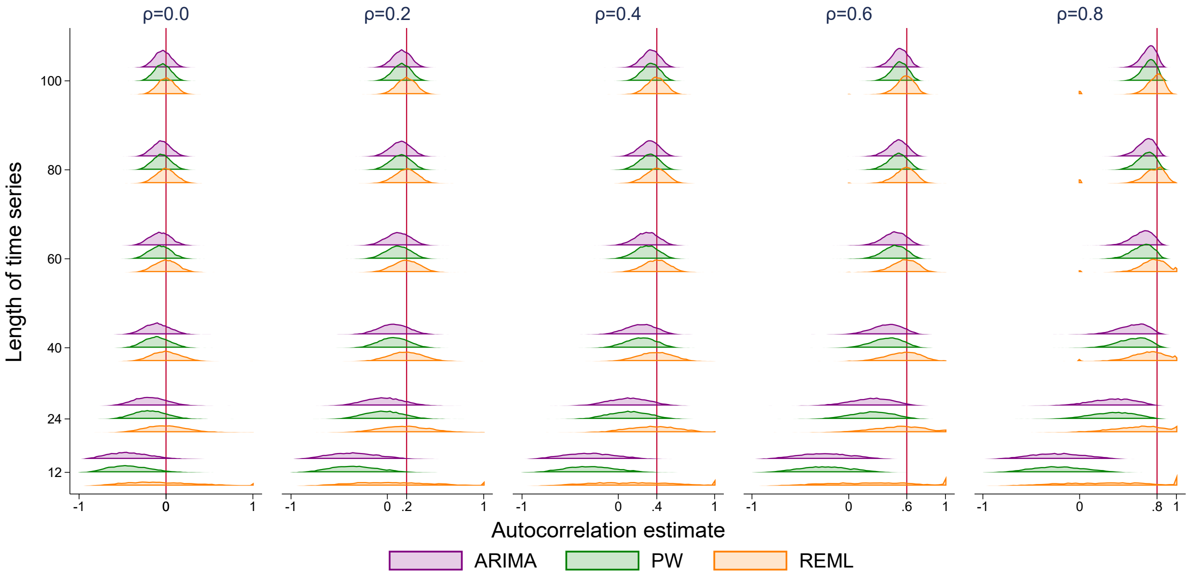


Figure 10: Autocorrelation coefficient estimates. The horizontal axis shows the estimate of autocorrelation coefficient. The vertical axis shows the length of the time series. The five vertical columns display the results for different values of autocorrelation ranging from 0 to 0.8 (the value of autocorrelation is shown by a vertical red line). Each coloured curve shows the distribution of autocorrelation coefficient estimates from 10,000 simulations. Each subset of four curves shows the results from a different analysis method for a given combination of time series length and autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. From top to bottom the methods are autoregressive integrated moving average (ARIMA) (purple), Prais-Winsten (PW) (green) and restricted maximum likelihood (REML) (orange).

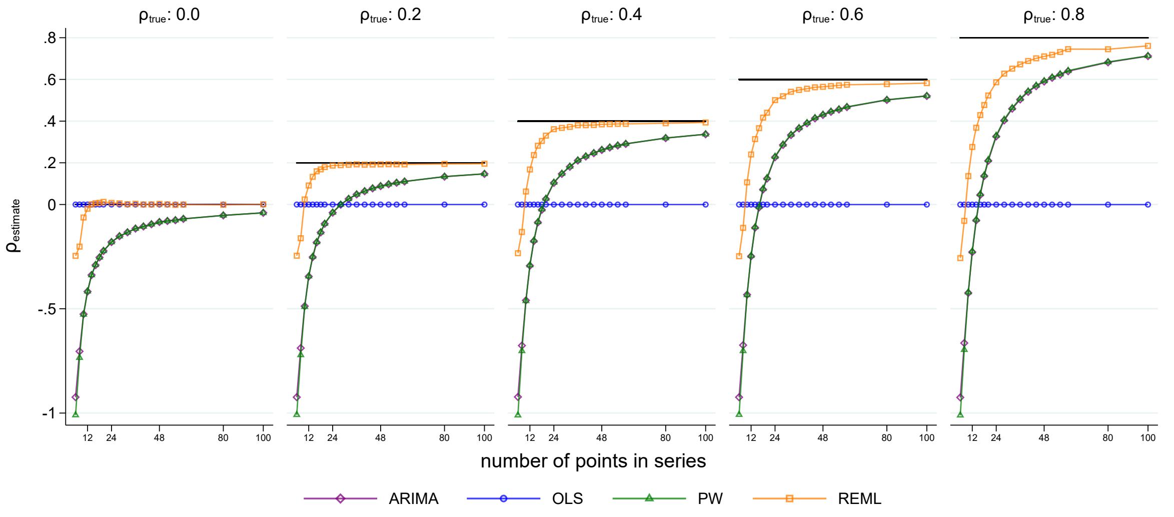


Figure 11: Autocorrelation coefficient estimates. The horizontal axis shows the length of the time series. The vertical axis shows the mean estimate of the autocorrelation coefficient across 10,000 simulations. The five plots display the results for different values of autocorrelation ranging from 0 to 0.8 (the true value of autocorrelation is shown by a horizontal black line). Each coloured point shows the mean autocorrelation estimate for a given combination of true autocorrelation coefficient and number of points in the data series. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; PW, Prais-Winsten; REML, restricted maximum likelihood.

3.5.1 Durbin-Watson test for autocorrelation

The DW test for detecting autocorrelation performed poorly except for long data series and large underlying values of autocorrelation (Figure 12). For series of moderate length (i.e. 48 points), with an underlying autocorrelation of 0.2, the DW test gave an “inconclusive” result in 30% of the simulations, incorrectly gave a value of no autocorrelation in 63% of the simulations, and only correctly identified that there was autocorrelation in 7% of the simulations. For shorter length series the percentage of simulations in which autocorrelation was correctly identified decreased (for a series length of 24 even at extreme magnitudes of autocorrelation (i.e. 0.8) positive autocorrelation was reported in only 26% of the simulations). For very short length series (fewer than 12 data points) the DW test gave an “inconclusive” result in over 75% of the simulations for all values of autocorrelation and always failed to identify that autocorrelation was present.

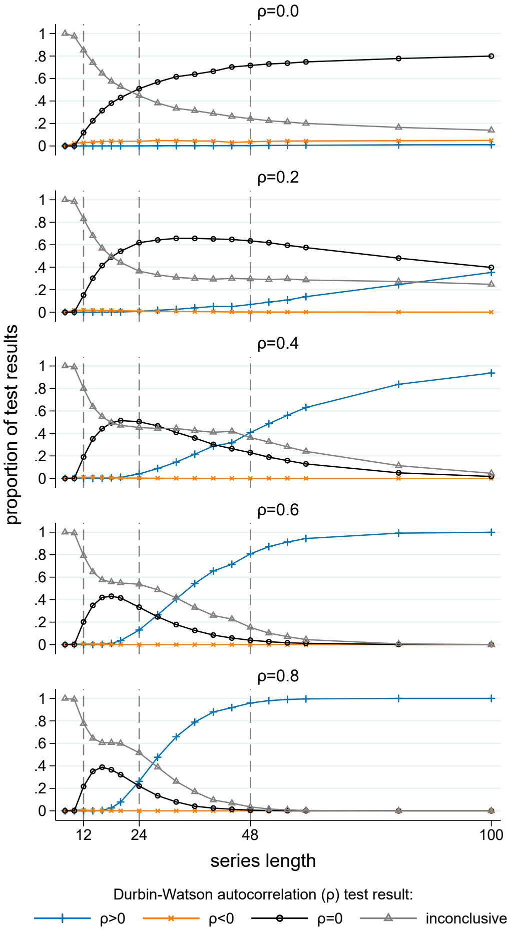


Figure 12: Durbin-Watson tests for autocorrelation. For each combination of length of data series and true magnitude of autocorrelation the Durbin Watson test results from 10,000 simulated data sets are summarised. The horizontal axis is the length of the data series, the vertical axis is the proportion of results indicating: ρ > 0 (blue), ρ < 0, (orange) ρ = 0 (black) and an inconclusive test (grey). Each graph shows results for a different magnitude of autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results.

3.6 Convergence of estimation methods

The number of the 10,000 simulations in which the estimation methods converged is presented in Supplementary 1.8. Most methods had no numerical convergence issues. The PW model failed to converge a small number of times (less than 7% of simulations) when there were only three data points pre- and post-interruption. The REML model regularly failed to converge (approximately 70% convergence) for short data series (fewer than 12 data points) at all values of autocorrelation, however convergence improved substantially as the number of points in the series increased. In addition, convergence issues for REML occurred more frequently for higher values of autocorrelation, unless the series length was large.

3.7 Analysis of motivating example

We re-analysed the ITS study (introduced in Section 2) using each of the statistical methods evaluated in the simulation study to estimate the effect of terminal room cleaning with dilute bleach on *C difficile* rates. Estimates of level and slope change (along with their confidence intervals and p-values) and autocorrelation are presented in Table 3. The point estimates for level and slope change are similar across methods, but notably, the width of the confidence intervals vary considerably. The confidence intervals are narrower for OLS, NW and PW, but wider for REML (with and without the Satterthwaite adjustment) and ARIMA. For level change, this led to corresponding p-values that ranged from 0.002 to 0.095; and for the slope change, p-values ranging from 0.069 to 0.531. Estimates of autocorrelation also varied, with REML yielding an estimate of 0.23, while ARIMA and PW yielded much lower estimates of 0.07. The DW statistic was 1.86, indicating no autocorrelation. Such differences in confidence interval width and p-values may impact on the interpretation of the results.

Table 3: Level- and slope-change point estimates with 95% confidence intervals (CIs), p-values and estimate of magnitude of lag-1 autocorrelation ($\hat{\rho}_{est}$) from C difficile infection data using a range of statistical methods. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

	Level change		**Slope change**		$\hat{\rho}_{est}$
	Estimate (CI)	**p-value**	**Estimate (CI)**	**p-value**	
ARIMA	-0.42 (-0.89,0.05)	0.079	-0.03 (-0.11,0.06)	0.531	0.07
OLS	-0.44 (-0.76,-0.13)	0.008	-0.03 (-0.07,0.02)	0.201	N/A
NW	-0.44 (-0.71,-0.17)	0.002	-0.03 (-0.06,0.00)	0.069	N/A
PW	-0.42 (-0.75,-0.09)	0.014	-0.03 (-0.08,0.02)	0.251	0.07
REML	-0.37 (-0.72,-0.01)	0.044	-0.02 (-0.08,0.03)	0.390	0.23
REML-Satt	-0.37 (-0.82,0.09)	0.095	-0.02 (-0.10,0.05)	0.437	N/A

4 Discussion

4.1 Summary and discussion of key findings

Interrupted time series studies are commonly used to evaluate the effects of interventions or exposures. The results of our simulation study provide insight into how a set of statistical methods perform under a range of scenarios which included different level and slope changes, varying lengths of series and magnitudes of autocorrelation. We chose to examine statistical methods that are commonly used in practice for interrupted time series studies (Ramsay, Matowe et al. 2003, Jandoc, Burden et al. 2015, Hudson, Fielding et al. 2019, Ewusie, Soobiah et al. 2020, Turner, Karahalios et al. 2020), and those performing well in the general, non-interrupted, time series literature (Cheang and Reinsel 2000, Alpargu and Dutilleul 2003).

Not surprisingly, we found that the statistical methods all yielded unbiased estimates of both level and slope change for all values of model shape, length of series and autocorrelation. Confidence interval coverage, however, was generally below the nominal 95% level, except in particular circumstances for specific methods. The REML method with and without the Satterthwaite adjustment had improved confidence interval coverage compared with the other statistical methods, particularly for slope change. An exception to this was for very small series (fewer than 12 points), where the OLS method had better coverage than the other methods, even in the presence of large underlying autocorrelation. Coverage improved for most methods with increasing series length (with the exception of OLS and NW in some circumstances). REML with the Satterthwaite adjustment to the d.f. was the only method that yielded at least the nominal level of confidence interval coverage, however it was overly conservative in some scenarios, with a resultant reduction in power compared with other methods.

Autocorrelation was systematically underestimated by all statistical methods, with estimates of autocorrelation being particularly biased (and often negative) for small time series (fewer than 24 points). This underestimation of autocorrelation had a detrimental impact on the estimates of SE, which were too small, and in turn, this led to confidence interval coverage that was less than the nominal 95% level. This can be seen in Figure 13 (level change) and Supplementary 1.9 (slope change), where a relationship between the magnitude of bias in the estimates of autocorrelation and confidence interval coverage is clearly evident. Ideally the confidence interval coverage should be at the nominal 95% level with no bias in autocorrelation (the intersection of the dashed lines in Figure 13). For short time series, the severe underestimation of autocorrelation led to poorer confidence interval coverage than had autocorrelation been ignored, as is the case with OLS.

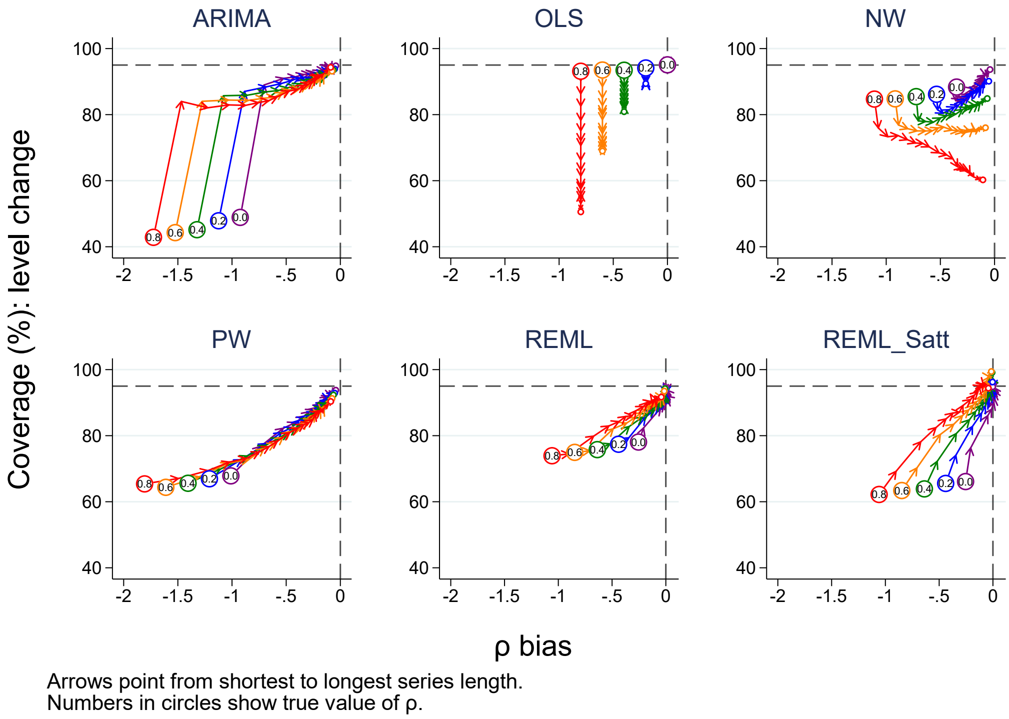


Figure 13: Bias in autocorrelation estimate versus coverage for level change. The horizontal axis shows the bias in the autocorrelation estimate. The vertical axis shows the percentage coverage. The horizontal dashed line indicates 95% coverage, the vertical dashed line indicates no bias in the estimate of autocorrelation. Each colour represents a different value of underlying autocorrelation, ranging from zero (purple) to 0.8 (red), with each value displayed in a circle at the smallest series length (six points). The arrows point from shortest to longest series length, with the small circles at the end of each line showing coverage at a series length of 100 data points. Each data point shows the mean value from 10,000 simulations for a given combination of autocorrelation coefficient and number of points in the series. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

We included REML due to its potential to reduce bias in the variance parameters compared with maximum likelihood. Although the ARIMA model fitted in our simulations used maximum likelihood estimation, the model-based SEs were generally more similar to the empirical SEs for the ARIMA method compared with the REML method (where the model-based SEs were generally smaller than the empirical SEs). ARIMA confidence interval coverage was similar to REML for level change, though REML showed improved confidence interval coverage for slope change. Further, the REML method yielded less biased estimates of autocorrelation than the other methods, even for small series lengths.

The only method to yield overly conservative confidence intervals was the REML with SW adjustment to the t-distribution d.f.. When deciding whether to use the Satterthwaite adjustment, consideration therefore needs to be made between the trade-off in the risk of type I and type II errors. A further issue we identified with the Satterthwaite adjustment was that the adjusted d.f. were very small in some series, leading to nonsensible confidence intervals. To limit this issue we set a minimum value of 2 for the d.f., but other choices could be adopted.

The DW test is the most commonly used test to identify autocorrelation and is often used when series are short (Hudson, Fielding et al. 2019, Turner, Karahalios et al. 2020). Some authors use the test as part of a two-stage analysis strategy where they first test for autocorrelation, and depending on the result of the test, either use a method that attempts to adjust for autocorrelation or not. This type of two-stage approach is used in other contexts, such as testing for carryover in crossover trials. The findings of our simulation study underscore why such two stage approaches fail and are discouraged; namely, due to their failure to detect the presence of a statistic when it exists (i.e., their high type II error rate). In our case, we found that for short series (fewer than 12 data points), the DW test failed to identify autocorrelation when it was present, and for moderate length series (i.e. 48 points), with an underlying autocorrelation of 0.2, autocorrelation was only detected in 7% of the simulations.

4.2 Comparisons with other studies

We are not aware of other simulation studies that have examined the performance of statistical methods for interrupted time series studies. However, other simulation studies have investigated the performance of methods for general time series, and our findings align with these. Alpargu and Dutilleul (Alpargu and Dutilleul 2003) concluded from their simulation study examining the performance of REML, PW and OLS for lag(1) time series data over a range of series lengths (from 10 to 200), that REML is to be preferred over OLS and PW in estimating slope parameters. Cheang and Reinsel (Cheang and Reinsel 2000) examined the performance of ML and REML for estimating linear trends in lag(1) time series data of length 60 and 120 (both with and without seasonal components) and concluded that the REML estimator yielded better confidence interval coverage for the slope parameter, and less biased estimates of autocorrelation. Smith and McAleer (Smith and McAleer 1994) examined the performance of the NW estimator for time series of length 100 with lags of 1, 3 and 10, and found that it underestimated the SEs of the slope parameter.

4.3 Strengths and Limitations

The strengths of our study include that we have used many combinations of parameter estimates and statistical methods. Our parameter values were informed by characteristics of real world ITS studies (Turner, Karahalios et al. 2020). We planned and reported our study using the structured approach of Morris et al (Morris, White et al. 2019) for simulation studies, and we generated a large number of data sets per combination to minimise MCSE.

As with all simulation studies, there are limitations to the applicability of findings. All data series were based on a random number generator and results may change given a different set of series, however, this is unlikely to be problematic given our MCSE was < 0.5% for all potential values of coverage and type I error rate. Our findings are only applicable to the scenarios in which they were generated, and so may not apply to ITS studies with different characteristics, such as unequal numbers of time points in the pre- and post-interruption segments, non-constant variance or different lags of autocorrelation.

4.4 Implications for practice

We found that all methods yielded unbiased estimates of the level and slope change, however, the methods differed in their performance in terms of confidence interval coverage and estimation of the autocorrelation parameter. Confidence interval coverage was primarily determined by the length of the time series and the underlying magnitude of autocorrelation. In practice, however, most analysts will only have knowledge of the length of the time series to guide in the choice of method. In rare cases, knowledge of the likely size of the underlying autocorrelation may be available from a previous long time series study in a similar context, which could help inform their choice. In our review of ITS studies investigating public health interruptions or exposures, the magnitude of autocorrelation was almost never explicitly specified (1%, 3/230 time series) (Turner, Karahalios et al. 2020). Analysis of data extracted from the ITS studies included in this review using the REML method yielded a median autocorrelation 0.2 (IQR: 0 to 0.6, n=180); however, as shown from the simulation study, the estimates of autocorrelation (on which these summary statistics are based) are likely to be underestimated.

From the statistical methods and scenarios we examined, we found that for small time series (approximately 12 points or under), in the absence of a method that performs well adjusting for autocorrelation in such short series, OLS is the recommended method. For longer time series, REML is recommended. If the analyst has knowledge that the underlying autocorrelation is likely to be large, then using REML with the Satterthwaite adjustment may be advantageous. However, when the Satterthwaite adjustment yields d.f. lower than 2, we recommend replacing these with 2 to mitigate nonsensible confidence intervals. When REML doesn’t converge, ARIMA provides a reasonable alternative as, with the exception of REML, it yields higher confidence interval coverage than the other methods. Given most methods will yield confidence intervals that are too small, with type I error rates greater than 5%, borderline findings of statistical significance for the regression parameters should be cautiously interpreted; these may be due to chance rather than as a result of the interruption.

Estimates of autocorrelation from long series can be useful to inform sample size calculations and analytical decisions in future studies. We recommend reporting the REML estimates of the autocorrelation coefficient when possible. We only recommend using the DW test for detecting underlying autocorrelation in long time series (longer than 100 data points) and recommend against its use as part of a two-stage or stepwise approach to determine whether to use a statistical method that adjusts for autocorrelation.

In terms of study design, we recommend using at very minimum 24 points data points. With this number of points, confidence interval coverage close to the nominal 95% level can be achieved using REML with the Satterthwaite adjustment (when underlying autocorrelation is between 0 and 0.6). With fewer data points, poor confidence interval coverage is likely, irrespective of method.

4.5 Implications for future research

Although we investigated the statistical methods most commonly observed in reviews of ITS studies (Ramsay, Matowe et al. 2003, Jandoc, Burden et al. 2015, Hudson, Fielding et al. 2019, Ewusie, Soobiah et al. 2020, Turner, Karahalios et al. 2020), there is scope for further research examining other statistical methods, such as robust methods (Cruz, Bender et al. 2017) or Bayesian approaches where the uncertainty in the estimate of autocorrelation could be incorporated. We investigated one small-sample adjustment (Satterthwaite) though others, such as Kenward-Roger (Kenward and Roger 1997), which adds a correction to the SE of regression parameter estimates, could also be examined. Further investigation of how the methods perform for scenarios other than those we investigated would be valuable. For example, when there are unequal numbers of points pre- and post-interruption, lags greater than 1, and where the autocorrelation and error variance differ between the pre and post interruption periods.

4.6 Conclusion

We undertook a simulation study to examine the performance of a set of statistical methods to analyse ITS data under a range of scenarios that included different level and slope changes, varying lengths of series and magnitudes of autocorrelation. We found that all methods yielded unbiased estimates of the level and slope change, however, the magnitude of autocorrelation was underestimated by all methods. This generally led to SEs that were too small and confidence interval coverage that was less than the nominal level. The DW test for the presence of autocorrelation performed poorly except for long series and large underlying autocorrelation. Care is needed when interpreting results from all methods, given the confidence intervals will generally be too narrow. Further research is required to determine and develop methods that perform well in the presence of autocorrelation, especially for short series.

5 Acknowledgements

5.1 Acknowledgements

This work forms part of SLT’s PhD, which is supported by an Australian Postgraduate Award administered through Monash University, Australia.

5.2 Availability of data and materials

The data that supports the findings of this study were generated by the simulation code available in the supplementary material of this article.

5.3 Competing interests

The authors declare that they have no competing interests.

5.4 Funding

This work was supported by the Australian National Health and Medical Research Council (NHMRC) project grant (1145273). SLT was funded through an Australian Postgraduate Award administered through Monash University, Australia. JEM is supported by an NHMRC Career Development Fellowship (1143429). The funders had no role in study design, decision to publish, or preparation of the manuscript.

5.5 Author contributions

JEM and ABF conceived the study and all authors contributed to its design. SLT designed and wrote the computer code and ran and analysed the simulations. SLT wrote the first draft of the manuscript, with contributions from JEM. SLT, JEM, AK, ABF, MT contributed to revisions of the manuscript and take public responsibility for its content.

6 Supplementary files

There are two files that supplement this study.

6.1 Supplementary File 1

This file contains supplementary graphs as described in the text.

File: STurner_ITS_Simulation_Supplementary_File_1.docx

6.2 Supplementary File 2

This file contains the data and computer code used to analyse the motivating example, the computer code used to create and analyse the simulated data sets, and the computer code used to plot the graphs in the manuscript (including Supplementary File 1), available via figshare: https://doi.org/10.26180/13284329 (Turner 2020).

Appendix 1 Statistical method details

Appendix 1.1 Ordinary Least Squares

Model (1) can be written in a matrix form as:

$$Y = X\beta + \varepsilon \qquad (3)$$

where *Y* and ε are *n* × *1* vectors whose *t*th element is y_t and ε_t respectively, *X* is the *n* × *4* design matrix with $t'th\ row\ \big(1, t, D_t, D_t I(t - T_1)\big)$, and $\epsilon_t \sim N(0, \sigma^2)$. The OLS estimator of β is $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$, and $Var(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}$.

Appendix 1.2 Newey West

The NW estimator (lag-1) of β is just the OLS estimator, $\hat{\beta}_{NW} = \hat{\beta}_{OLS}$, but with a sandwich variance estimator of the form

$$\widehat{Var}\big(\hat{\beta}_{NW}\big) = (X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1} \qquad (5)$$

where:

$$X'\hat{\Omega}X = X'\hat{\Omega}_0X + \frac{n}{n-k}\frac{1}{2}\sum_{t=2}^{n} \hat{e}_t\hat{e}_{t-1}(x_t'x_{t-1} + x_{t-1}'x_t) \qquad (6)$$

$$X'\hat{\Omega}_0X = \frac{n}{n-k}\sum_i \hat{e}_i^2\, x_i'x_i \qquad (7)$$

$$\hat{e}_i = y_i - x_i\hat{\beta}_{OLS} \qquad (8)$$

where *X* is the same $n \times 4$ design matrix as specified for OLS above. The central term in the variance expression allows for empirical determination of autocorrelation and heteroskedasticity (StataCorp 2017).

Appendix 1.3 Generalised Least Squares

In the Cochrane-Orcutt and Prais-Winsten methods, from the equations (1) and (2), the dependent and independent variables are transformed to create a new model in which the error terms are uncorrelated:

$$Y_t^* = Y_t - \rho Y_{t-1} \qquad (9a)$$

$$X_t^* = X_t - \rho X_{t-1} \qquad (9b)$$

Then fit $Y_t^* = X_t^* \beta + w_t$, where

$$w_t = \varepsilon_t - \rho \varepsilon_{t-1} \sim N(0, \sigma^2) \qquad (10)$$

using OLS, and iterate until convergence.

Generally, the correlation is unknown, and must first be estimated. An estimate of autocorrelation at each iteration can be obtained using the OLS residuals e_t from fitting Equation (2) as above:

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_{t-1} e_t}{\sum_{t=2}^{n} e_{t-1}^2} \qquad (11)$$

The CO method discards the first observation, while the PW method retains the first observation, but applies the following transformation (Prais 1954):

$y_1^* = \sqrt{1-\rho^2} y_1 \; and \; X_1^* = \sqrt{1-\rho^2} X_1$, where X_1 is the first row of X. *(12)*

Appendix 1.4 Autoregressive integrated moving average

The ARIMA model includes parameters that model observations and error terms from previous time points. In an ARIMA model with first order autocorrelation only, i.e. ARIMA(1,0,0), equations (1) and (2) are fit simultaneously by maximum likelihood.

Appendix 1.5 Durbin-Watson test for autocorrelation

The Durbin-Watson test statistic is given by:

$$D = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad (13)$$

For test statistic values under two, D is compared to lower (d_L) and upper (d_U) bounds, leading to either a conclusive or inconclusive result. For test statistic values over two, 4-D is compared to the lower and upper bounds and a conclusive $H_{alternative}$ indicates the presence of negative autocorrelation:

$$If\ D > d_U, conclude\ H_o$$

$$If\ D < d_L, conclude\ H_{alternative}$$

$$If\ d_L \leq D \leq d_U, inconclusive$$

Lower (d_L) and upper (d_U) bounds can be found in tables online or in textbooks, e.g. Kutner et al (2008).

Appendix 2 Definitions of performance measures

Table 4: Definitions of performance measures. Where θ represents the parameter under investigation, $\hat{\theta}$ being the estimate of that parameter, $\bar{\theta}$ being the mean value of the estimate, n_{sim} being the number of simulations (in this study, 10,000), p_i being the p-value of estimate i and α being the significance level (Morris, White et al. 2019).

Performance measure	Definition	Estimate
Bias	$E[\hat{\theta}] - \theta$	$\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\theta}_i - \theta$
Empirical standard error	$\sqrt{Var(\hat{\theta})}$	$\sqrt{\frac{1}{n_{sim} - 1} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \bar{\theta})^2}$
Mean square error	$E\left[(\hat{\theta}_i - \theta)^2\right]$	$\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \theta)^2$
Coverage	$Pr(\hat{\theta}_{low} \leq \theta \leq \hat{\theta}_{upp})$	$\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1(\hat{\theta}_{low,i} \leq \theta \leq \hat{\theta}_{upp,i})$
Power	$Pr(p_i \leq \alpha)$	$\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1(p_i \leq \alpha)$

References

Alpargu, G. and P. Dutilleul (2003). "Efficiency and Validity Analyses of Two-Stage Estimation Procedures and Derived Testing Procedures in Quantitative Linear Models with AR(1) Errors." Communications in Statistics - Simulation and Computation **32**(3): 799-833.

Box, G. E. P. a. (2016). Time series analysis : forecasting and control, Hoboken, New Jersey : Wiley.

Brennan, S., McDonald, S., McKenzie, J., Cheng, A., Green, S., Allen, K., & Reid, J. (2017). Systematic review of antimicrobial surfaces to reduce infection rates in hospitalised populations. Cochrane Australia. Melbourne, Australia, Cochrane Australia**:** 21.

Cheang, W.-K. and G. C. Reinsel (2000). "Bias Reduction of Autoregressive Estimates in Time Series Regression Model through Restricted Maximum Likelihood." Journal of the American Statistical Association **95**(452): 1173-1184.

Cruz, M., M. Bender and H. Ombao (2017). "A robust interrupted time series model for analyzing complex health care intervention data." Statistics in Medicine **36**(29): 4660-4676.

Durbin, J. and G. S. Watson (1950). "Testing for Serial Correlation in Least Squares Regression: I." Biometrika **37**(3/4): 409-428.

Ewusie, J., C. Soobiah, E. Blondal, J. Beyene, L. Thabane and J. Hamid (2020). "Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review." Journal of Multidisciplinary Healthcare **13**: 411-423.

Gebski, V., K. Ellingson, J. Edwards, J. Jernigan and D. Kleinbaum (2012). "Modelling interrupted time series to evaluate prevention and control of infection in healthcare." Epidemiology and Infection **140**(12): 2131-2141.

Hacek, D. M., A. M. Ogle, A. Fisher, A. Robicsek and L. R. Peterson (2010). "Significant impact of terminal room cleaning with bleach on reducing nosocomial Clostridium difficile." American journal of infection control **38**(5): 350-353.

Hudson, J., S. Fielding and C. R. Ramsay (2019). "Methodology and reporting characteristics of studies using interrupted time series design in healthcare." BMC Medical Research Methodology **19**(1): 137.

Huitema, B. E. (2011). Analysis of covariance and alternatives statistical methods for experiments, quasi-experiments, and single-case studies. Hoboken, N.J., Hoboken, N.J. : Wiley.

Huitema, B. E. and J. W. McKean (2007). "Identifying Autocorrelation Generated by Various Error Processes in Interrupted Time-Series Regression Designs." Educational and Psychological Measurement **67**(3): 447-459.

Jandoc, R., A. M. Burden, M. Mamdani, L. E. Lévesque and S. M. Cadarette (2015). "Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations." Journal of Clinical Epidemiology **68**(8): 950-956.

Kenward, M. G. and J. H. Roger (1997). "Small sample inference for fixed effects from restricted maximum likelihood." Biometrics **53**(3): 983.

Kontopantelis, E., T. Doran, D. A. Springate, I. Buchan and D. Reeves (2015). "Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis." BMJ : British Medical Journal **350**: h2750.

Kutner, M., C. Nachtscheim, J. Neter, W. Li and H. Senter (2008). Applied linear statistical models. M. Kutner, C. Nachtscheim, J. Neter, W. Li and H. Senter. **103:** 880-880.

Lopez Bernal, J., S. Cummins and A. Gasparrini (2016). "Interrupted time series regression for the evaluation of public health interventions: a tutorial." International Journal of Epidemiology: dyw098.

Morris, T. P., I. R. White and M. J. Crowther (2019). "Using simulation studies to evaluate statistical methods." Statistics in Medicine **38**(11): 2074-2102.

Nelson, B. K. (1998). "Statistical methodology: V. Time series analysis using autoregressive integrated moving average (ARIMA) models." Academic emergency medicine : official journal of the Society for Academic Emergency Medicine **5**(7): 739.

Newey, W. K. and K. D. West (1987). "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix." Econometrica **55**: 703.

Penfold, R. B. and F. Zhang (2013). "Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements." Academic Pediatrics **13**(6): S38-S44.

Prais, S. J., Winsten, C.B. (1954). Trend estimators and serial correlation. Cowles Commision. Y. University.

Ramsay, C. R., L. Matowe, R. Grilli, J. M. Grimshaw and R. E. Thomas (2003). "Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies." International Journal of Technology Assessment in Health Care **19**(4): 613-623.

Satterthwaite, F. E. (1946). "An Approximate Distribution of Estimates of Variance Components." Biometrics Bulletin **2**(6): 110-114.

Singer, J. and J. Willett (2003). Applied longitudinal data analysis: modeling change and event occurrence.

Smith, J. and M. McAleer (1994). "Newey-West covariance matrix estimates for models with generated regressors." Applied Economics **26**(6): 635-640.

Stata (2017). Stata Statistical Software. College Station, TX, Statcorp LLC. **15**.

StataCorp (2017). Stata 15 Base Reference Manual. College Station, TX, Stata Press.

Thompson, W. A. (1962). "The Problem of Negative Estimates of Variance Components." The Annals of Mathematical Statistics **33**(1): 273-289.

Turner, S. L. (2020). Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study - Code and Data.

Turner, S. L., A. Karahalios, A. B. Forbes, M. Taljaard, J. M. Grimshaw, A. C. Cheng, L. Bero and J. E. McKenzie (2020). "Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: a review." Journal of Clinical Epidemiology **122**: 1-11.

Wagner, A. K., S. B. Soumerai, F. Zhang and D. Ross-Degnan (2002). "Segmented regression analysis of interrupted time series studies in medication use research." Journal of Clinical Pharmacy and Therapeutics **27**(4): 299-309.

White, I. R. (2010). "Simsum: analyses of simulation studies including Monte Carlo error." Stata Journal **10**: 369+.

Chapter 6. Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series

This chapter presents the findings of an empirical evaluation that compared the effect estimates when different statistical methods were applied to 190 real-world datasets. The simulation study (Chapter 5) provided insight into how statistical methods perform against a known truth (58). The empirical evaluation provides evidence of whether the choice of statistical method matters in practice and the degree in which it may do so.

The ITS identified in the review (Chapter 3) formed the repository of ITS data series included in the empirical evaluation. Time series data were obtained from the studies via different methods; using data that had been published along with the study, via email contact with authors, and via digital data extraction from the graphs included in the studies. A segmented linear regression model was fitted to each series, treating the outcome as continuous, including a single interruption, and allowing for lag-1 autocorrelation. These time series were analysed using the set of statistical methods evaluated in the simulation study (Chapter 5). Pairwise comparisons between the methods were made in terms of estimates of level- and slope-change and their standard errors, CIs and p-values. Estimates of the magnitude of autocorrelation yielded by the various methods were also compared. Implications of the findings were considered for practice and future research. A repository of 184 published and digitally extracted real-world time series was collated and made publicly available on the online repository figshare: https://doi.org/10.6084/m9.figshare.13297136 (59).

Chapter 6 is presented as a manuscript, which has been submitted to *BMC Medical Research Methodology*.

Additional files referred to in the manuscript are appended to this thesis as follows:

Location in thesis	**Referred to in manuscript**	**Content of appendix**
Appendix H	Additional File 1	Computer code to analyse data sets for the empirical evaluation
Appendix I	Additional File 2	Citation details of the 200 studies and the source of dataset collected from each (via publication, email or digital extraction)

Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series

Simon L Turner[1], Amalia Karahalios[1], Andrew B Forbes[1], Monica Taljaard[2,3], Jeremy M Grimshaw[2,3,4], Joanne E McKenzie[1*]

[1]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia.

[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada. 1053 Carling Ave, Ottawa.

[3]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada. 600 Peter Morand Crescent, Ottawa, Ontario K1G 5Z3.

[4]Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada. Roger Guindon Hall, 451 Smyth Rd.

*Corresponding Author:

Joanne McKenzie
mail: Level 4, 553 St. Kilda Road, Melbourne, 3004, Australia
email: joanne.mckenzie@monash.edu
ph: +61 3 9903 0380

Abstract

Background

The Interrupted Time Series (ITS) is a quasi-experimental design commonly used in public health to evaluate the impact of interventions or exposures. Multiple statistical methods are available to analyse data from ITS studies, but no empirical investigation has examined how the different methods compare when applied to real-world datasets.

Methods

A random sample of 200 ITS studies identified in a previous methods review were included. Time series data from each of these studies was sought. Each dataset was re-analysed using six statistical methods. Point and confidence interval estimates for level and slope changes, standard errors, p-values and estimates of autocorrelation were compared between methods.

Results

From the 200 ITS studies, including 230 time series, 190 datasets were obtained. We found that the choice of statistical method can importantly affect the level and slope change point estimates, their standard errors, width of confidence intervals and p-values. Statistical significance (categorised at the 5% level) often differed across the pairwise comparisons of methods, ranging from 4% to 25% disagreement. Estimates of autocorrelation differed depending on the method used and the length of the series.

Conclusions

The choice of statistical method in ITS studies can lead to substantially different conclusions about the impact of the interruption. Pre-specification of the statistical method is encouraged, and naive conclusions based on statistical significance should be avoided.

Keywords

Autocorrelation, Interrupted Time Series, Public Health, Segmented Regression, Statistical Methods, Empirical study

1 Background

Randomised trials are the gold standard design for investigating the impact of public health interventions, however, they cannot always be used. For example, interventions that impact an entire country, or those that have occurred historically, may preclude the ability to randomize or include control groups (1). An alternative non-randomised design that may be considered in such circumstances is an interrupted time series (ITS) (2-4). In an ITS design, data are collected at multiple time points both before and after an interruption (i.e. an intervention or exposure). Modelling of the data in the pre-interruption period allows estimation of the underlying secular trend, which when modelled correctly and extrapolated into the post-interruption time period, yields a counterfactual for what would have occurred in the absence of the interruption. Differences between the counterfactual and observed data at various points post interruption can be estimated (e.g. immediate and long-term effects), having accounted for the underlying secular trend.

A characteristic of data collected over time is that the data points tend to be correlated (5). This correlation – referred to as autocorrelation or serial correlation – can be positive (whereby data points close together in time are more similar than data points further apart) or, infrequently, negative (whereby data points close together are more dissimilar than data points further apart). Autocorrelation may be observed between consecutive data points or over longer periods of time (e.g. seasonal effects). This characteristic of the data needs to be considered when designing and analysing ITS studies. If positive autocorrelation is present, larger sample sizes are required to provide power at the desired level (6) and if autocorrelation is not accounted for in the statistical analysis, standard errors may be underestimated (7).

Segmented linear regression models are often fitted to ITS data using a range of estimation methods (8-11). Commonly ordinary least squares (OLS) is used to estimate the model parameters (10); however, the method does not account for autocorrelation. Other statistical methods are available that attempt to account for autocorrelation in different ways (e.g. correction of standard errors, directly modelling the errors).

Turner et al undertook a statistical simulation study examining the performance of statistical methods for analysing ITS data, where the methods were those commonly used in practice or had shown potential to perform well (12). This simulation study provided insight into how these statistical methods performed under different scenarios, including different level and slope changes, varying magnitudes of underlying autocorrelation and series lengths. In combination with these findings, evidence from an empirical evaluation can provide a more comprehensive understanding of how the methods operate. In particular, empirical evaluations – in which methods are applied to real-world data sets and the results are compared – allow assessment of whether the choice of method matters in practice, and the degree to which they may do so.

To our knowledge, there has been no study that has empirically compared different methods for analysing ITS data when applied to a large sample of real-world data sets. We therefore undertook such an evaluation, where we aimed to compare level and slope change estimates, their standard errors, confidence intervals and p-values, and estimates of autocorrelation, obtained from the set of statistical methods used in the Turner et al simulation study (12).

2 Methods

2.1 Repository of ITS studies

A sample of 200 ITS studies identified in a previous methods review were eligible for inclusion in the current study (10). In brief, we randomly selected ITS studies indexed on PubMed between the years 2013 to 2017. The criteria for inclusion were: 1) studies in which there were at least two segments separated by a clearly defined interruption with at least three points in each segment; 2) observations were collected on a group of individuals at each time point; and 3) the study investigated the impact of an interruption that had public health implications.

For each of the 200 studies, the first reported ITS of each outcome type (binary, continuous, count or proportion) was included, resulting in 230 ITS. Data were collected on the study characteristics and design of the ITS studies, types of outcomes, models used, statistical methods employed, effect measures reported, and the properties of included graphs. Further details of the study methods are available in the study protocol and results papers (10, 13).

2.2 Methods to obtain time series data

Time series data from the included studies were obtained using three methods. First, we collated datasets that were reported in the published paper or its supplement (e.g. time series data reported in tables, or as text files). Second, we contacted all authors for whom we were able to obtain contact details to request datasets. We requested only aggregate level data (i.e. not individual participant data) and in the circumstance where a study included multiple series, we only sought data from the first time series reported in the paper to reduce respondent burden.

We sent an initial email request on the 13th December 2018 and a follow-up email on the 24th January 2019. Third, we digitally extracted datasets from published graphs using the software WebPlotDigitizer (14). This graphical data extraction tool has been found to accurately estimate the position of points on a graph (15).

If multiple datasets from the above methods were available for a particular time series, we selected the dataset generated using the following hierarchy: (i) published data, (ii) contact with authors, and (iii) digitally extracted. We checked the data provided by authors against the information reported in the publication. Where there was a discrepancy, we re-contacted the authors to query the provided data.

2.3 Interrupted time series model

We fitted segmented linear regression models to each dataset using the parameterisation of Huitema and McKean (7) (Equation1, Figure 1):

$$Y_t = \beta_0 + \beta_1 t + \beta_2 D_t + \beta_3 [t - T_I] D_t + \varepsilon_t \qquad (1)$$

where Y_t represents the outcome that is measured at time point t of N time points (1 to n_1 measurements during the pre-interruption stage, and $n_1 + 1$ to n_2 measurements in the post-interruption stage), with the interruption occurring at time T_I. D_t is an indicator variable that represents the post-interruption interval: coded as 0 in the pre-interruption period, and as 1 in the post-interruption period. The model parameters (βs) represent the baseline intercept (β_0); pre-interruption slope (β_1); change in level at the interruption (β_2), and the change in slope (β_3). The model can be extended to accommodate more than one interruption with the inclusion of terms representing additional segments.

The error term ε_t allows for deviation from the fitted model. In a first order (lag-1) autocorrelation model, the error at time point t (ε_t) is influenced by only the previous data point as $\varepsilon_t = \rho\varepsilon_{t-1} + w_t$, where ρ is the magnitude of autocorrelation (ranging from -1 to 1) and w_t represents normally distributed “white noise” $w_t \sim N(0, \sigma^2)$. Longer lags can be modelled or accommodated, but here we restrict our focus to lag-1.

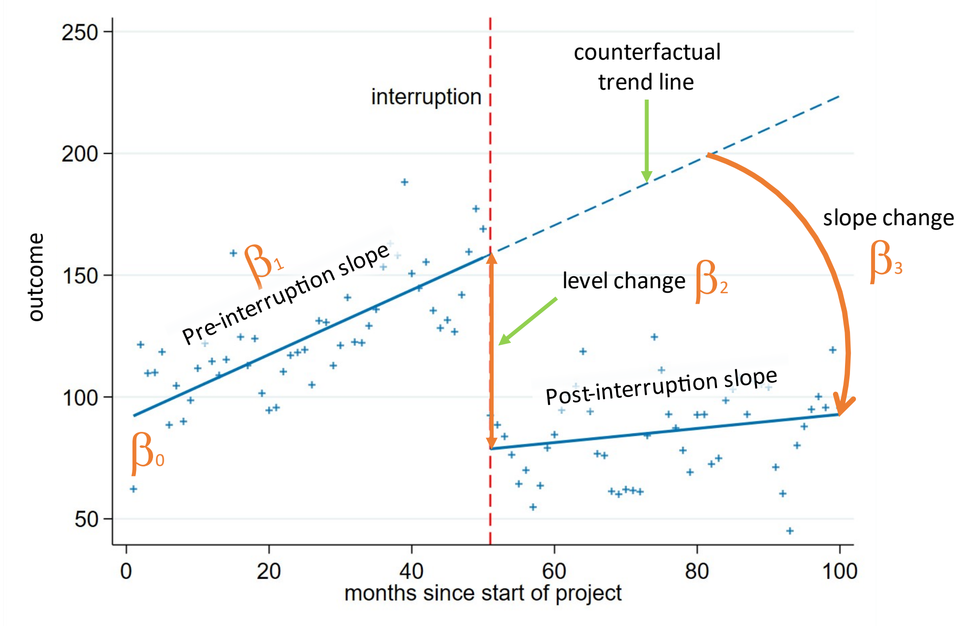


Figure 1: Graphical depiction of a segmented linear regression model fitted to ITS data. Secular trends (indicated by solid blue lines) for the pre and post interruption periods (indicated by the vertical dashed line) are estimated from the data (indicated by blue crosses). A counterfactual trend line (extrapolation of the pre-interruption trend line shown as a dashed blue line) is compared with the post interruption trend to estimate the immediate and longer term impact of the interruption. Model parameters are indicated as the intercept (β_0); pre-interruption slope (β_1); change in level at the interruption (β_2), and the change in slope (β_3).

2.4 Interrupted time series analysis methods

Six statistical methods were used to analyse the ITS datasets assuming first order autocorrelation (lag-1) (Table 1). The methods were chosen because they have commonly been used in practice (8-11) or because of they have been shown (through numerical simulation) to have improved performance (12). The methods were:

- ordinary least squares regression (OLS), which provides no adjustment for autocorrelation, and in the presence of positive autocorrelation will yield standard errors that are too small (16);
- OLS with Newey-West standard errors (NW), which yield OLS estimates of the model regression parameters, but with standard errors that are adjusted for autocorrelation (17);
- Prais-Winsten (PW), a generalised least squares method, which provides an extension of OLS where the assumption of independence across observations is relaxed (18, 19);
- restricted maximum likelihood (REML) (with and without the small sample Satterthwaite approximation (Satt)), which addresses bias in maximum likelihood estimators of variance components by separating the log-likelihood into two terms (one of which is only dependent on variance parameters) and using the appropriate number of degrees of freedom (d.f.) (20, 21); and,
- autoregressive integrated moving average (ARIMA), which explicitly models the influence of previous time points by including regression coefficients from lagged values of the dependent variable and errors (22).

Table 1: Statistical methods, adjustments for autocorrelation and abbreviations used.

Statistical method	Autocorrelation adjustment	Abbreviation
Ordinary least squares	None	OLS
	Newey-West standard error adjustment with lag-1 autocorrelation	NW
Generalised least squares	Prais-Winsten	PW
Restricted maximum likelihood	Lag-1 autocorrelation model	REML
	Lag-1 autocorrelation model with small sample Satterthwaite approximation	REML-Satt
Autoregressive integrated moving average	Lag-1 autocorrelation model (i.e. ARIMA(1,0,0))	ARIMA

2.5 Analysis of the ITS datasets

We implemented the segmented linear regression model (Equation 1, Section 2.3) by setting up datasets for each ITS study with the following variables:

- outcome variable;
- time variable t, beginning at 1 and incrementing by 1 up to time point N;
- an interruption time indicator D_t; coded 0 pre-interruption and 1 post-interruption; and,
- a slope change variable $[t - T_I]D_t,$ equal to zero at the time of the interruption (T_I) and incrementing by 1 up to time point N.

We used information provided in the corresponding manuscript to determine the interruption time. In studies with multiple interruptions, we only included the first interruption (and adjacent periods). In studies with a transition period, we extended the model to include an additional segment for the transition period; however, when calculating the level and slope changes, we ignored this segment (further details available in Appendix 1).

We analysed each dataset using the six estimation methods described in Section 2.4. For REML with the Satterthwaite approximation, when the computed degrees of freedom were less than two, we substituted these with the value two to avoid overly conservative confidence limits and hypothesis tests. We only included analyses for which the estimate of autocorrelation was strictly between -1 and +1. The datasets were analysed in Stata 15 (23) (see Additional File 1 for analysis code).

2.6 Comparison of results from the different ITS analysis methods

The results of interest were point estimates of the immediate level change (β_2) and slope change (β_3), their associated standard errors, confidence intervals and p-values, and the estimated lag-1 autocorrelation. Across the ITS studies, different outcomes were measured, necessitating the need to standardise the estimates of slope and level change for comparison across the datasets. This was achieved for each dataset by dividing parameter estimates by the root mean square error (RMSE) estimated from a segmented linear regression model using OLS. We also standardised the direction of effect. This was achieved for each pairwise comparison of methods by multiplying both estimates by -1 if the first method's estimate was less than zero. We also repeated these analyses standardising to the direction of the second method's estimate.

2.6.1 Estimates of level and slope changes, and their standard errors

We compared the level and slope change point estimates with their standard errors using visual displays and tabulation. Specifically, we used Bland Altman scatter plots (24) to assess pairwise agreement in the results (standardised estimates of level change, slope change, and their standard errors) between the different statistical methods. For each pairwise comparison, the difference in the two estimates was plotted against the average of the two estimates (e.g. 'difference in estimates of level change from OLS and PW' versus 'average of estimates of level change from OLS and PW'). In the case of the standard errors, we first log-transformed these to remove the relationship between the variability of the differences and the magnitude of the standard errors (24). The mean difference and limits of agreement (average difference ±1.96×standard deviation of the differences) were calculated and overlaid on the plots. These pairwise comparisons were displayed in a matrix of plots to show comparisons of each method with all others. Plots in the top triangle of the matrix illustrate agreement between the effect estimates (either level change or slope change), and plots in the bottom triangle illustrate the agreement between the standard errors.

We also investigated whether series length impacted the difference in level and slope change estimates between each pair of methods. A matrix of scatterplots of the differences versus the (log) length of series (overlaid with a local regression (LOESS) smoothed curve) for each pairwise method comparison was used to visually examine this relationship.

2.6.2 Confidence Intervals

We visually compared the width of the confidence intervals from the different statistical methods. For each dataset and pairwise comparison, a ratio of the confidence interval widths from the two methods was calculated and then scaled so that the comparison method confidence interval spanned -0.5 to 0.5.

2.6.3 p-values

We compared the p-values of the effect estimates between the methods by categorising the p-values based on commonly used levels of statistical significance. First, we categorised the p-values at the 5% level of statistical significance (i.e. < 5%, ≥ 5%), and second, we categorised p-values using a finer gradation (i.e. p-value < 1%, 1% ≤ p-value < 5%, 5% ≤p-value < 10%, p-value ≥ 10%). For each pairwise comparison between methods, we calculated the percentage of datasets where there was agreement in the categories of statistical significance (i.e. the percentage of datasets where the p-value for the effect estimate was < 0.05 for both methods *or* the p-value was ≥ 0.05 for both methods). Further, we calculated kappa statistics to assess agreement beyond chance. We use the following adjectives when describing the results: 0.41-0.6 moderate agreement, 0.61-0.8 substantial agreement, 0.81-1.0 almost perfect agreement (25).

2.7 Autocorrelation coefficient estimates

We calculated and tabulated medians and interquartile ranges for estimates of lag-1 autocorrelation for the three methods that yield these estimates (ARIMA, PW, REML). The summary statistics are reported for all series as well as being restricted to series with ≥ 24 points and series with ≥ 100 points, in order to assess whether series length impacted the magnitude of the estimates. A scatterplot of autocorrelation versus (log) length of series (overlaid with a LOESS curve) was used to visually examine this relationship. A further scatter plot was generated that depicted the REML estimates of autocorrelation along with their confidence intervals.

3 Results

3.1 Time series dataset acquisition

Of the 230 ITS identified in the review (10) we obtained 10/230 (4%) datasets directly from the publication (e.g. time series data reported in tables), 50/230 (22%) through email contact with the authors, and 184/230 (80%) through digital data extraction. For some series (n = 47), multiple datasets from the different sources were available (Figure 2). Using our hierarchy for selecting the source of the dataset when multiple series were available resulted in 190 unique datasets, with 8/190 (4%) sourced directly from the publication, 45/190 (24%) through email contact with authors, and 137/190 (72%) from digital data extraction.

We were unable to obtain 40 of the 230 ITS included in the review because the data were not reported in the paper, could not be obtained from authors, or could not be digitally extracted. For two datasets, a segmented linear regression model was not appropriate to fit, and these were excluded. Five of the datasets obtained from the authors could not be used: three due to errors in the data; two as segmented linear regression models were not appropriate to fit. Forty-six of the datasets could not be digitally extracted, 27 studies included graphs with insufficient resolution to digitally extract data; 8 studies had no graph; 8 studies had summary data only (e.g. a summary graph showing a small number of annual figures was provided when monthly data was used in the analysis); and 3 studies had graphs but did not plot data points.

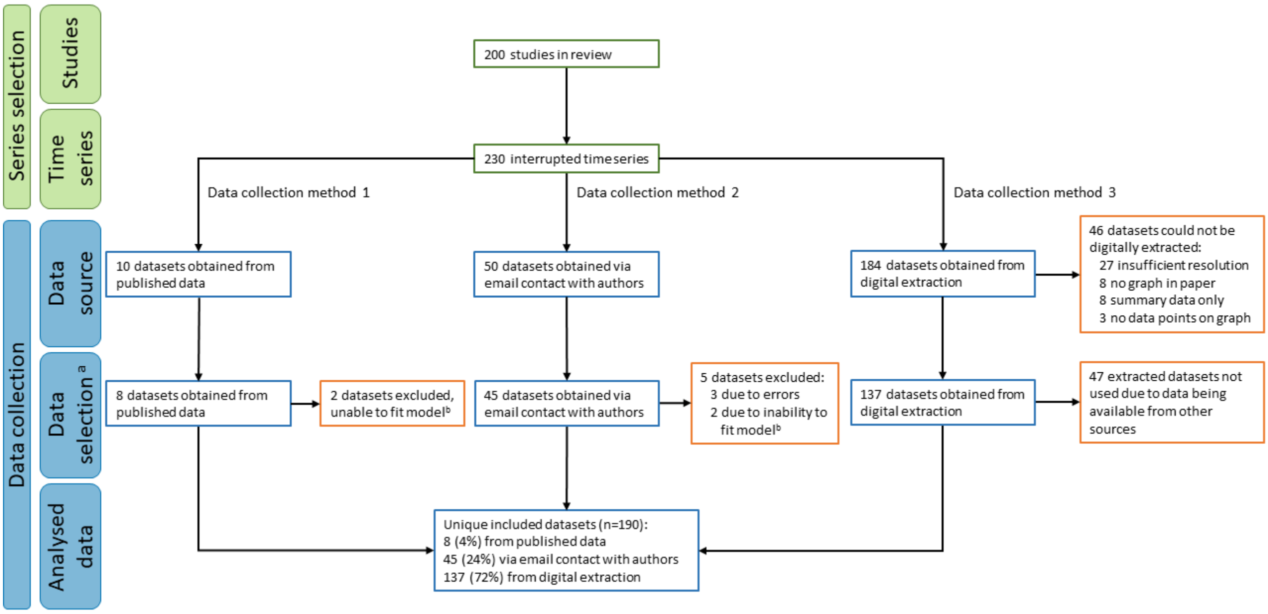

Series selection
Studies
Time series
Data collection
Data source
Data selection a
Analysed data
200 studies in review
230 interrupted time series
Data collection method 1
Data collection method 2
Data collection method 3
10 datasets obtained from published data
50 datasets obtained via email contact with authors
184 datasets obtained from digital extraction
46 datasets could not be digitally extracted:
27 insufficient resolution
8 no graph in paper
8 summary data only
3 no data points on graph
8 datasets obtained from published data
2 datasets excluded, unable to fit model b
45 datasets obtained via email contact with authors
5 datasets excluded:
3 due to errors
2 due to inability to fit model b
137 datasets obtained from digital extraction
47 extracted datasets not used due to data being available from other sources
Unique included datasets (n=190):
8 (4%) from published data
45 (24%) via email contact with authors
137 (72%) from digital extraction

Figure 2: Flowchart of selected datasets. Green boxes denote the number of included studies and time series, blue boxes denote the numbers corresponding to dataset collection, and orange boxes denote the numbers corresponding to dataset exclusion.

[a] *Hierarchy for data selection was (i) published data, (ii) contact with authors, and (iii) digital extraction*

[b] *An appropriate segmented linear regression model could not be used for some datasets*

3.2 Characteristics of the included ITS

The characteristics of the ITS studies with available datasets for re-analysis are compared to all 200 ITS studies in Table 2. No major differences were found. The types of study interventions were similar, as were the types of time intervals. The number of time points per series were lower in the studies with available datasets than in all ITS studies (median 41, Q1-Q3 25 to 71) versus 48 (30 to 100)). The length of the segments used to calculate the estimates for the first interruption were slightly shorter in the series with available data than in all series (16, IQR (10,28) versus 18 IQR (10, 34)).

Table 2: Characteristics of interrupted time series studies and series

Study level characteristics	**All ITS studies (n = 200)**		**ITS studies with available data (n = 166)[a]**	
	n	**%**	**n**	**%**
Type of interruption				
Exposure[a]	12	6	10	6
Intervention	188	94	180	95
Intervention type				
policy change	104	52	81	49
practice change	40	20	36	22
communication	29	15	24	14
organisation of care	13	7	12	7
clinical intervention	2	1	2	1
Time interval type				
daily	3	2	2	1
weekly	9	5	6	4
two weekly	1	1	1	1
monthly	120	60	96	58
quarterly	31	16	28	17
six monthly	3	2	3	2
annually	20	10	17	10
other	12	6	12	7
can't determine	1	1	1	1
Series level characteristics	**ITS (n=230)**		**ITS with available data (n=190)**	
	median	**IQR**	**median**	**IQR**
Number of time points per series	48	(30, 100)	41	(25, 71)
Number of time points in the segments used to calculate estimates for the first interruption	18	(10, 34)	16	(10, 28)

Abbreviation: ITS, interrupted time series; IQR, inter-quartile range.

[a] Our definition of an exposure is limited to exposures or events that are not under investigator control (e.g. earthquakes, financial crises, tsunamis, environmental chemicals). We use the term 'investigator' loosely to include researchers, clinicians and policy makers.

3.3 Comparison of results from the different ITS analysis methods

3.3.1 Estimates of level and slope changes, and their standard errors

The median values of the absolute value of the standardised effect estimates for level change ranged from 1.22 to 1.49 across the statistical methods (Table 3). For slope change, the median value of the absolute value of the standardised effect estimates was 0.13 for all statistical methods (Table 3). Pairwise comparisons were limited to a minimum of 171 datasets because at least one statistical method failed to converge, failed to yield standard errors or estimated the magnitude of autocorrelation to be outside the range -1 to +1 in 19 of the datasets (Table 4).

Table 3: Effect estimate summaries. The NW estimates are the same as OLS and the REML-Satt estimates are the same as REML, so these are not presented.

	N	Absolute value of effect estimate Level change Median (IQR)	 Slope change Median (IQR)
ARIMA	189	1.40 (0.63,2.90)	0.13 (0.05,0.26)
OLS	190	1.49 (0.60,3.03)	0.13 (0.06,0.27)
PW	189	1.33 (0.57,2.81)	0.13 (0.05,0.26)
REML	181	1.22 (0.47,2.56)	0.13 (0.05,0.25)

Abbreviations: IQR, interquartile range; ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; PW, Prais-Winsten; REML, restricted maximum likelihood.

Table 4: Number of available comparisons for the statistical methods investigated (n=190).

Number of comparisons	ARIMA	OLS	NW	PW	REML	REML-Satt
ARIMA	189	189	188	185	175	175
OLS		190	189	186	175	175
NW			189	186	174	174
PW				186	171	171
REML					175	175
REML-Satt						175

Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, OLS with Newey-West standard error adjustments; PW, Prais-Winsten; REML, restricted maximum likelihood; REML-Satt, restricted maximum likelihood with Satterthwaite small sample adjustment.

Pairwise comparisons of level change, slope change, and their standard errors for each of the five methods were made (Figures 3 and 4). REML with the Satterthwaite approximation was excluded from these comparisons because it only adjusts the width of the confidence intervals, and not the standard errors. There were small systematic differences in estimates of level change in the pairwise comparisons between the methods, REML had slightly smaller and OLS slightly larger effect estimates than the other methods (Figure 3, top triangle, and Table 5). The largest limits of agreement between all methods (REML vs OLS) were ±1.11. Expectedly, there was no difference in the standardised level change estimates between OLS and NW (since they use the same estimator for β_2) and a very small difference between PW and ARIMA (since their point estimation methods are almost equivalent). There were no systematic differences in slope change estimates between the methods (Figure 4, top triangle and Table 6). Limits of agreement for slope change were generally similar across the pairwise comparisons of methods (but again with the exceptions of the comparison between OLS and NW, and PW and ARIMA).

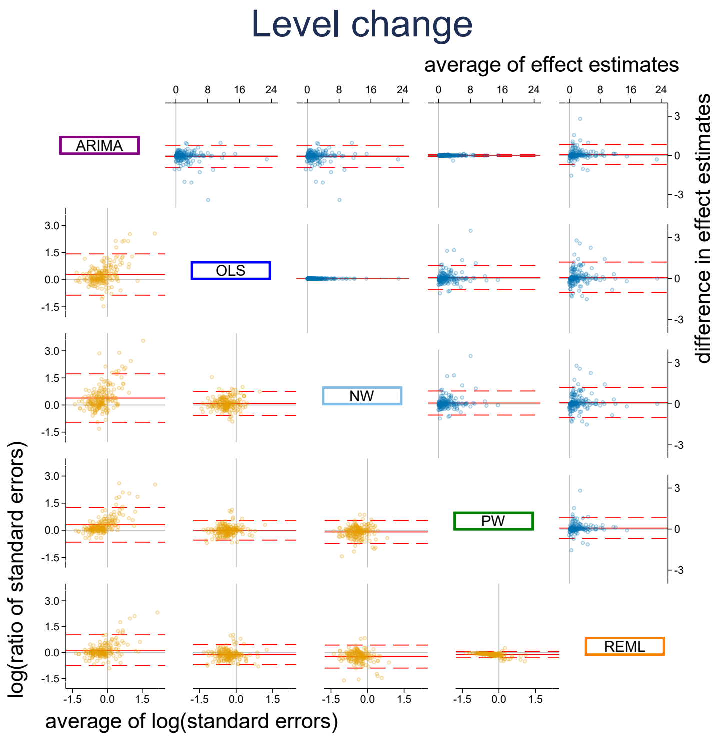
There were systematic differences in the estimates of standard error of level change across some pairwise comparisons of methods (Figure 3, bottom triangle, and Table 5). Notably, the ARIMA standard errors were systematically larger compared with all other methods; however, this difference was smaller when compared with REML (geometric mean ratio standard errors for level change of 1.15). Aside from the pairwise comparison between PW and REML, the limits of agreement between the methods showed that the methods could yield large differences in the standard errors, particularly so for ARIMA compared with the other methods. For example, the limits of agreement for ARIMA compared with NW showed that the differences in standard errors could be large, ranging from 61% smaller to 460% larger. Similar patterns were observed for slope change (Figure 4 bottom triangle, and Table 6).

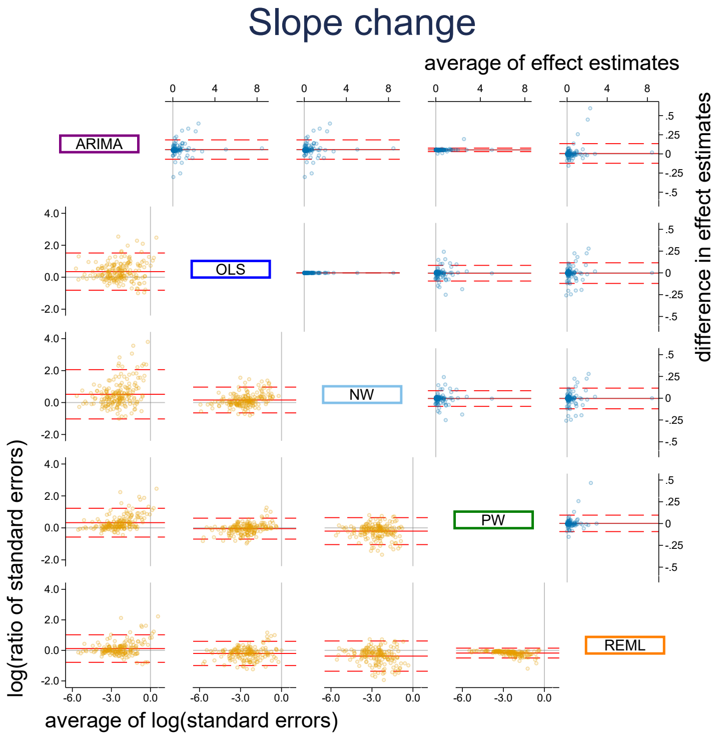
Table 5: Mean of differences in level change estimates between methods (row method - column method) (top triangle) and geometric mean ratio of standard errors for level change between methods (column method/row method) (shaded bottom triangle) with 95% limits of agreement. The OLS and NW level change estimates are the same, so the difference is not presented. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, OLS with Newey-West standard error adjustments; PW, Prais-Winsten; REML, restricted maximum likelihood. Note that REML with the Satterthwaite approximation is not presented because it only makes an adjustment to the confidence intervals, and not the standard errors.

Level Change	**Mean of differences in level change estimates between methods (95% limits of agreement)**				
Geometric mean ratio of standard errors for level change between methods	**ARIMA**	-0.08 (-0.94,0.78)	-0.08 (-0.94,0.78)	0.00 (-0.07,0.07)	0.07 (-0.69,0.84)
	1.34 (0.43,4.18)	**OLS**	-	0.07 (-0.81,0.95)	0.09 (-1.02,1.21)
	1.47 (0.39,5.60)	1.09 (0.56,2.11)	**NW**	0.07 (-0.81,0.95)	0.09 (-1.02,1.21)
	1.35 (0.52,3.53)	0.99 (0.58,1.69)	0.91 (0.48,1.71)	**PW**	0.07 (-0.69,0.82)
	1.15 (0.47,2.80)	0.89 (0.50,1.59)	0.79 (0.41,1.55)	0.89 (0.74,1.08)	**REML**

Table 6: Mean of differences in slope change estimates between methods (row method - column method) (top triangle) and geometric mean ratio of standard errors for level change between methods (column method/row method) (bottom triangle) with 95% limits of agreement. The OLS and NW slope change estimates are the same, so the difference is not presented. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, OLS with Newey-West standard error adjustments; PW, Prais-Winsten; REML, restricted maximum likelihood. Note that REML with the Satterthwaite approximation is not presented because it only makes an adjustment to the confidence intervals, and not the standard errors.

Slope Change	**Mean of differences in slope change estimates between methods (95% limits of agreement)**				
Geometric mean ratio of standard errors for slope change between methods	**ARIMA**	0.00 (-0.11,0.12)	0.00 (-0.11,0.12)	0.00 (-0.02,0.02)	0.01 (-0.12,0.13)
	1.40 (0.44,4.46)	**OLS**	-	0.00 (-0.09,0.09)	0.00 (-0.12,0.12)
	1.68 (0.36,7.90)	1.19 (0.53,2.65)	**NW**	0.00 (-0.09,0.09)	0.00 (-0.12,0.12)
	1.38 (0.56,3.38)	0.96 (0.50,1.86)	0.81 (0.35,1.91)	**PW**	0.00 (-0.09,0.10)
	1.12 (0.45,2.80)	0.82 (0.37,1.84)	0.68 (0.25,1.86)	0.84 (0.59,1.19)	**REML**





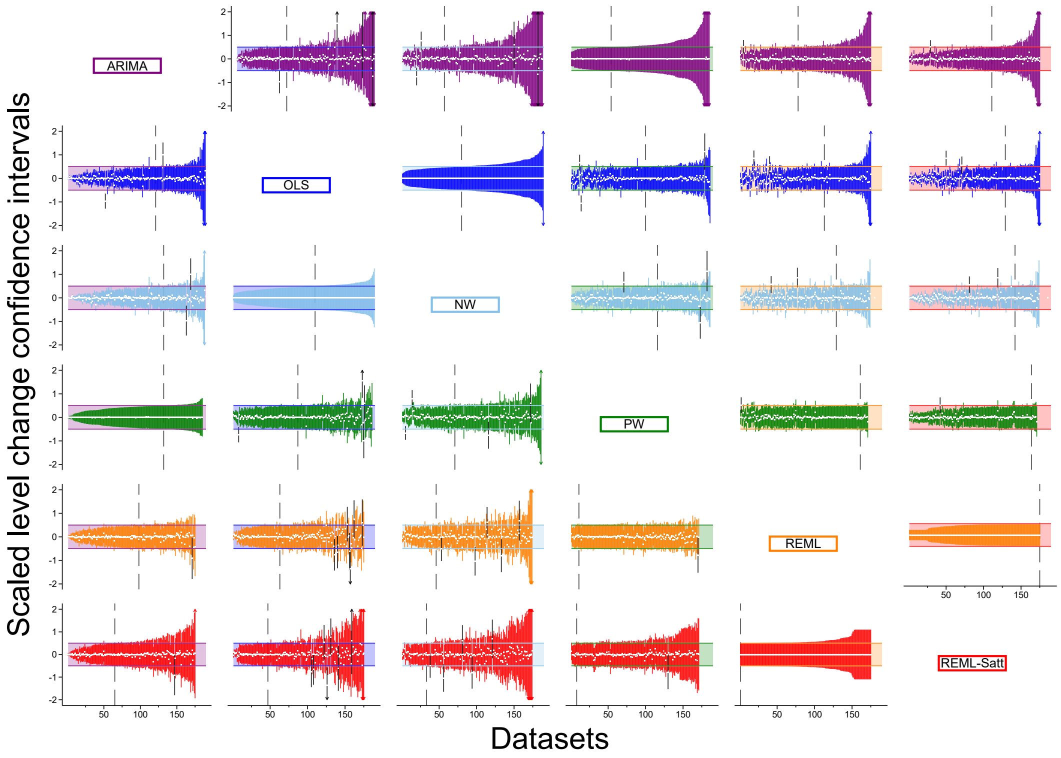
*Figures 3 and 4: Bland Altman plots of standardised level change (Figure 3) and slope change (Figure 4). Plots in the top triangle (blue points) show the difference in point estimates (row method – column method) on the vertical axis and average of the parameter estimates on the horizontal axis. Plots in the bottom triangle (orange points) show differences in standard errors on the vertical axis (= log(ratio of standard errors)) (column method – row method) and the average of the log of the standard errors on the horizontal axis. Red horizontal lines depict the average, red dashed lines depict the 95% limits of agreement (calculated as the average ±1.96*standard deviation of the differences). Grey lines indicate zero. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, OLS with Newey-West standard error adjustments; PW, Prais-Winsten; REML, restricted maximum likelihood. Note that REML with the Satterthwaite approximation is not presented because it only makes an adjustment to the confidence intervals, and not the standard errors.*

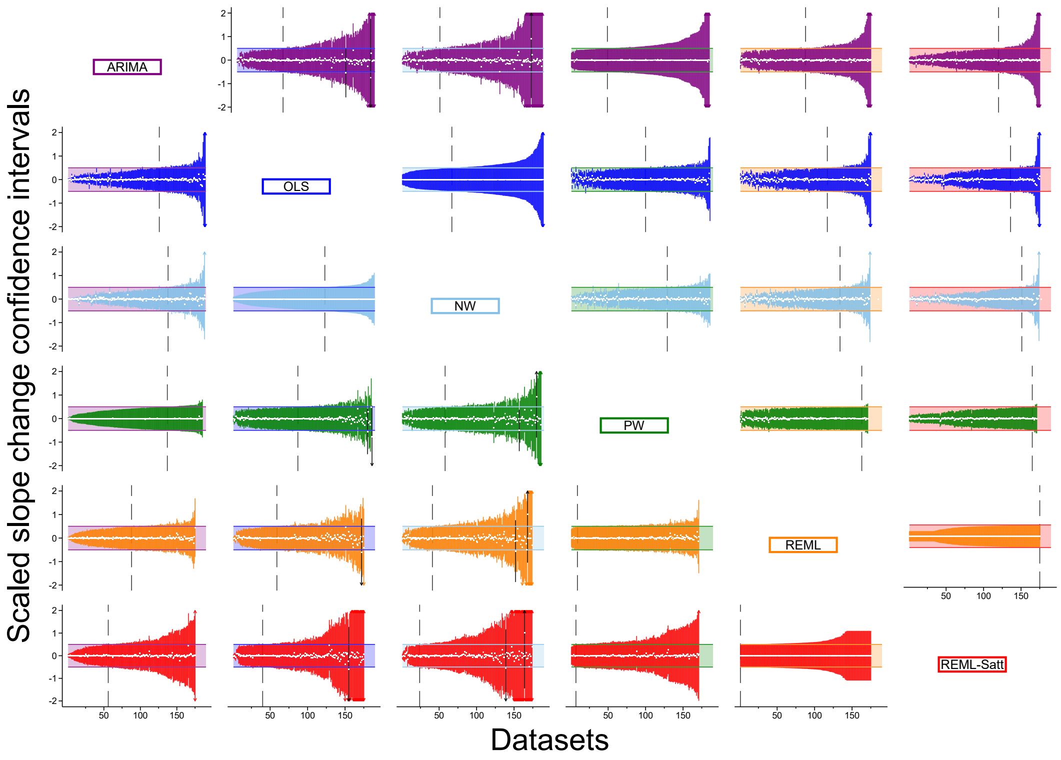
Our visual examination of the impact of series length on the differences in level change estimates between pairs of methods showed that series length was not associated with the differences, with the exception of comparisons with the REML method. For these comparisons, the variability of the differences decreased for longer series (Appendix 2). The variability in differences in slope change estimates for all pairwise comparisons between methods (except between ARIMA and PW), tended to decrease with increasing series length.

When we repeated the analysis standardising the direction of effect to the second method's estimate, we found the results did not importantly change (Appendix 3).

3.3.2 Confidence Intervals

Pairwise comparisons of the confidence interval width for the estimated level change between the methods reflected the patterns observed when comparing the standard errors (Figure 5). ARIMA generally yielded wider confidence intervals with 64%, 70% and 71% of the ARIMA confidence intervals being wider than OLS, NW and PW respectively. ARIMA confidence intervals widths were similar to REML. REML with the Satterthwaite confidence interval adjustment yielded the widest confidence intervals of all methods; only 37% of ARIMA confidence intervals were wider than REML with Satt. This pattern was also seen when comparing the confidence interval widths for the estimated slope change between the methods (Figure 6).





Figures 5 and 6: Pairwise confidence interval comparisons for level change (Figure 5) and slope change (Figure 6). Each plot displays up to 190 confidence intervals (CIs) (depicted as vertical lines), with each scaled so that the confidence interval from the reference method spans -0.5 to 0.5 (shaded area). The reference method is the column method (e.g. the plot in the second row, first column shows OLS CIs (blue) compared to ARIMA (purple)). Vertical lines falling entirely within the shaded area have smaller confidence intervals than the comparison (left of the vertical dashed line), while lines extending beyond the shaded area have larger confidence intervals than the comparison (right of the vertical dashed line). White dots indicate the point estimate. Black vertical lines indicate scenarios in which the point estimate from one method does not lie within the confidence interval of the other. Abbreviations: ARIMA, autoregressive integrated moving average, purple; OLS, ordinary least squares, blue; NW OLS with Newey-West standard error adjustments, light blue; PW, Prais-Winsten, light green; REML, restricted maximum likelihood, orange; REML-Satt, restricted maximum likelihood with Satterthwaite small sample adjustment, red.

3.3.3 p-values

The percentage agreement in statistical significance (dichotomised at the 5% significance level) for level change in the pairwise comparisons between methods ranged from 79.3% (NW versus REML-Satt) to 97.1% (PW versus REML) (Table 7). Corresponding kappa statistics ranged from 0.59 (moderate agreement) for NW versus REML-Satt to 0.94 (almost perfect agreement) for PW versus REML. Discordance in statistical significance in comparisons with REML-Satt and ARIMA arose because these methods yielded larger p-values (Figure 7). For example, in the comparison of NW with REML-Satt, 20% of NW analyses yielded a p-value ≤ 0.05 when the REML-Satt p-value was > 0.05, while only 1% of NW analysis yielded a p-value > 0.05 when the REML-Satt p-value was ≤ 0.05.

In general, the agreement was less for slope change compared with level change (Table 8). The percentage agreement in statistical significance (at the 5% significance level) for slope change in the pairwise comparisons between methods ranged from 75.3% (NW versus REML-Satt) to 93.6% (PW versus REML). Corresponding kappa statistics ranged from 0.50 (moderate agreement) for NW versus REML-Satt to 0.87 (almost perfect agreement) for PW versus REML. The direction of disagreement was similar to that of level change with ARIMA and REML-Satt methods yielding larger p-values more often than the other methods (Figure 8).

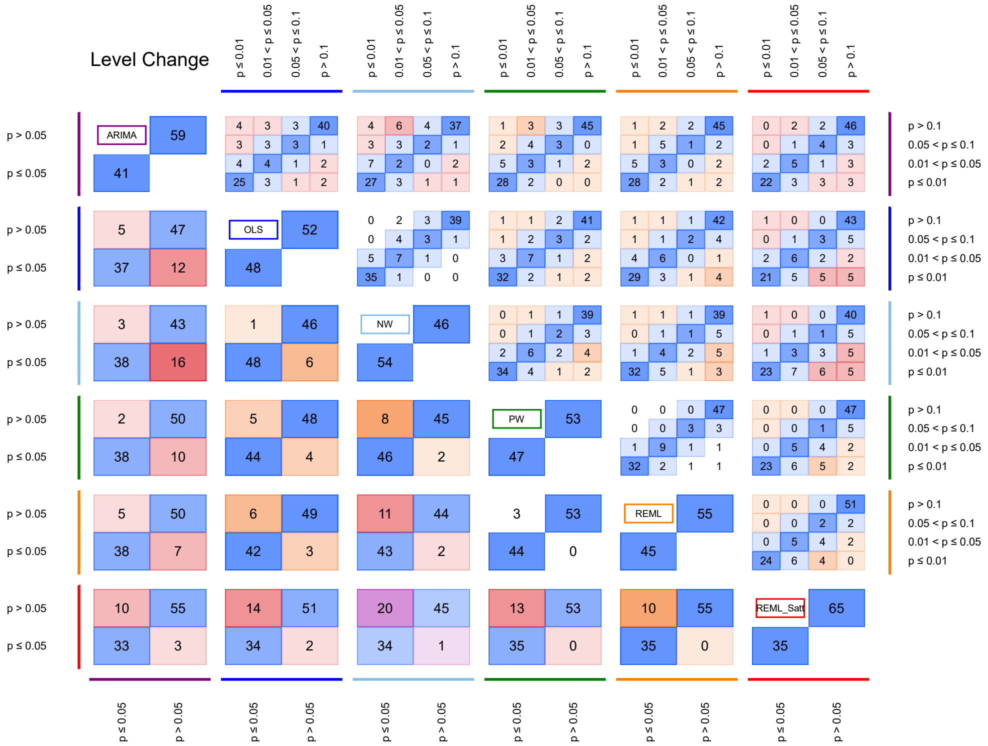
Table 7: Pairwise agreement in statistical significance of estimates of level change between statistical methods. P-values associated with estimates of level change were categorised at the 5% level of statistical significance (i.e. <5%, ≥5%). Cells in the upper triangle contain the percentage of series for which the p-value for level change was < 0.05 for both methods or the p-value was ≥0.05 for both methods. Denominators are reported in Table 4. Cells in the lower triangle (shaded) contain kappa statistics. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW OLS with Newey-West standard error adjustments; PW, Prais-Winsten; REML, restricted maximum likelihood; REML-Satt, restricted maximum likelihood with Satterthwaite small sample adjustment.

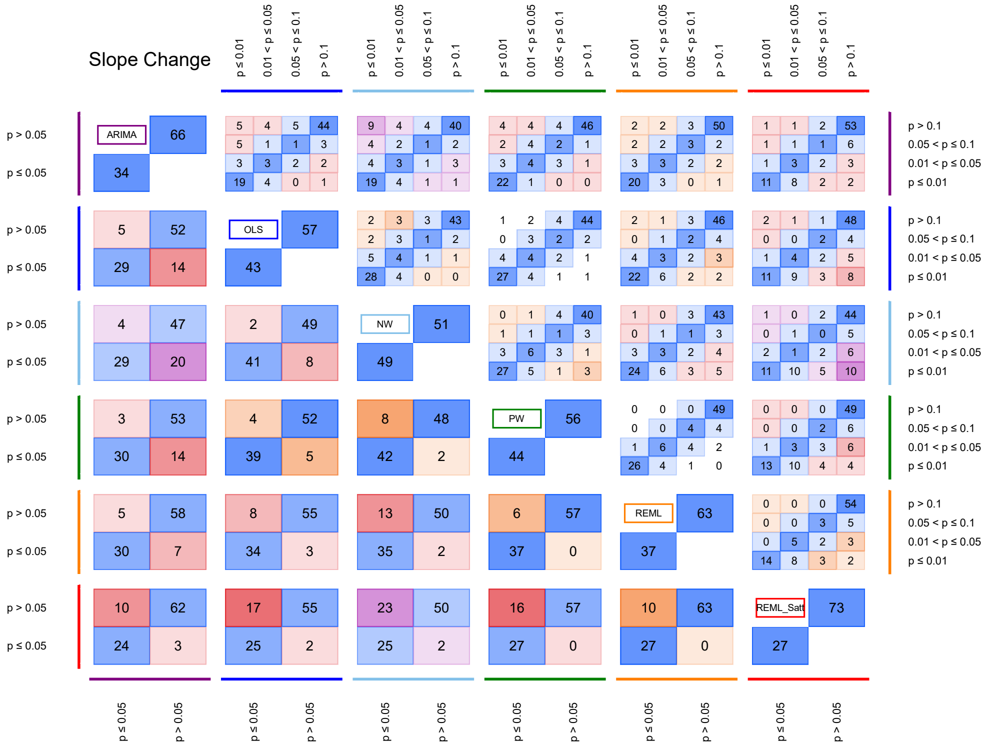
Level Change	**ARIMA**	83.1%	80.3%	88.1%	88.0%	87.4%	**Percentage of agreement**
Kappa statistic	0.66	**OLS**	93.1%	91.4%	90.9%	84.6%	
	0.61	0.86	**NW**	90.3%	86.8%	79.3%	
	0.76	0.83	0.81	**PW**	97.1%	87.1%	
	0.76	0.82	0.74	0.94	**REML**	90.3%	
	0.74	0.69	0.59	0.74	0.80	**REML-Satt**	

Table 8: Pairwise agreement in statistical significance of estimates of slope change between statistical methods. P-values associated with estimates of level change were categorised at the 5% level of statistical significance (i.e. ≤5%, >5%). Cells in the upper triangle contain the percentage of series for which the p-value for level change was ≤ 0.05 for both methods or the p-value was > 0.05 for both methods. Denominators are reported in Table 4. Cells in the lower triangle (shaded) contain kappa statistics. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW OLS with Newey-West standard error adjustments; PW, Prais-Winsten; REML, restricted maximum likelihood; REML-Satt, restricted maximum likelihood with Satterthwaite small sample adjustment.

Slope Change	**ARIMA**	81.5%	76.1%	82.7%	88.0%	86.3%	**Percentage of agreement**
Kappa statistic	0.61	**OLS**	89.4%	90.3%	89.1%	80.6%	
	0.52	0.79	**NW**	90.3%	85.1%	75.3%	
	0.64	0.80	0.81	**PW**	93.6%	83.6%	
	0.74	0.77	0.70	0.87	**REML**	90.3%	
	0.68	0.58	0.50	0.65	0.78	**REML-Satt**	

Our examination of agreement using a finer gradation of statistical significance categories showed that when there was discordance between methods, this generally occurred in an adjacent category (e.g. one method with a p-value ≤ 0.01 and the comparison method with 0.01 ≤ p-value < 0.05). However, there were some examples where there was discordance in non-adjacent categories. For level change these comparisons were ARIMA versus NW, NW versus REML-Satt, and OLS versus REML and REML-Satt (Figure 7), while for slope change these comparisons were the same, but also with the addition of PW versus REML-Satt (Figure 8). The p-values yielded from ARIMA and REML-Satt were generally larger than the other methods, and by contrast, the p-values for NW, and to a lesser extent OLS, tended to be smaller (Appendix 4).





Figures 7 and 8: Pairwise agreement in statistical significance of estimates of p-value comparisons for level change (Figure 7) and slope change (Figure 8). For plots in the bottom triangle, p-values were categorised at the 5% level of significance (i.e. ≤5%, >5%), while for plots in the top triangle, p-values were categorised using a finer gradation of p-value ≤ 0.01, 0.01<p-value≤0.05, 0.05<p-value≤0.1, p-value>0.1. Each cell within a plot contains the percentage of datasets falling within the row and column defined significance levels. Concordant results are shown in blue. Discordant results are shown as either white (0-5% discordance), orange (5-10% discordance), red (10-20% discordance) or purple (over 20% discordance). For example, comparing ARIMA to OLS (row 2, column 1, bottom triangle) shows that for 12% of the datasets the ARIMA method yields a p-value > 0.05 while the OLS method yields a p-value ≤ 0.05. Numbers may not add to 100 due to rounding. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW OLS with Newey-West standard error adjustments; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite adjustment.

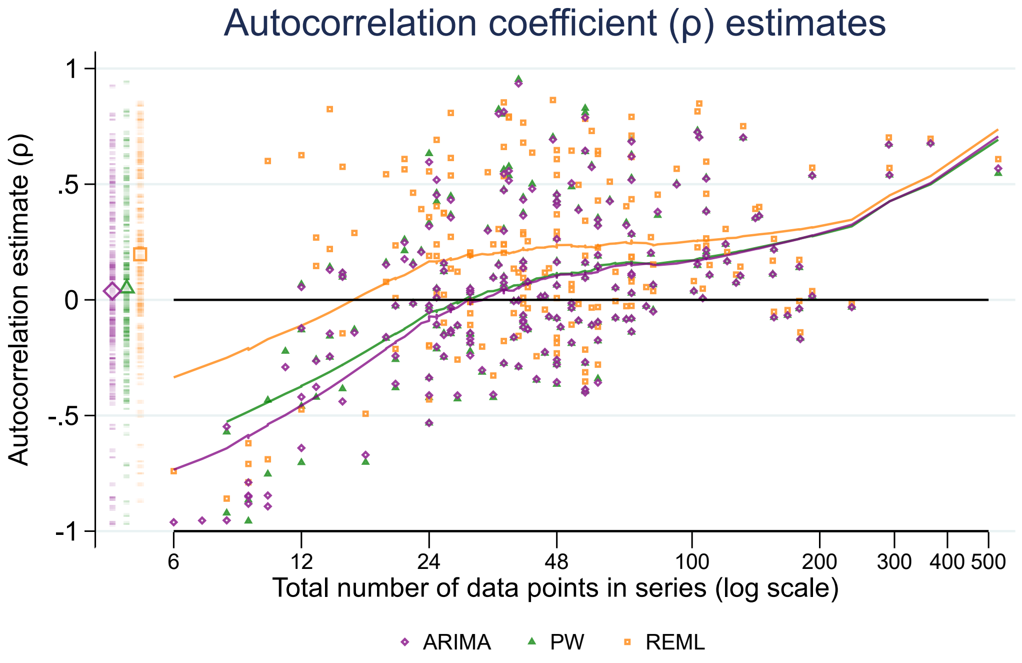
3.4 Autocorrelation coefficient estimates

Three of the statistical methods (ARIMA, PW, REML) yielded estimates of autocorrelation (Table 9, Figure 9). The REML method estimated consistently larger magnitudes of autocorrelation than the other methods (median and inter-quartile range (IQR) of 0.2 (-0.01, 0.54) compared with 0.04 (-0.15, 0.30) for ARIMA and 0.05 (-0.14, 0.33) for PW). When restricting the examination of autocorrelation to datasets where all three methods could be compared (n = 171 datasets), the summary statistics were essentially unchanged.

The difference between REML and the other methods was more pronounced for shorter series (Table 9, Figure 9). All methods tended to yield negative values for short data series (fewer than approximately 12 data points). In longer data series (≥100 data points) all methods yielded similar estimates.

Table 9: Autocorrelation coefficient estimates (REML estimates of -1 and 1 are excluded, PW estimates < -1 are excluded)

	Autocorrelation coefficient (ρ) estimate					
	All available datasets		**Series with ≥ 24 points**		**Series with ≥ 100 points**	
Statistical method	**N**	**median (IQR)**	**N**	**median (IQR)**	**N**	**median (IQR)**
ARIMA	189	0.04 (-0.15,0.30)	154	0.07 (-0.10,0.36)	31	0.19 (0.04,0.54)
PW	186	0.05 (-0.14,0.33)	155	0.07 (-0.10,0.38)	31	0.19 (0.04,0.54)
REML	175	0.20 (-0.01,0.54)	147	0.20 (-0.01,0.53)	31	0.23 (0.08,0.57)
	Restricted to datasets where all methods can be compared					
ARIMA	171	0.05 (-0.14,0.30)	147	0.06 (-0.11,0.35)	31	0.19 (0.04,0.54)
PW	171	0.05 (-0.14,0.31)	147	0.07 (-0.11,0.35)	31	0.19 (0.04,0.54)
REML	171	0.20 (-0.01,0.54)	147	0.20 (-0.01,0.53)	31	0.23 (0.08,0.57)



Dashes on left represent the autocorrelation estimate distributions
the large symbol shows the median value.

Figure 9: Autocorrelation coefficient estimates. Scatterplot showing the autocorrelation estimate on the vertical axis and length of data series on the (log scale) horizontal axis. LOESS lines are overlaid to show trends in autocorrelation coefficient with data series length. Dashed lines on the left show the distribution of the estimates with overlaid symbols showing the median value. Abbreviations: ARIMA, autoregressive integrated moving average; PW, Prais-Winsten; REML, restricted maximum likelihood.

Confidence intervals for the REML estimates of autocorrelation show that for most studies with fewer than 48 data points the confidence limits extend below and above zero (Figure 10). For longer series, as expected, the confidence intervals are narrow, with many excluding no and negative autocorrelation estimates.

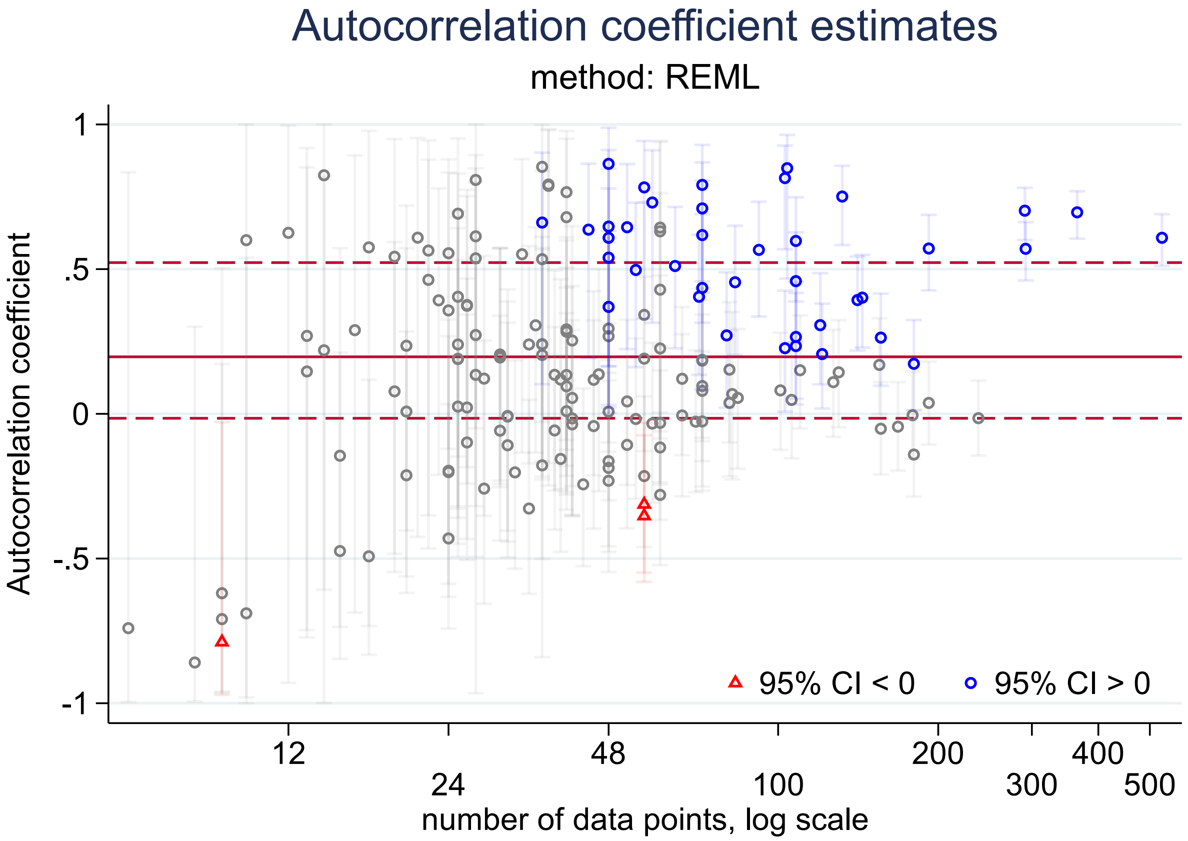


Figure 10: Autocorrelation coefficient estimates using the restricted maximum likelihood (REML) method. Data from 172 datasets. Red horizontal lines show the median and IQR of 0.2 (-0.02, 0.52). Blue circular markers indicated 95% confidence intervals that lie entirely above zero, red triangular markers indicate 95% confidence interval that lie entirely below zero.

4 Discussion

4.1 Summary and discussion of key findings

We re-analysed 190 ITS using six statistical methods and compared estimates of immediate level change, slope change, their associated standard errors, confidence intervals and p-values, and the estimated lag-1 autocorrelation. We found important inconsistency in these estimates across the methods, such that the interpretation of the findings in some series may differ depending on the chosen method.

On average, there were small systematic differences in estimates of level change across the statistical methods, with OLS yielding slightly larger estimates, and REML slightly smaller estimates compared with the other methods. For slope change, all methods yielded, on average, similar estimates. For some pairwise comparisons, the limits of agreement indicated large differences could arise. This was particularly notable in the comparisons between REML and the other methods. There were systematic differences in the standard errors between most methods, and the limits of agreement also indicated large differences could arise. ARIMA yielded systematically larger standard errors compared with all other methods, although the difference with REML was not as large. Of note, the PW yielded, on average, similar standard errors as OLS. This was perhaps surprising given PW provides adjustment for autocorrelation (which OLS does not), and in a numerical simulation study investigating the performance of these methods, PW was shown to perform better than OLS for data series approximately longer than 24 points (12). The results in our empirical investigation therefore likely reflect the influence of shorter data series.

The differences in point estimates and standard errors led to differences in the confidence interval widths, p-values, and statistical significance. Reflecting the pattern observed with standard errors, the ARIMA confidence intervals were wider compared with the other methods. However, REML with the Satterthwaite adjustment, which adjusts the t-distribution degrees of freedom used in the calculation of the confidence interval to account for uncertainty in estimation of the standard error, yielded the widest confidence intervals.

Our results show that naively basing conclusions on statistical significance could lead to a qualitatively different interpretation. There was important discordance in statistical significance (at the 5% level) across many of the pairwise method comparisons. As expected, the discordance was greatest between the methods that yielded larger standard errors or adjusted for uncertainty in estimation of the standard error (i.e. ARIMA, and REML with SW, respectively) and the other methods.

For long series (≥100 data points), all methods yielded similar estimates of autocorrelation. The methods yielded different estimates with short to medium length series (i.e. < 100 data points), with the ARIMA and OLS autocorrelation estimates being substantially smaller than REML. Given the true underlying autocorrelation would not be expected to vary by series length, the stability of the REML estimates over the different series lengths is suggestive of it being the preferable estimator, which has been shown in numerical simulation studies to be the case (12, 26).

The magnitude of autocorrelation estimates from these ITS public health datasets, with a median of 0.23 (IQR 0.08 to 0.57, restricted to series with ≥ 100 data points, n = 31 REML method), indicate that autocorrelation should not be ignored in the design or analysis of ITS studies. Despite this, in nearly 50% (113/230) of the series included in the review, autocorrelation was not considered, or the method to adjust for autocorrelation could not be determined (10). Furthermore, only 1.5% (3/200) studies provided evidence of a sample size calculation, and only two of these considered autocorrelation. Similar findings have also been observed in other systematic reviews. Jandoc et al. (8) found that only 146/220 (66.4%) ITS studies reported testing for autocorrelation, Hudson et al. (11) found that 63/115 (55%) considered autocorrelation, Ewusie et al. (9) found that only 812/1365 (59.5%) checked for autocorrelation and Hategeka et al. (27) similarly found that 66/120 (55%) checked or adjusted for autocorrelation.

4.2 Strengths and limitations

There are several strengths to our study. First, the repository of ITS studies was randomly sampled from PubMed, thus the findings are likely to be generalisable to ITS studies indexed in this database. Second, we used a variety of methods to obtain the time series data to optimise the number of datasets retrieved, which resulted in a large percentage of datasets being retrieved (190/230; 83%). Finally, we investigated a range of statistical methods, including those commonly used in practice (8-11), and compared their results using metrics of interest to researchers (point estimates, standard errors, confidence intervals, p-values, statistical significance) to provide a comprehensive picture of how the methods compared.

One limitation of this study is that our findings may not be generalisable to ITS studies outside of public health. For example, this would be the case if influencing characteristics (e.g. series length) of ITS studies in public health differ to other disciplines. Another limitation of our study is that we fitted a segmented linear regression model, assuming lag-1 autocorrelation, to all datasets. This model may have differed to that used in the original publication, and furthermore, may not have been the best fitting model. However, our re-analysis was not intended to specifically address the research question(s) of the original publications, but as a means of comparing different statistical methods.

4.3 Implications for practice

Our research has shown that in this set of ITS studies, the choice of statistical method can importantly affect the findings. This could lead to ‘bias in the selection of the reported result’ (28), where the reported result is chosen based on its magnitude, direction of effect, or statistical significance. Publication of protocols with detailed statistical analysis plans provide a mechanism for study authors to engender trust in the reported results (e.g. when there is consistency between the planned and used analysis methods). Protocols also allow readers to assess whether there were any changes to the analysis, and if so, what the legitimacy of those changes were. While protocols and statistical analysis plans are now common for randomised trials (29), in our review of ITS studies, none of the 200 studies reported having a published protocol. Protocols can be published in a peer-reviewed journal, published on a pre-print server (e.g. medRxiv), or registered in an online registry (e.g. open science framework).

Given the results can vary importantly, the selected statistical method needs to be carefully chosen considering the characteristics of the ITS. For example, Turner et al (12) found through a numerical simulation study that the length of the series is an important factor for deciding on the statistical method. Sensitivity analyses that use an alternative method might also be considered.

Finally, we recommend that time series data, including dates of the interruptions and any transition periods be made available alongside the publication. At a minimum, any plots of ITS data should follow graphing recommendations to facilitate data extraction using digitising software (30).

4.4 Implications for future research

Future research examining factors that may modify the magnitude of autocorrelation (e.g. type of outcome) would be useful. Knowledge of these factors would facilitate informed predictions about the likely magnitude of autocorrelation for an individual ITS study with particular characteristics, which could be used to more accurately determine the required sample size. Similar research has been undertaken investigating factors that modify intra-cluster correlations (ICCs) in cluster randomised trials, which has led to generalizable ‘rules-of-thumb’ on the selection of ICCs for sample size calculations in cluster trials (31).

5 Conclusion

ITS studies are commonly used in public health research to assess the impact of an intervention or exposure. A range of statistical methods are available to analyse ITS, and our study has shown that the choice of method can importantly affect the level and slope change estimates, their standard errors, width of confidence intervals and p-values. These differences may lead to qualitatively different conclusions being drawn about the impact of the interruption. Pre-specification of the statistical method is encouraged, and naive conclusions based on statistical significance should be avoided.

6 Declarations

6.1 Ethics approval and consent to participate

Not applicable

6.2 Consent for publication

Not applicable

6.3 Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on request.

6.4 Competing interests

The authors declare that they have no competing interests.

6.5 Funding

This work was supported by the Australian National Health and Medical Research Council (NHMRC) project grant (1145273). SLT is funded through an Australian Postgraduate Award administered through Monash University, Australia. JEM is supported by an NHMRC Career Development Fellowship (1143429). JMG holds a Canada Research Chair in Health Knowledge Uptake and Transfer and a Canadian Institute of Health Research (CIHR) Foundation grant (FDN 143269). The funders had no role in study design, decision to publish, or preparation of the manuscript.

6.6 Authors' contributions

JEM conceived the study and all authors contributed to its design. SLT collected the data by emailing authors and digitally extracting the data. SLT analysed the data and wrote the first draft of the manuscript, with contributions from JEM. SLT, JEM, AK, ABF, MT, JMG contributed to revisions of the manuscript and take public responsibility for its content.

6.7 Acknowledgements

We wish to thank all of the authors who generously contributed datasets for this study (Additional File 2).

7 Additional Material

Two files add additional material to this manuscript.

7.1 Additional File 1

The first additional file includes the Stata 15 (23) computer code used to analyse the data sets and can be found in the file “Additional_File_1.docx”.

7.2 Additional File 2

The second additional file contains a list of the studies that contributed data via publication, email or digital data extraction and can be found in the file “Additional_File_2.docx”.

8 References

1. Lopez Bernal J, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. International Journal of Epidemiology. 2016:dyw098.
2. Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. Journal of Clinical Pharmacy and Therapeutics. 2002;27(4):299-309.
3. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ : British Medical Journal. 2015;350:h2750.
4. Penfold RB, Zhang F. Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements. Academic Pediatrics. 2013;13(6):S38-S44.
5. Biglan A, Ary D, Wagenaar A. The Value of Interrupted Time-Series Experiments for Community Intervention Research. Prevention Science. 2000;1(1):31-49.
6. Zhang F, Wagner AK, Ross-Degnan D. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. Journal of Clinical Epidemiology. 2011;64(11):1252-61.
7. Huitema BE, McKean JW. Identifying Autocorrelation Generated by Various Error Processes in Interrupted Time-Series Regression Designs. Educational and Psychological Measurement. 2007;67(3):447-59.
8. Jandoc R, Burden AM, Mamdani M, Lévesque LE, Cadarette SM. Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. Journal of Clinical Epidemiology. 2015;68(8):950-6.
9. Ewusie J, Soobiah C, Blondal E, Beyene J, Thabane L, Hamid J. Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review. Journal of Multidisciplinary Healthcare. 2020;13:411-23.
10. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Cheng AC, et al. Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: a review. Journal of Clinical Epidemiology. 2020;122:1-11.

11. Hudson J, Fielding S, Ramsay CR. Methodology and reporting characteristics of studies using interrupted time series design in healthcare. BMC Medical Research Methodology. 2019;19(1):137.

12. Turner SL, Forbes AB, Karahalios A, Taljaard M, McKenzie JE. Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study. medRxiv. 2020:2020.10.12.20211706.

13. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Cheng AC, 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.

14. Rohatgi A. WebPlotDigitizer. 4.2 ed. San Francisco, California, USA2019.

15. Drevon D, Fursa SR, Malcolm AL. Intercoder Reliability and Validity of WebPlotDigitizer in Extracting Graphed Data. Behavior Modification. 2017;41(2):323-39.

16. Kutner M, Nachtscheim C, Neter J, Li W, Senter H. Applied linear statistical models. In: Kutner M, Nachtscheim C, Neter J, Li W, Senter H, editors. 2008. p. 880-.

17. Newey WK, West KD. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica. 1987;55:703.

18. StataCorp. Stata 15 Base Reference Manual. College Station, TX: Stata Press; 2017.

19. Prais SJ, Winsten, C.B. Trend estimators and serial correlation. In: University Y, editor. Cowles Commision1954.

20. Thompson WA. The Problem of Negative Estimates of Variance Components. The Annals of Mathematical Statistics. 1962;33(1):273-89.

21. Satterthwaite FE. An Approximate Distribution of Estimates of Variance Components. Biometrics Bulletin. 1946;2(6):110-4.

22. Nelson BK. Statistical methodology: V. Time series analysis using autoregressive integrated moving average (ARIMA) models. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine. 1998;5(7):739.

23. Stata. Stata Statistical Software. 15 ed. College Station, TX: Statcorp LLC; 2017.

24. Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research. 1999;8(2):135-60.

25. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20(1):37-46.

26. Cheang W-K, Reinsel GC. Bias Reduction of Autoregressive Estimates in Time Series Regression Model through Restricted Maximum Likelihood. Journal of the American Statistical Association. 2000;95(452):1173-84.

27. Hategeka C, Ruton H, Karamouzian M, Lynd LD, Law MR. Use of interrupted time series methods in the evaluation of health system quality improvement interventions: a methodological systematic review. BMJ Global Health. 2020;5(10):e003567.

28. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898.

29. van Rosmalen BV, Alldinger I, Cieslak KP, Wennink R, Clarke M, Ahmed Ali U, et al. Worldwide trends in volume and quality of published protocols of randomized controlled trials. HPB (Oxford, England). 2019;21:S666-S.

30. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Korevaar E, et al. Creating effective interrupted time series graphs: Review and recommendations. Research synthesis methods. 2020.

31. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. Clinical Trials. 2005;2(2):99-107.

9 Abbreviations

ARIMA	Autoregressive Integrated Moving Average
d.f.	Degrees of freedom
ICC	Intra-correlation coefficient
IQR	Inter-quartile range
ITS	Interrupted Time Series
LOESS	Local Regression
OLS	Ordinary Least Squares
NW	Newey-West
PW	Prais-Winsten
REML	Restricted Maximum Likelihood
REML-Satt	Restricted Maximum Likelihood with the small sample Satterthwaite approximation
RMSE	Root Mean Square Error
SE	Standard Error

Appendices

Appendix 1: Interrupted time series with a transition period

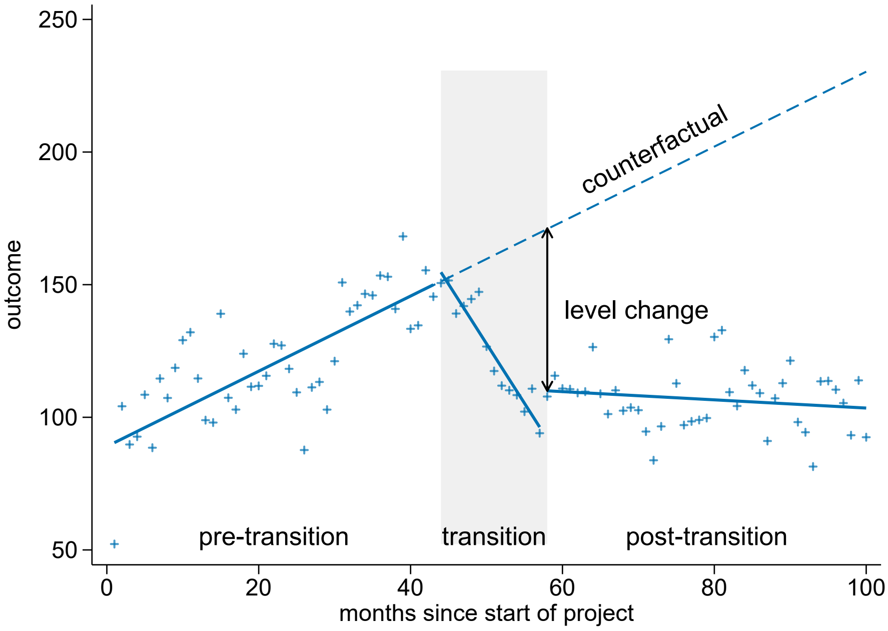


Figure 11: Interrupted time series with a transition period. The level change is calculated as the vertical distance between the counterfactual trend line and the post-transition trend line at the beginning of the post-transition series. The slope change is calculated as the change in slope between the counterfactual trend line and post-transition trend line.

Appendix 2: Difference in level and slope change by length of time series

We investigated the impact of series length on the difference in level and slope change estimates between each pair of methods. A matrix of scatterplots of the differences in level change (or slope change) versus the (log) length of series (overlaid with a local regression (LOESS) smoothed curve) for each pairwise method comparison is presented in Figure 12.

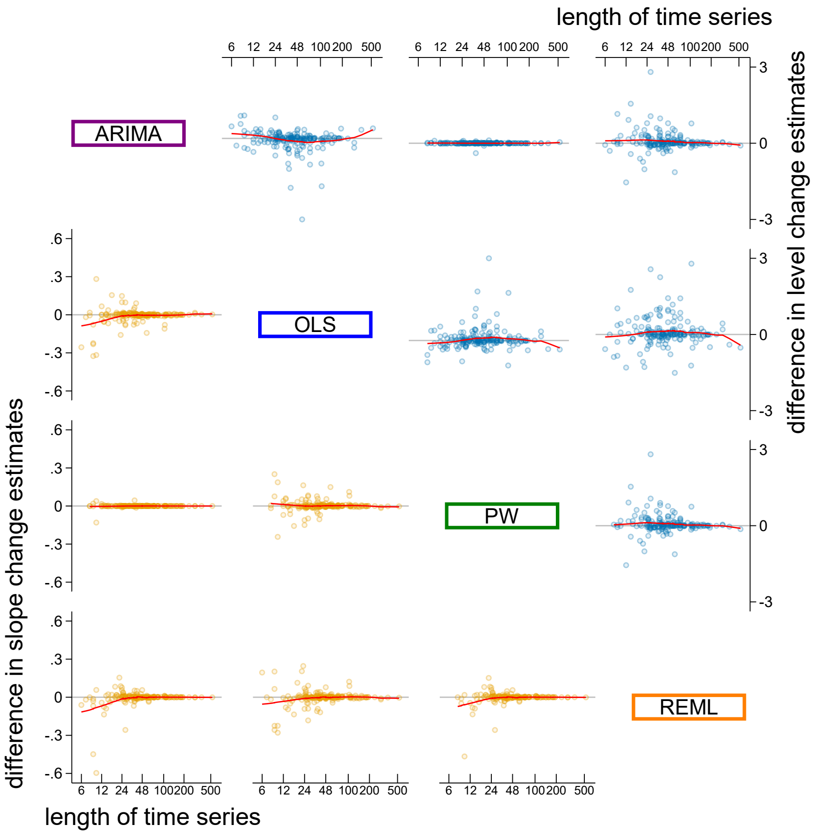


*Figure 12: Scatter plot of standardised level change versus series length. The vertical axis shows the effect estimate of the row method – the effect estimate of the column method. The horizontal axis shows the length of the time series (using a log scale). Blue dots represent situations in which the direction of effect was the same (both positive or both negative), while orange crosses represent situations in which the direction of effect was not the same. Red horizontal lines depict the average, red dashed lines depict the 95% limits of agreement (calculated as the average ±1.96*standard deviation of the differences). Grey lines indicate zero. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; PW, Prais-Winsten; REML, restricted maximum likelihood. Note that NW is not presented as it yields identical effect estimates to OLS, similarly REML with the Satterthwaite approximation is not presented because it yields identical effect estimates to REML.*

Appendix 3: Standardising the direction of effect

We standardised the direction of effect for each pairwise comparison of methods by multiplying both estimates by -1 if the first method's estimate was less than zero. Given the choice of first method was arbitrary, we repeated these analyses by standardising to the direction of the second method's estimate. Table 10 presents equivalent information to that presented in the top triangles of Tables 5 and 6, except with the standardisation reversed. The direction of standardisation did not affect mean differences in slope change estimates between methods, and had a small, but unimportant, impact for level change.

Table 10: Mean of differences in level change estimates between methods (column method - row method) (top triangle) and geometric mean ratio of standard errors for level change between methods (row method/column method) (shaded bottom triangle) with 95% limits of agreement. The OLS and NW level change estimates are the same, so the difference is not presented.

Level Change	**Mean of differences in level change estimates between methods (95% limits of agreement)**				
Mean of differences in slope change estimates between methods (95% limits of agreement)	**ARIMA**	0.05 (-0.81,0.92)	0.05 (-0.81,0.92)	0.00 (-0.07,0.07)	-0.11 (-0.86,0.64)
	-0.01 (-0.12,0.11)	**OLS**	-	-0.09 (-0.97,0.78)	-0.17 (-1.25,0.91)
	-0.01 (-0.12,0.11)	-	**NW**	-0.09 (-0.97,0.78)	-0.17 (-1.25,0.91)
	0.00 (-0.02,0.02)	0.00 (-0.09,0.09)	0.00 (-0.09,0.09)	**PW**	-0.11 (-0.84,0.63)
	-0.01 (-0.13,0.12)	0.00 (-0.12,0.12)	0.00 (-0.12,0.12)	0.00 (-0.10,0.09)	**REML**

Appendix 4: Detailed p-value comparisons

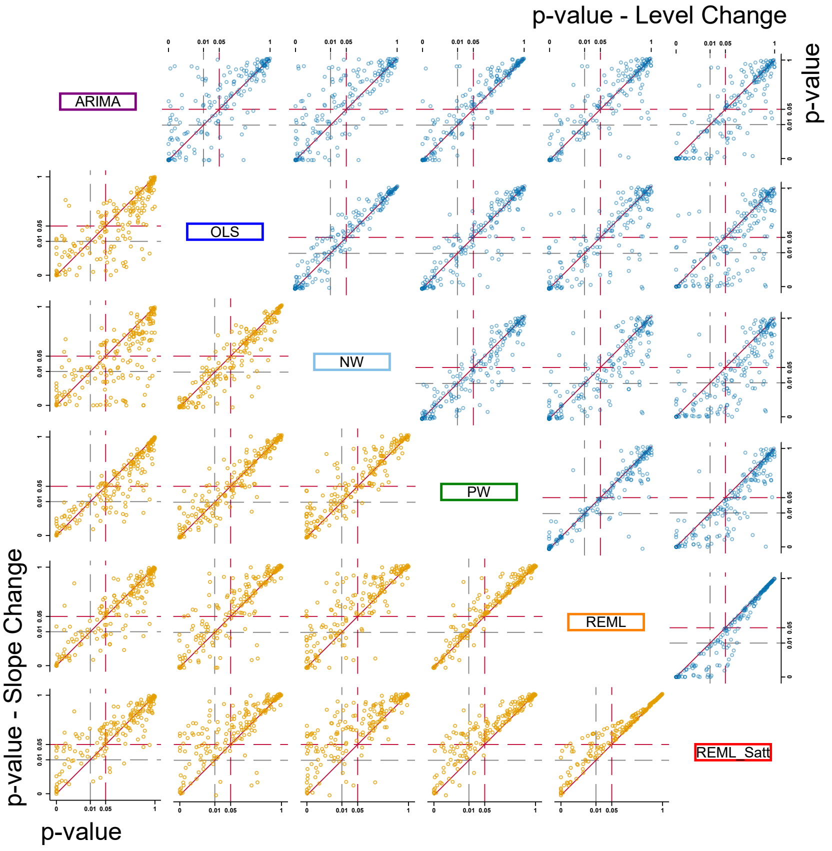


Figure 13: Pairwise comparisons of p-values between all statistical methods. The top triangle refers to level change p-values, the bottom triangle refers to slope change p-values. Dashed red lines indicate p-values of 0.05, dashed grey lines indicate p-values of 0.01. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW OLS with Newey-West standard error adjustments; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite adjustment.

Chapter 7. Summary and conclusions

Interrupted time series designs are important for examining the impact of public health interruptions. They are particularly useful in public health research when randomised trials are not always feasible (1-8). When appropriately designed and analysed, the results from ITS studies can contribute valuable information to inform health and policy decisions. As repositories of data continue to accumulate (e.g. registry data, administrative data), there will be greater opportunities for this design to be used. Given this, it is vital to understand how researchers are conducting, analysing and reporting ITS studies; and, to examine how the statistical methods used to analyse ITS perform, and which characteristics of ITS may compromise their performance. This knowledge can underpin development of tools and guidance for researchers and identify where further statistical methods' development is required. The aim of the research presented in this thesis was to assess the design, reporting quality and statistical methods used in ITS studies investigating interruptions with public health implications, and to provide tools and guidance to improve each of these elements.

The research presented in this thesis consists of a series of interlinked studies. First, a review of ITS studies was undertaken (Chapter 3). The review assessed design characteristics, statistical models, estimation methods, and quality of reporting in 200 ITS studies investigating interruptions with public health implications. The results of this review highlighted deficiencies in the reporting of ITS studies and informed the subsequent studies. Drawing from the visualisation literature, a set of recommendations for graphing the results of ITS were proposed (Chapter 4). ITS graphs from studies included in the review were assessed against these recommendations, and the recommendations were then applied to two graphs from the published literature. Computer code was developed and provided to enable researchers to implement the graphing recommendations. The statistical methods identified in the review were compared in a comprehensive simulation study using parameter values representative of real-world data. The findings of the simulation study led to recommendations about the choice of statistical methods (Chapter 5). A comparison of the level and slope change estimates (and associated statistics) when the statistical methods were applied to 190 real-world datasets was undertaken. Estimates of autocorrelation were calculated and summarised. An open access repository of the ITS datasets was collated for use in future statistical research (Chapter 6).

A visual depiction of the linkage between the thesis chapters and their outputs is displayed in Figure 4. The following sections summarise the key findings from each of the chapters, discuss overall findings, outline implications for ITS researchers, and conclude with proposals for further research.

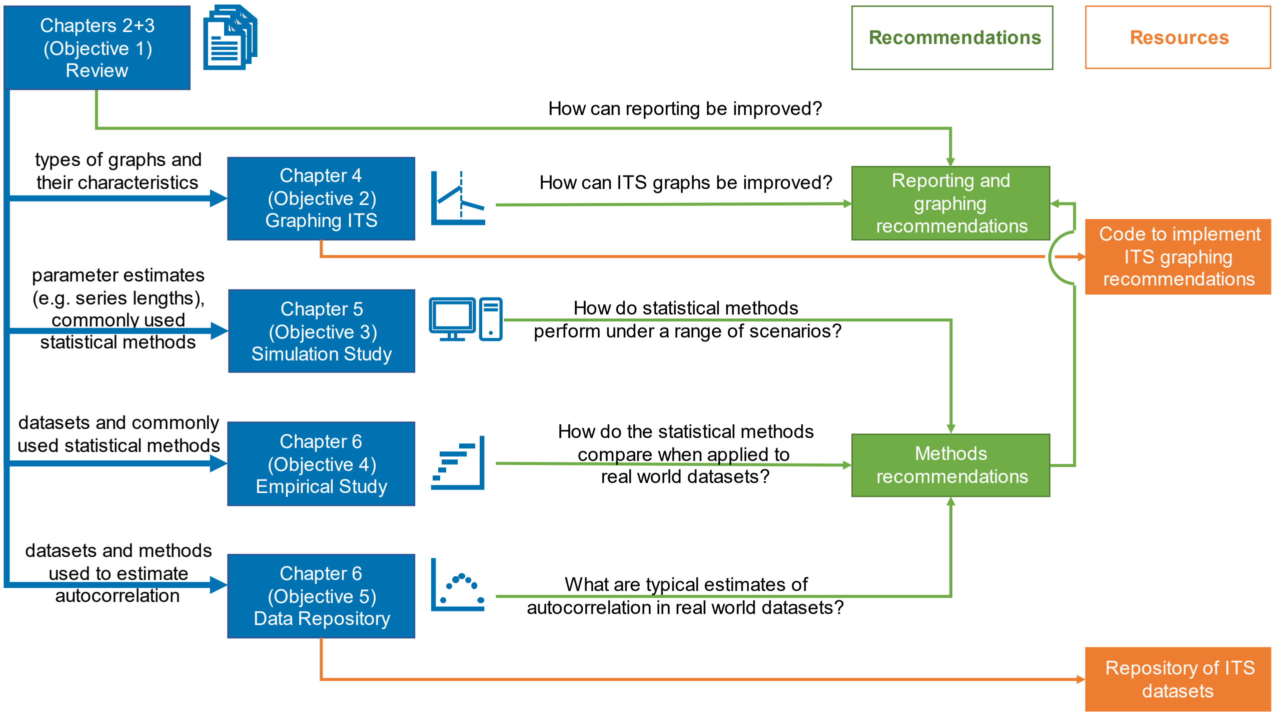

Chapters 2+3
(Objective 1)
Review
Recommendations
Resources
How can reporting be improved?
types of graphs and their characteristics
Chapter 4
(Objective 2)
Graphing ITS
How can ITS graphs be improved?
Reporting and graphing recommendations
Code to implement ITS graphing recommendations
parameter estimates (e.g. series lengths), commonly used statistical methods
Chapter 5
(Objective 3)
Simulation Study
How do statistical methods perform under a range of scenarios?
datasets and commonly used statistical methods
Chapter 6
(Objective 4)
Empirical Study
How do the statistical methods compare when applied to real world datasets?
Methods recommendations
datasets and methods used to estimate autocorrelation
Chapter 6
(Objective 5)
Data Repository
What are typical estimates of autocorrelation in real world datasets?
Repository of ITS datasets

Figure 4: A visual depiction of the thesis studies, how they interlink, and their outputs. Blue arrows show how the review informed subsequent studies, green boxes show the recommendations, orange boxes show the developed resources.

7.1 Summary of thesis chapters

7.1.1 Chapters 2 and 3 – Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: protocol and review

Chapter 2 presented the protocol for a review of published ITS studies (2013-2017) and Chapter 3 presented the results of the review. The review aimed to examine the design characteristics, statistical methods, and completeness of reporting of a random sample of 200 ITS studies (including 230 series) examining interruptions with public health impacts.

Key findings from the review included the following:

- Nearly all the studies evaluated the impact of an intervention (94%, 188/200), and more than half investigated policy change interventions (52%, 103/200).
- Time series lengths were frequently short, with a median of 48 points (IQR 30,100), and two thirds of the time series were composed of two segments (66%, 152/230).
- Fewer than a quarter of analyses included a control series (24%, 48/230).
- A graph depicting the time series data was almost always included (93%, 214/230).
- The most common statistical analysis methods were linear regression (31%, 72/230) and ARIMA (19%, 43/230); however, in almost one fifth of the series, the statistical method could not be determined (17%, 40/230).

- When a non-ARIMA method was used (n=187), in almost two thirds of analyses it could not be determined if a method for handling autocorrelation was considered or employed (60%, 113/187). When linear regression was used (n=72), autocorrelation was not adjusted for in over a third of the series (35%, 25/72), and in almost one fifth of these analyses it could not be determined whether there was adjustment for autocorrelation (19%, 14/72).
- Estimates of the magnitude of autocorrelation were almost never reported (1%, 3/230).
- The two most common effect measures reported were the level change at the time of the interruption (70%, 160/230) and slope change following the interruption (54%, 125/230).
- Using the counterfactual trend to estimate differences at times other than the immediate time of the interruption was reported in only a quarter of analyses (25%, 58/230).
- For over one third of the reported effect estimates, a measure of precision was not reported (CI or SE) (37%, 311/852).

7.1.2 Chapter 4 - Creating effective interrupted time series graphs: review and recommendations

Nearly all the publications included in the review (Chapter 3) included a graphical display of ITS data. Inspection of these graphs highlighted that a wide variety of graphical displays were used, and that, surprisingly, many did not actually clearly depict the time series data or the results. This motivated a formal examination of the graphs used in publications of ITS studies. The aim of the research presented in Chapter 4 was to formulate recommendations for graphing ITS data (informed from seminal data visualisation resources) and to assess whether the graphs from ITS studies included in the review (Chapter 3) met these recommendations. To aid application, the recommendations were demonstrated using two examples, and computer code was provided to produce ITS graphs meeting the recommendations.

Key results of this research included:

- Recommendations for the graphing of ITS data were proposed. The recommendations pertained to the following graph characteristics: data points, timing of interruption, pre- and post-interruption trend lines, counterfactual, additional lines and general graph components.
- The recommendation to plot the data points was met by fewer than two thirds of graphs in the review (60%, 130/217).
- The recommendation to indicate the timing of the interruption with a vertical line or shading was met in approximately three quarters of the graphs (73%, 158/217).
- The recommendation to plot the fitted pre- and post-interruption trend lines was only met in 48% (103/217) of graphs.
- The recommendation to plot the counterfactual was rarely met (17%, 37/217).

7.1.3 Chapter 5 - Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study

Chapter 5 presented a simulation study investigating the performance of a set of statistical methods used in analysing ITS data. The parameter values used in constructing the simulated datasets, and the statistical methods evaluated, were informed by the findings of the review (Chapter 3). Eight hundred different scenarios were simulated using a simple model with a continuous outcome and a single interruption at the mid-point of the series. The datasets were constructed with varying level and slope changes, lengths of series, magnitudes of lag-1 autocorrelation and with constant variance. The statistical methods examined included OLS, NW, PW, ARIMA, REML and REML with the small sample Satterthwaite approximation.

Key findings from this research included the following:

- All statistical methods yielded unbiased estimates of level change and slope change.
- The OLS and NW methods underestimated standard errors in the presence of autocorrelation.
- All methods systematically underestimated the magnitude of autocorrelation; however, the REML method was more accurate than the other methods at estimating the magnitude of autocorrelation.
- For all parameter combinations, CI coverage was always lower than the nominal 95% level except:
    - the OLS method in the absence of autocorrelation, and
    - the Satterthwaite method (which was often overly-conservative).
- The method with the highest 95% CI coverage was OLS for short time series (approximately 12 points or under) and REML (with or without the small sample Satterthwaite approximation) for longer time series.
- The DW test for autocorrelation was often inconclusive or incorrect except for longer data series (approximately 80 points or over) when there was either no underlying autocorrelation or larger values of underlying autocorrelation (greater than approximately 0.5).

7.1.4 Chapter 6 - Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series

Chapter 6 presented the findings of an empirical evaluation comparing the results when different statistical methods were applied to 190 real-world datasets. A segmented linear regression model was fitted to each series, treating the outcome as continuous, including a single interruption, and allowing for lag-1 autocorrelation. This study examined the same statistical methods as those included in the simulation study (Chapter 5) and provided complementary evidence about how the statistical methods operate. The ITS identified in the review (Chapter 3) formed the repository of ITS data series included in the empirical evaluation.

Key findings from this research included the following:

- 190 unique datasets were obtained from the 230 datasets identified in the review. Digital data extraction provided 72% (137/190) of these datasets, 24% (45/190) were obtained via email contact with authors and 4% (8/190) were extracted directly from the publications.
- There were very small systematic differences between pairwise comparisons of the six methods in terms of level change and no systematic differences between pairwise comparisons in terms of slope change. The 95% limits of agreement indicated that for some datasets, large differences in effect estimate could arise based on the statistical method used.
- There were systematic differences in the pairwise comparisons of standard errors; ARIMA had notably larger SEs than the other methods. The 95% limits of agreement showed that there could be large differences in SE estimates between methods.
- The ARIMA method generally yielded more conservative CIs than the other methods. The OLS and NW methods generally yielded less conservative CIs than the other methods.
- The percentage agreement in statistical significance (dichotomised at the 5% significance level) for the level change estimates between methods ranged from 79.3% (NW versus REML-Satt) to 97.1% (PW versus REML). For slope change the percentage agreement ranged from 75.3% (NW versus REML-Satt) to 93.6% (PW versus REML).
- Differences in the effects, and their associated statistics, estimated by the different statistical methods could lead to differences in the conclusions drawn about the impact of an interruption.
- Autocorrelation was often found in the included studies with a median of 0.23 (IQR 0.08 to 0.57, restricted to series with ≥ 100 data points, n=31, REML method). Length of series was found to impact the size of the estimated autocorrelation for PW and ARIMA, but not for REML.

7.2 Overall discussion

The aim of this thesis was to advance the existing body of knowledge on the design, analysis and reporting of ITS studies used in public health research. The research considered continuous outcomes, segmented linear regression, and lag-1 autocorrelation, since these models are commonly fitted in practice, but there has been a paucity of research examining their performance (53). Future research that proposes to address more complex models and outcome types is outlined in section 7.4.

The key findings of this thesis were that, for ITS studies that have evaluated the impact of interventions or exposures on public health outcomes, details of the statistical methods and considerations of autocorrelation were rarely reported; graphs often did not meet core recommendations; and, the statistical methods commonly used generally did not operate as desired, and in practice, can yield importantly different effect estimates and associated statistics. In this section, issues related to these findings are discussed.

Autocorrelation is a key consideration in the design and analysis of ITS studies. The reporting of the statistical methods used in the studies included in the review was often incomplete, making it difficult to determine whether autocorrelation was acknowledged or accounted for (Chapter 3). Poor reporting of the handling of autocorrelation in ITS studies has also been found in reviews investigating other types of interruptions (9, 17-19, 24). Furthermore, the importance of reporting correlation coefficients has been recognised for other designs (e.g. cluster randomised trials (60)) but this has not gained recognition for ITS designs. The findings from this thesis therefore indicate that more work is needed to communicate the importance of autocorrelation in ITS designs.

Many methods are available to estimate autocorrelation and to detect the presence of autocorrelation (26, 61-63). The DW test was often used to identify autocorrelation in the reviewed studies (Chapter 3); however, this test often fails to identify autocorrelation in time series of lengths that are commonly used in practice (Chapter 5). The REML method was found to accurately estimate the magnitude of autocorrelation for series longer than approximately 12 points (Chapter 5). Applying the REML method to the public health ITS datasets in the review showed that autocorrelation was frequently present and should not be ignored in the analysis (Chapter 6). Until now, there has been no guidance in the literature as to the typical values of autocorrelation; the work presented in Chapter 6 assists in filling this gap.

Another key consideration in the design and analysis of ITS studies is the length of the time series (Chapter 5). Advice on the appropriate length of time series for an ITS study has been proposed in several tutorial papers (2, 3, 10, 25, 47); however, the suggested minimum number of data points varies. The review of ITS studies in this thesis found that a large proportion of studies used short time series (median 48, IQR 30 to 100) (Chapter 3). Time series shorter than approximately 24 points are unlikely to allow adequate modelling of the counterfactual post-interruption trend (leading to wide distributions of possible observed effect estimates), or accurate estimation of the magnitude of autocorrelation. Furthermore, CI coverage is likely to be less than the 95% nominal level (Chapter 5).

The performance of the statistical methods is heavily influenced by the key factors mentioned above, magnitude of autocorrelation and series length. Until now, no studies have compared the performance of different statistical methods in the analysis of ITS. The findings of the simulation study (Chapter 5) provided some guidance for selection of appropriate statistical method when dealing with continuous data with a single interruption at the mid-point of the series, but also highlighted the need for development of statistical methods for short ITS (fewer than 12 data points). OLS had better CI coverage than other methods for very small series lengths, while for longer series the REML method had the best CI coverage (Chapter 5). When different statistical methods were applied to real world datasets, important differences were sometimes observed in the effect estimates and associated statistics. These differences mean that the choice of statistical method may qualitatively alter the interpretation of the effect of the interruption (Chapter 6).

The choice of effect measures and the completeness of reporting of the resulting effect estimates are core for understanding the impact of an interruption. The most commonly reported effect measures are the level change at the time of the interruption and slope change post-interruption (17, 53). However, a strength of the ITS design is that long term estimates of the impact can also be estimated, though the review demonstrated that this was rarely done (Chapter 3). When effect estimates were reported, they often did not include an estimate of precision (SE or CI), which is necessary for accurate interpretation of the impact of an interruption, and to enable inclusion of effects in a meta-analysis (28). However, the research in this thesis demonstrates that even when CIs are provided, care should be taken in their interpretation as, in the presence of autocorrelation, CI coverage was systematically lower than the 95% nominal value (Chapter 5), making reported 95% CIs incorrectly too precise. Similarly, borderline findings of statistical significance for effect estimates should be cautiously interpreted.

An accurate graph depicting the source time series data and results is an important aspect of the reporting for ITS studies. Effect measures, such as the level change at the time of the interruption and slope change post-interruption, are easily depicted on an appropriate graph and allow a reader to visualise the impact of the interruption (Chapter 4). The impact of the interruption can also be seen throughout the post-interruption period by visual inspection of the difference between the post-interruption and counterfactual trend lines. In addition to aiding interpretation of the study, a well-designed graph that includes data points allows for accurate data extraction. This is important for replication of the study results and for systematic reviewers who may wish to re-analyse the data (24, 28). A re-analysis may be desirable as the effect measure of interest may not have been reported, a measure of precision may not have been reported, or there may have been an incorrect analysis (Chapters 5 and 6). Complete reporting is necessary for other researchers to understand what was done, to allow assessment of the potential for bias in the study design, and facilitates replication of the study (24, 64-67). The publication and use of reporting guidelines, such as those proposed by Jandoc et al. (9) and under development by Lopez-Bernal (68) may also improve the reporting of ITS studies (69-71).

7.3 Recommendations for ITS researchers

In this section, recommendations for the design and analysis of ITS studies are suggested. The design and analysis recommendations stem from the findings of the simulation study (Chapter 5) and empirical evaluation (Chapter 6). As such, these recommendations are directly applicable to the model and scenarios investigated (i.e. continuous outcomes, single interruption, constant variance and lag-1 autocorrelation). Caution is required in generalising to different scenarios (e.g. different outcome types, different model structures).

7.3.1 Design recommendations

- A minimum of 24 data points is recommended. Using fewer points than this results in poor CI coverage irrespective of the statistical method employed.
- Assume the presence of autocorrelation in sample size calculations.
- Publishing a study protocol detailing the proposed analysis strategy is recommended, either in a peer-reviewed journal, pre-print server or online registry.

7.3.2 Analysis recommendations

- The DW test to detect the presence of autocorrelation is not recommended.
- For small series lengths (fewer than 12 points) OLS is the recommended statistical method, for longer time series, REML is recommended. If REML does not converge, ARIMA is recommended.

7.3.3 Reporting recommendations

- Report a measure of precision (SE or CI) with effect estimates and avoid dichotomous conclusions based on statistical significance.
- Reporting the magnitude of autocorrelation is recommended (estimated using REML).
- Consider reporting the longer-term effects of the interruption as well as the immediate impacts.
- The inclusion of a graph is recommended, plotting the data points, the interruption time, and lines for the fitted pre- and post-interruption trend, as well as the counterfactual post-interruption trend.

7.4 Future research

The individual chapters include recommendations for future research. In the following these are summarised and expanded upon.

- The statistical methods examined in this thesis were applied to continuous outcomes. Examination of how statistical methods perform for other outcomes types would be valuable.
- The most frequently used statistical methods identified in the review (Chapter 3) were evaluated in this thesis, with the addition of the REML method. However, there is scope to examine other statistical methods (72, 73) or Bayesian approaches, where the uncertainty in the estimate of autocorrelation could be incorporated directly into the estimation procedure.
- The simulation and empirical evaluation research in this thesis only considered lag-1 autocorrelation. Further research investigating the effects of autocorrelation is required, including examination of lags greater than 1 (e.g. seasonal effects).
- Further work examining factors that may modify the magnitude of autocorrelation (e.g. type of intervention or outcome) would be helpful for developing general rules about the likely magnitude of autocorrelation, which would aid in sample size calculation. This could include examining the impact of adjusting for seasonality on estimates of lag-1 autocorrelation.
- Future research that compares bias and efficiency when simple (e.g. models where there is only adjustment for lag-1 autocorrelation) versus more complex models (e.g. models including adjustment for time-varying predictors) would be valuable. This research may be particularly valuable because for many ITS studies, it may be the case that time-varying predictors are not available, and the only option is to model the autocorrelation.

- In the simulation study (Chapter 5), datasets with an equal number of points in the pre- and post-interruption segments, and constant error variance throughout the series, were generated. Statistical power is likely to be affected by the ratio of pre- to post-interruption series length (50), and further work is required to determine the extent of the impact of unequal pre- and post-interruption series lengths and non-constant error variance.
- The research presented in this thesis focused on single-series ITS designs. Research investigating more complex designs, such as multi-location or controlled ITS would be valuable. For example, this could include identifying the range of statistical methods available, and then evaluating their performance using numerical simulation, as has been done in this thesis.
- An important decision in an ITS design is the choice of time period over which the data are aggregated. Aggregating over longer time periods may reduce variation in the data, but at the cost of reducing the number of data points. Future work should investigate the preferred balance between variance reduction and series length.
- Finally, there are many effect measures that can be used to quantify the impact of an interruption. The research in this thesis has focused on the most commonly used measures (level change and slope change); however, the examination of other measures would be valuable.

References

1. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ : British Medical Journal. 2015;350:h2750.
2. Lagarde M. How to do (or not to do) ... Assessing the impact of a policy change with routine longitudinal data. Health Policy and Planning. 2011;27(1):76-83.
3. Lopez Bernal J, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. International Journal of Epidemiology. 2016:dyw098.
4. Sanson-Fisher RW, Bonevski B, Green LW, D'este C. Limitations of the Randomized Controlled Trial in Evaluating Population-Based Health Interventions. American journal of preventive medicine. 2007;33(2):155-61.
5. Biglan A, Ary D, Wagenaar A. The Value of Interrupted Time-Series Experiments for Community Intervention Research. Prevention Science. 2000;1(1):31-49.
6. Craig P, Cooper C, Gunnell D, Haw S, Lawson K, Macintyre S, et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. Journal of Epidemiology and Community Health (1979-). 2012;66(12):1182-6.
7. Victora CG, Habicht J-P, Bryce J. Evidence-Based Public Health: Moving Beyond Randomized Trials. American Journal of Public Health. 2004;94(3):400-5.
8. Penfold RB, Zhang F. Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements. Academic Pediatrics. 2013;13(6):S38-S44.
9. Jandoc R, Burden AM, Mamdani M, Lévesque LE, Cadarette SM. Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. Journal of Clinical Epidemiology. 2015;68(8):950-6.
10. Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. Journal of Clinical Pharmacy and Therapeutics. 2002;27(4):299-309.
11. Katikireddi SV, Der G, Roberts C, Haw S. Has Childhood Smoking Reduced Following Smoke-Free Public Places Legislation? A Segmented Regression Analysis of Cross-Sectional UK School-Based Surveys. Nicotine & Tobacco Research. 2016;18(7):1670-4.
12. Nazif-Munoz JI, Quesnel-Vallée A, van den Berg A. Did Chile's traffic law reform push police enforcement? Understanding Chile's traffic fatalities and injuries reduction. Injury Prevention. 2015;21(3):159-65.
13. McPhedran SCM, G. Lethal firearm-related violence against Canadian women: did tightening gun laws have an impact on women's health and safety? Violence and victims. 2013;28(5):875-83.
14. Abegaz T, Berhane Y, Worku A, Assrat A. Effectiveness of an improved road safety policy in Ethiopia: an interrupted time series study. BMC public health. 2014;14(1):539.
15. Scherb HH, Mori K, Hayashi K. Increases in perinatal mortality in prefectures contaminated by the Fukushima nuclear power plant accident in Japan. Medicine. 2016;95(38):e4958.
16. Vokó Z, Pitter JG. The effect of social distance measures on COVID-19 epidemics in Europe: an interrupted time series analysis. GeroScience. 2020;42(4):1075.
17. Hudson J, Fielding S, Ramsay CR. Methodology and reporting characteristics of studies using interrupted time series design in healthcare. BMC Medical Research Methodology. 2019;19(1):137.
18. Hategeka C, Ruton H, Karamouzian M, Lynd LD, Law MR. Use of interrupted time series methods in the evaluation of health system quality improvement interventions: a methodological systematic review. BMJ Global Health. 2020;5(10):e003567.
19. Ewusie J, Soobiah C, Blondal E, Beyene J, Thabane L, Hamid J. Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review. Journal of Multidisciplinary Healthcare. 2020;13:411-23.
20. Lopez Bernal J, Cummins S, Gasparrini A. The use of controls in interrupted time series studies of public health interventions. International Journal of Epidemiology. 2018;47(6):2082-93.

21. Linden A. A matching framework to improve causal inference in interrupted time-series analysis. Journal of Evaluation in Clinical Practice. 2018;24(2):408-15.

22. Huitema BE. Analysis of covariance and alternatives statistical methods for experiments, quasi-experiments, and single-case studies. 2nd ed. ed. Hoboken, N.J.: Hoboken, N.J. : Wiley; 2011.

23. Huitema BE, Mckean JW. Design Specification Issues in Time-Series Intervention Models. Educational and Psychological Measurement. 2000;60(1):38-58.

24. Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. International Journal of Technology Assessment in Health Care. 2003;19(4):613-23.

25. Gebski V, Ellingson K, Edwards J, Jernigan J, Kleinbaum D. Modelling interrupted time series to evaluate prevention and control of infection in healthcare. Epidemiology and Infection. 2012;140(12):2131-41.

26. Huitema BE, McKean JW. Identifying Autocorrelation Generated by Various Error Processes in Interrupted Time-Series Regression Designs. Educational and Psychological Measurement. 2007;67(3):447-59.

27. Liu W, Ye S, Barton BA, Fischer MA, Lawrence C, Rahn EJ, et al. Simulation-based power and sample size calculation for designing interrupted time series analyses of count outcomes in evaluation of health policy interventions. Contemporary clinical trials communications. 2020;17:100474.

28. Korevaar E, Karahalios A, Forbes AB, Turner SL, McDonald S, Taljaard M, et al. Methods used to meta-analyse results from interrupted time series studies: A methodological systematic review protocol. F1000Res. 2020;9:110.

29. Boers M. Designing effective graphs to get your message across. Annals of the Rheumatic Diseases. 2018;77(6):833.

30. Cleveland WS. The elements of graphing data. Rev. ed. ed. Murray Hill, N.J.: Murray Hill, N.J. : AT&T Bell Laboratories; 1994.

31. Few S. Show me the numbers : designing tables and graphs to enlighten. 2nd ed. ed. Burlingame, Calif.: Burlingame, Calif. : Analytics Press; 2012.

32. Lane DM, Sándor A. Designing Better Graphs by Including Distributional Information and Integrating Words, Numbers, and Images. Psychological Methods. 2009;14(3):239-57.

33. Tufte ER. Beautiful evidence. Cheshire, Conn.: Cheshire, Conn. : Graphics Press; 2006.

34. Tufte ER. The visual display of quantitative information. 2nd ed. ed. Cheshire, Conn.: Cheshire, Conn. : Graphics Press; 2001.

35. Tufte ER. Visual explanations : images and quantities, evidence and narrative. Cheshire, Conn.: Cheshire, Conn. : Graphics Press; 1997.

36. Yau N. Visualize this : the FlowingData guide to design, visualization, and statistics. Ebooks C, editor. Indianapolis, Ind.: Indianapolis, Ind. : Wiley Pub.; 2011.

37. Cheang W-K, Reinsel GC. Bias Reduction of Autoregressive Estimates in Time Series Regression Model through Restricted Maximum Likelihood. Journal of the American Statistical Association. 2000;95(452):1173-84.

38. Kutner M, Nachtscheim C, Neter J, Li W, Senter H. Applied linear statistical models. In: Kutner M, Nachtscheim C, Neter J, Li W, Senter H, editors. 2008. p. 880-.

39. Newey WK, West KD. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica. 1987;55:703.

40. Smith J, McAleer M. Newey-West covariance matrix estimates for models with generated regressors. Applied Economics. 1994;26(6):635-40.

41. Alpargu G, Dutilleul P. Efficiency and Validity Analyses of Two-Stage Estimation Procedures and Derived Testing Procedures in Quantitative Linear Models with AR(1) Errors. Communications in Statistics - Simulation and Computation. 2003;32(3):799-833.

42. Nelson BK. Statistical methodology: V. Time series analysis using autoregressive integrated moving average (ARIMA) models. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine. 1998;5(7):739.

43. Box GEPa. Time series analysis : forecasting and control. 5th ed. ed: Hoboken, New Jersey : Wiley; 2016.

44. Harrop JW, Velicer WF. A Comparison of Alternative Approaches to the Analysis of Interrupted Time-Series. Multivariate Behavioral Research. 1985;20(1):27-44.
45. Singer J, Willett J. Applied longitudinal data analysis: modeling change and event occurrence2003. xx-xx p.
46. Thompson WA. The Problem of Negative Estimates of Variance Components. The Annals of Mathematical Statistics. 1962;33(1):273-89.
47. EPOC. EPOC resources for review authors2017 20/02/2020. Available from: epoc.cochrane.org/resources/epoc-resources-review-authors.
48. SAS/ETS 14.1 User's Guide. 14.1 ed. Cary, NC: SAS Institute Inc.; 2015.
49. McLeod AI, Vingilis ER. Power Computations for Intervention Analysis. Technometrics. 2005;47(2):174-81.
50. Zhang F, Wagner AK, Ross-Degnan D. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. Journal of Clinical Epidemiology. 2011;64(11):1252-61.
51. SAS Institute Inc. SAS OnlineDoc(R). Cary, NC: SAS Institute Inc; 2005.
52. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Cheng AC, 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.
53. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Cheng AC, et al. Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: a review. Journal of Clinical Epidemiology. 2020;122:1-11.
54. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Korevaar E, et al. Creating effective interrupted time series graphs: Review and recommendations. Research synthesis methods. 2020.
55. Turner SL, Forbes AB, Karahalios A, Taljaard M, McKenzie JE. Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study. medRxiv. 2020:2020.10.12.20211706.
56. Stata. Stata Statistical Software. 15 ed. College Station, TX: Statcorp LLC; 2017.
57. Turner SL, Forbes AB, Karahalios A, Taljaard M, McKenzie JE. Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study - Code and Data2020.
58. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38(11):2074-102.
59. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw J, McKenzie JE. Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series - published and extracted data2020.
60. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345(sep04 1):e5661-e.
61. Cumby R, Huizinga J. Testing the autocorrelation structure of disturbances in ordinary least squares and instrumental variables regressions. Econometrica (1986-1998). 1992;60(1):185.
62. Awosoga OA, McKean JW, Huitema BE. Simple Robust Tests for Autocorrelated Errors in Time Series Design Intervention Models. Communications in Statistics - Theory and Methods. 2014;43(13):2629-41.
63. Bence JR. Analysis of Short Time Series: Correcting for Autocorrelation. Ecology. 1995;76(2):628-39.
64. Aytug ZG, Rothstein HR, Zhou W, Kern MC. Revealed or Concealed? Transparency of Procedures, Decisions, and Judgment Calls in Meta-Analyses. Organizational Research Methods. 2012;15(1):103-33.
65. Katherine SB, John PAI, Claire M, Brian AN, Jonathan F, Emma SJR, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience. 2013;14(5):365.
66. Moonesinghe R, Khoury MJ, Janssens ACJW. Most Published Research Findings Are False—But a Little Replication Goes a Long Way (Essay). PLoS Medicine. 2007;4(2):e28.
67. Simons DJ. The Value of Direct Replication. Perspectives on psychological science. 2014;9(1):76-80.

68. Lopez Bernal J. Framework for Enhanced Reporting of Interrupted Time Series (FERITS) 2018 [updated updated 22 May 2018. Available from: http://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-observational-studies/#92

69. Agha RA, Fowler AJ, Limb C, Whitehurst K, Coe R, Sagoo H, et al. Impact of the mandatory implementation of reporting guidelines on reporting quality in a surgical journal: A before and after study. International journal of surgery (London, England). 2016;30:169-72.

70. Panic N, Leoncini E, de Belvis G, Ricciardi W, Boccia S. Evaluation of the Endorsement of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) Statement on the Quality of Published Systematic Review and Meta-Analyses.(Research Article). PLoS ONE. 2013;8(12):e83138.

71. Leclercq V, Beaudart C, Ajamieh S, Rabenda V, Tirelli E, Bruyère O. Meta-analyses indexed in PsycINFO had a better completeness of reporting when they mention PRISMA. Journal of clinical epidemiology. 2019;115:46-54.

72. Cruz M, Bender M, Ombao H. A robust interrupted time series model for analyzing complex health care intervention data. Statistics in Medicine. 2017;36(29):4660-76.

73. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics. 1997;53(3):983.

74. Cummins S, Lopez Bernal J, Gasparrini A. The use of controls in interrupted time series studies of public health interventions. International journal of epidemiology. 2018;47(6):2082-93.

Appendix A. Additional file accompanying Chapter 2 – Review data extraction items

"Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: protocol for a review"

Simon L. Turner, Amalia Karahalios, Andrew B. Forbes, Monica Taljaard, Jeremy M. Grimshaw, Allen Cheng, Lisa Bero, Joanne E. McKenzie

The following table contains the data extraction items used for the review.

Additional File 1: Data Extraction items.

Item	**Study Level**
	<standard items such as author, title, journal etc.>
1	Why was an ITS study design/analysis chosen?
2	Type of intervention
3	Date/time of intervention(s)
4	Longer description of intervention
5	Number of study observations
6	Study observation notes
7	Main software used
8	Additional software used
9	Software details
10	Were there sample size calculations?
11	Sample size calculation details

	Outcome Level
12	Outcome selection hierarchy
13	Further details of outcome selection
14	Any details about subpopulation
15	At the individual level, what was the outcome?
16	What type of outcome is this?
17	Any further individual level details?
18	Text description of group level outcome
19	What type of outcome is this?
20	Any further group level details?
21	Was this outcome standardised?
22	Any standardisation details?
23	Type of time interval
24	Any details on time interval?

25	Was there any evidence of model pre-specification
26	Any details of the pre-specification
27	Was the intervention time pre-specified?
28	Any further intervention time details
29	Was a delay from intervention to its impact mentioned?
30	Was there a search for the time of impact?
31	What was/were the time(s) of impact?
32	Details of how delayed impact was dealt with
33	What general shape is the model?
34	Further notes on the shape of the model
35	Statistical model used
36	Further details of the statistical model
37	Is there any evidence of checking for model fit?
38	Details of checking for model fit
39	Details on parameterisation
40	Were there any additional analyses?
41	Details of further analyses
42	What presentation was used for results
43	Was a graph included
44	Is data from other groups available?
45	If other data was supplied please attach here
46	Please take a screenshot and attach the image of the graph here
47	Was autocorrelation mentioned
48	How was autocorrelation detected
49	How was autocorrelation handled
50	What was the estimate of autocorrelation parameter rho
51	Any further details on autocorrelation
52	Was seasonality mentioned
53	How was seasonality detected
54	How was seasonality modelled
55	Any further details on seasonality
56	Was non-stationarity mentioned
57	Which test was used to detect non-stationarity
58	Any further details on non-stationarity
59	Were outliers mentioned
60	What method was used to handle outliers

61	Any further details on outliers
62	Was there a control group
63	How was the control group incorporated?
64	Was there any forecasting

	Effect measures
65	Effect measure
66	Further details on effect measure
67	Metric
68	Numeric value of effect measure
69	Lower confidence interval
70	Upper confidence interval
71	Standard deviation
72	p-value
73	Notes on the effect measure
74	Was there any mention of a ceiling/floor effect
75	Details of the ceiling/floor effect

	Segments
76	Segment number
77	Segment description
78	What type of time interval is used in this segment
79	Is the timing of the segment clearly defined
80	How many intervals are in the segment
81	How many observations are in the segment
82	Please summarise any further segment details…

Appendix B. Additional file 1 accompanying Chapter 3 – Deviations, additions and amendments to the protocol

“Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: A review”

Simon L Turner, Amalia Karahalios, Andrew B Forbes, Monica Taljaard, Jeremy M Grimshaw, Allen C Cheng, Lisa Bero, Joanne E McKenzie

The following tables contain deviations, additions and amendments to the review protocol (52).

Table B: Additions to the data extraction form

Data extraction item	Data extraction item label	Response options
What type of control/comparison was used	Type of control	Response options are from Lopez Bernal et al 2018(74): • Behaviour (a group of individuals who never performed the behaviour being investigated) • Characteristic (a group not targeted by an intervention [e.g. for an intervention targeted at only males, controls may be selected from females]) • Historical (a historical cohort that has not been targeted by the intervention [e.g. historical years, where those years are not included in the intervention time series]) • Location (a location similar to the study location that is not targeted by the intervention [e.g. different hospital, different geographical area]) • Outcome (the same group but a different outcome that is predicted to be unaffected by the intervention, but would be affected by confounding events)
How was the control/comparison used?	Reporting of control	• Narrative (where differences in the series were described, or effect measures were stated separately for each group, or both, but no formal statistical comparison was undertaken) • Statistical (where the control series is formally incorporated in the statistical model) • Unclear
How was any delay between interruption and impact dealt with?	Impact delay method	• Delay (where the delay was acknowledged and included in pre- or post-interruption segment) • Excluded (where a separate segment was used for the delay time period, but this was *excluded* from the analysis)

		• Segment (where a separate segment was used for the delay time period, and this was *included* in analysis) • Sensitivity (where the delay was modelled as part of a sensitivity analysis, but ignored in main analysis) • Unclear
Was the standard error reported for item?	Standard error reported	• Yes • No • Unclear
Was a confidence interval reported for item?	Confidence interval reported	• Yes • No • Unclear
How was any Durbin Watson calculation reported?	Durbin Watson reporting	• No detail (where a statement is made that the Durbin Watson test was carried out, but with no further detail provided) • None detected (where a statement is made that the Durbin Watson test showed no evidence of autocorrelation) • Fully reported (where the actual Durbin Watson statistic is reported)
Was the segment used in the analysis? For each segment, an assessment was made as to whether it was included in the analysis.	Segment used	• Yes • No • Unclear
Method for handling autocorrelation in non-ARIMA models	Autocorrelation_handled	• Adjusts for autocorrelation • No adjustment (not required based on test or sensitivity analysis) • Can't determine/not considered

(74) Lopez Bernal J, Cummins S, Gasparrini A. The use of controls in interrupted time series studies of public health interventions. *International journal of epidemiology* 2018;47(6):2082-93. doi: 10.1093/ije/dyy135

Table C:Clarification and elaboration of the data extraction processes

Clarifications/elaborations of process	Detail of the clarification or elaboration from the protocol
Abstract screening	Abstract screening was performed on an initial sample of 200 from each year. If 40 included studies were not identified from this initial sample in a given year, further abstracts were screened until another potential 20 inclusions were obtained. These were then discussed in a group. This process continued until there were at least 40 included studies per year. Studies were included in order of their randomly generated number.
Outcome selection	The following elaboration was developed to aid in the selection of outcomes: If graph(s) exists, choose the first outcome listed as primary outcome (described as “primary outcome” in the abstract or methods section). If not… Graph(s) exists, choose the first outcome mentioned in the abstract (in words or numerically, independent of the analysis method) If not… Graph(s) exists, choose the first outcome mentioned in the paper (i.e. methods, results, discussion). If no graphs exist, check paper for mentioned outcomes of that type… Graph(s) don’t exist, choose the first outcome listed as primary outcome (described as “primary outcome” in the abstract or methods section). If not… Graph(s) don’t exist, choose the first outcome mentioned in the abstract (in words or numerically, independent of the analysis method) If not… Graph(s) don’t exist, choose the first one mentioned in the paper (i.e. methods, results, discussion).

Primary outcome designation	If there is only one outcome it should be considered the primary outcome for the outcome selection process.
Subgroup analysis selection	If the first mentioned analysis is from a subgroup, extract information related to this subgroup. Note that there are other subgroup analyses available, and of what type.
Unadjusted/Adjusted selection	If both unadjusted and adjusted analyses are reported, information should be collected on both, and a note made as to whether the analyses are unadjusted or adjusted.
Other/Unclear potential reclassification	When the response options *Other* or *Unclear* are selected for any items, these should be discussed by the group for the purpose of deciding whether they may be classified using any of the other response options.

Table D:Clarification, changes and elaboration to data extraction items

Clarification / change / elaboration of item	**Data extraction item label**	**Detail of the clarification or elaboration to the data extraction item**
Outcome type classification: Defined Daily Dose	Type of outcome	The outcome defined daily dose (DDD) should be classified as a continuous data type. “Defined Daily Dose (DDD): The assumed average maintenance dose per day for a drug used for its main indication in adults.” (Reference: WHO http://www.who.int/medicines/regulation/medicines-safety/toolkit_ddd/en/)
Outcome type classification: rates	Type of outcome	Although all data points in a time series may be viewed as rates, since each point represents a summary over a period of time (e.g. months), only classify the aggregate outcomes type as ‘rate’ if the measure is expressed per time period (e.g. DDDs per bed-day or incidence of X per 100000 person-years).
Clarification regarding the recording of delayed effects	Was a delay from interruption to its impact mentioned?	Delayed effects refer to the time delay between the interruption and the impact of that interruption on the outcome. If the date given for the interruption was at an earlier time point than the modelled impact of the interruption we recorded that there was a delay.

Choosing the statistical model used in the analysis	Statistical model used	The statistical method should not be assumed unless details are explicitly stated. For example, presentation of the statistical model (e.g. $Y_t = \beta_0 + \beta_1 t + \beta_2 D_t + \beta_3[t - T_1]D_t + \varepsilon_t$) without further elaboration of the method should be rated as ‘can’t determine’. Similarly, if *only* the statistical software is listed (e.g. *itsa* in Stata, or *PROC AUTOREG* in SAS), the statistical method should be rated ‘can’t determine’, since multiple options are available to fit different methods within each package. The response option ‘other’ should be used when a method is clearly reported, but is not one of the predefined response options. In papers where the response option ‘can’t determine’ or ‘other’ is selected for the statistical analysis, these should discussed with multiple authors. In instances where multiple methods are reported, but it is not clear which method was used for the selected series, the latter method should be selected. Examples of classification decisions follow:

Example	Selected response option	Rationale/action
The reported effect estimates are odds ratios, but there is no mention in the methods (or elsewhere) that logistic regression (or other method) was used.	“Can’t determine”	The authors incompletely reported the analysis method.
It was not possible to determine if an ordinary least squares (OLS) or generalised least squares (GLS) method was used.	“Can’t determine”	The authors incompletely reported the analysis method.

The model parameterisation is provided, but without further information provided about the statistical analysis method.	“Can’t determine”	The model parametrisation does not define the statistical estimation method.
Only the statistical package is reported, with no further detail of the statistical method (e.g. “To account for autocorrelation PROC AUTOREG was used”).	“Can’t determine”	For a particular statistical package, there are often multiple options available that implement different statistical methods.
The statistical method is reported as “segmented linear regression”, but does not mention autocorrelation.	“OLS” or “Can’t determine”	Discussed by the author group to decide whether there is enough information to rate this as “OLS” or “can’t determine”.
The authors report using Newey-West standard errors, but without specifically stating they used ordinary least squares.	“OLS”	Newey-West standard errors are applied to ordinary least squares regression.
When multiple methods were reported, and it was not clear which method was used for the selected series, we selected the latter method (e.g. “Ordinary least squares regression was used”… “Autocorrelation was found so we adjusted using Prais-Winsten.”).	“GLS”	The decision to extract the latter method was arbitrary.

The reported statistical method is Prais Winsten or Cochrane Orcutt.	“GLS”	These statistical methods are types of generalised least squares.
The statistical method is completely described, but does not fall within one of the predefined response options.	“Other”	Discussed by the author group to confirm the selected response option.

Change to statistical model response options	Statistical model used	Two of the response options (OLS, GLS) for this item were modified as a result of the peer-review process. Specifically, the response options OLS and GLS were re-categorised into the following categories: • Linear regression without adjustment for autocorrelation [previously OLS; OLS with Newey-West standard errors] • Linear regression with adjustment for autocorrelation [previously GLS; GLS - Prais-Winsten; GLS - Cochrane-Orcutt] • Linear regression where it cannot be determined if there was adjustment for autocorrelation [previously OLS]
Clarification regarding non-stationarity	Was non-stationarity mentioned	If the Box Jenkins method was referenced, we rated the ‘non-stationarity’ item as ‘yes’.
Elaboration regarding recording the number of time points per segment	Segment number of intervals	Note the number of points in each time segment (e.g. pre-intervention, transition, post-intervention) when reported or can be determined (e.g. from graphs), as well as which of these segments were used in the analysis.

Appendix C. Additional file 2 accompanying Chapter 3 – Review search terms

“Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: A review”

Simon L Turner, Amalia Karahalios, Andrew B Forbes, Monica Taljaard, Jeremy M Grimshaw, Allen C Cheng, Lisa Bero, Joanne E McKenzie

The following table contains the search terms used in the review (52).

Review Search Terms

Search (#)	Search terms
1	Interrupted time series analysis (MeSH term)
2	“Interrupted time series” (title/abstract)
3	“Change point” (title/abstract)
4	“Segmented regression” (title/abstract)
5	“Segmented linear regression” (title/abstract)
6	“Repeated measures study” (title/abstract)
7	“Piecewise regression” (title/abstract)
8	“Time-series intervention” (title/abstract)
9	“Phase design” (title/abstract)
10	“Multiple baseline” (title/abstract)
11	“ARIMA” (title/abstract)
12	“Integrated moving average” (title/abstract)
13	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12

Abbreviations: MeSH, medical subject headings; ARIMA, autoregressive integrated moving average

Appendix D. Additional file 3 accompanying Chapter 3 – Citation details of the 200 studies from which data were extracted

"Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: A review"

Simon L Turner, Amalia Karahalios, Andrew B Forbes, Monica Taljaard, Jeremy M Grimshaw, Allen C Cheng, Lisa Bero, Joanne E McKenzie

Citation details of the 200 studies from which data were extracted.

1. Abegaz T, Berhane Y, Worku A, et al. Effectiveness of an improved road safety policy in Ethiopia: an interrupted time series study. BMC Public Health 2014;14(1) doi: 10.1186/1471-2458-14-539
2. Adams AS, Soumerai SB, Zhang F, et al. Effects of Eliminating Drug Caps on Racial Differences in Antidepressant Use Among Dual Enrollees With Diabetes and Depression. Clinical Therapeutics 2015;37(3):597-609. doi: 10.1016/j.clinthera.2014.12.011
3. Aiken AM, Wanyoro AK, Mwangi J, et al. Changing Use of Surgical Antibiotic Prophylaxis in Thika Hospital, Kenya: A Quality Improvement Intervention with an Interrupted Time Series Design. PLoS ONE 2013;8(11):e78942. doi: 10.1371/journal.pone.0078942
4. Akhtar S, Ziyab AH. Impact of the Penalty Points System on Severe Road Traffic Injuries in Kuwait. Traffic Injury Prevention 2013;14(7):743-48. doi: 10.1080/15389588.2012.749466
5. Alexandridis AA, McCort A, Ringwalt CL, et al. A statewide evaluation of seven strategies to reduce opioid overdose in North Carolina. Injury Prevention 2017;24(1):48-54. doi: 10.1136/injuryprev-2017-042396
6. Alpert HR, Carpenter D, Connolly GN. Tobacco industry response to a ban on lights descriptors on cigarette packaging and population outcomes. Tobacco Control 2017;27(4):390-98. doi: 10.1136/tobaccocontrol-2017-053683
7. Andersen SE, Knudsen JD. A managed multidisciplinary programme on multi-resistant Klebsiella pneumonia in a Danish university hospital. BMJ Quality & Safety 2013;22(11):907-15. doi: 10.1136/bmjqs-2012-001791
8. Andrade AL, Minamisava R, Policena G, et al. Evaluating the impact of PCV-10 on invasive pneumococcal disease in Brazil: A time-series analysis. Human Vaccines & Immunotherapeutics 2016;12(2):285-92. doi: 10.1080/21645515.2015.1117713
9. Armah G, Pringle K, Enweronu-Laryea CC, et al. Impact and Effectiveness of Monovalent Rotavirus Vaccine Against Severe Rotavirus Diarrhea in Ghana. Clinical Infectious Diseases 2016;62(suppl 2):S200-S07. doi: 10.1093/cid/ciw014
10. Barber C, Gagnon D, Fonda J, et al. Assessing the impact of prescribing directives on opioid prescribing practices among Veterans Health Administration providers. Pharmacoepidemiology and Drug Safety 2016;26(1):40-46. doi: 10.1002/pds.4066
11. Barocas DA, Mallin K, Graves AJ, et al. Effect of the USPSTF Grade D Recommendation against Screening for Prostate Cancer on Incident Prostate Cancer Diagnoses in the United States. Journal of Urology 2015;194(6):1587-93. doi: 10.1016/j.juro.2015.06.075
12. Baskerville NB, Brown KS, Nguyen NC, et al. Impact of Canadian tobacco packaging policy on use of a toll-free quit-smoking line: an interrupted time-series analysis. CMAJ Open 2016;4(1):E59-E65. doi: 10.9778/cmajo.20150104

13. Been JV, Mackay DF, Millett C, et al. Impact of smoke-free legislation on perinatal and infant mortality: a national quasi-experimental study. Scientific Reports 2015;5(1) doi: 10.1038/srep13020
14. Been JV, Szatkowski L, van Staa T-P, et al. Smoke-free legislation and the incidence of paediatric respiratory infections and wheezing/asthma: interrupted time series analyses in the four UK nations. Scientific Reports 2015;5(1) doi: 10.1038/srep15246
15. Bell S, Davey P, Nathwani D, et al. Risk of AKI with Gentamicin as Surgical Prophylaxis. Journal of the American Society of Nephrology 2014;25(11):2625-32. doi: 10.1681/asn.2014010035
16. Bendzsak AM, Baxter NN, Darling GE, et al. Regionalization and Outcomes of Lung Cancer Surgery in Ontario, Canada. Journal of Clinical Oncology 2017;35(24):2772-80. doi: 10.1200/jco.2016.69.8076
17. Berkowitz SA, Percac-Lima S, Ashburner JM, et al. Building Equity Improvement into Quality Improvement: Reducing Socioeconomic Disparities in Colorectal Cancer Screening as Part of Population Health Management. Journal of General Internal Medicine 2015;30(7):942-49. doi: 10.1007/s11606-015-3227-4
18. Bernat DH, Maldonado-Molina M, Hyland A, et al. Effects of Smoke-Free Laws on Alcohol-Related Car Crashes in California and New York: Time Series Analyses From 1982 to 2008. American Journal of Public Health 2013;103(2):214-20. doi: 10.2105/ajph.2012.300906
19. Blais E, Carnis L. Improving the safety effect of speed camera programs through innovations: Evidence from the French experience. Journal of Safety Research 2015;55:135-45. doi: 10.1016/j.jsr.2015.08.007
20. Bobo WV, Epstein RA, Hayes RM, et al. The effect of regulatory advisories on maternal antidepressant prescribing, 1995–2007: an interrupted time series study of 228,876 pregnancies. Archives of Women's Mental Health 2013;17(1):17-26. doi: 10.1007/s00737-013-0383-6
21. Boden DG, Agarwal A, Hussain T, et al. Lowering levels of bed occupancy is associated with decreased in hospital mortality and improved performance on the 4-hour target in a UK District General Hospital. Emergency Medicine Journal 2015;33(2):85-90. doi: 10.1136/emermed-2014-204479
22. Boel J, Andreasen V, Jarløv JO, et al. Impact of antibiotic restriction on resistance levels of Escherichia coli: a controlled interrupted time series study of a hospital-wide antibiotic stewardship programme. Journal of Antimicrobial Chemotherapy 2016;71(7):2047-51. doi: 10.1093/jac/dkw055
23. Bonander C, Nilson F, Andersson R. The effect of the Swedish bicycle helmet law for children: An interrupted time series study. Journal of Safety Research 2014;51:15-22. doi: 10.1016/j.jsr.2014.07.001
24. Borde JP, Kern WV, Hug M, et al. Implementation of an intensified antibiotic stewardship programme targeting third-generation cephalosporin and fluoroquinolone use in an emergency medicine department. Emergency Medicine Journal 2014;32(7):509-15. doi: 10.1136/emermed-2014-204067
25. Bowden JA, Dono J, John DL, et al. What happens when the price of a tobacco retailer licence increases? Tobacco Control 2013;23(2):178-80. doi: 10.1136/tobaccocontrol-2012-050615
26. Bozorgmehr K, Razum O. Effect of Restricting Access to Health Care on Health Expenditures among Asylum-Seekers and Refugees: A Quasi-Experimental Study in Germany, 1994–2013. PLOS ONE 2015;10(7):e0131483. doi: 10.1371/journal.pone.0131483
27. Branas CC, Kastanaki AE, Michalodimitrakis M, et al. The impact of economic austerity and prosperity events on suicide in Greece: a 30-year interrupted time-series analysis. BMJ Open 2015;5(1):e005619-e19. doi: 10.1136/bmjopen-2014-005619

28. Bugden S, Friesen KJ, Falk J. Voluntary warnings and the limits of good prescribing behavior: the case for de-adoption of meperidine. Journal of Pain Research 2015:879. doi: 10.2147/jpr.s96625
29. Burke LK, Brown CP, Johnson TM. Historical Data Analysis of Hospital Discharges Related to the Amerithrax Attack in Florida. Perspect Health Inf Manag 2016;13(Fall):1c-1c.
30. Busch SH, McGinty EE, Stuart EA, et al. Was federal parity associated with changes in Out-of-network mental health care use and spending? BMC Health Services Research 2017;17(1) doi: 10.1186/s12913-017-2261-9
31. Cairns KA, Jenney AWJ, Abbott IJ, et al. Prescribing trends before and after implementation of an antimicrobial stewardship program. The Medical Journal of Australia 2013;198(5):262-66. doi: 10.5694/mja12.11683
32. Carracedo-Martínez E, Pia-Morandeira A, Figueiras A. Trends in celecoxib and etoricoxib prescribing following removal of prior authorization requirement in Spain. Journal of Clinical Pharmacy and Therapeutics 2016;42(2):185-88. doi: 10.1111/jcpt.12490
33. Cecil E, Bottle A, Sharland M, et al. Impact of UK Primary Care Policy Reforms on Short-Stay Unplanned Hospital Admissions for Children With Primary Care-Sensitive Conditions. The Annals of Family Medicine 2015;13(3):214-20. doi: 10.1370/afm.1786
34. Chandran A, Pérez-Núñez R, Bachani AM, et al. Early Impact of a National Multi-Faceted Road Safety Intervention Program in Mexico: Results of a Time-Series Analysis. PLoS ONE 2014;9(1):e87482. doi: 10.1371/journal.pone.0087482
35. Chang C-H, Lin J-W, Wu L-C, et al. National Antiviral Treatment Program and the Incidence of Hepatocellular Carcinoma and Associated Mortality in Taiwan. Medical Care 2013;51(10):908-13. doi: 10.1097/mlr.0b013e3182a502ba
36. Chen IL, Lee C-H, Su L-H, et al. Effects of implementation of an online comprehensive antimicrobial-stewardship program in ICUs: A longitudinal study. Journal of Microbiology, Immunology and Infection 2018;51(1):55-63. doi: 10.1016/j.jmii.2016.06.007
37. Cheng C-L, Chao P-H, Hsu JC-S, et al. Utilization patterns of Antihyperuricemic Agents Following Safety Announcement on Allopurinol and Benzbromarone by Taiwan Food and Drug Administration. Pharmacoepidemiology and Drug Safety 2013;23(3):309-13. doi: 10.1002/pds.3550
38. Cheng J, Benassi P, de Oliveira C, et al. Impact of a mass media mental health campaign on psychiatric emergency department visits. Canadian Journal of Public Health 2016;107(3):e303-e11. doi: 10.17269/cjph.107.5265
39. Chua K-P, Shrime MG, Conti RM. Effect of FDA Investigation on Opioid Prescribing to Children After Tonsillectomy/Adenoidectomy. Pediatrics 2017;140(6):e20171765. doi: 10.1542/peds.2017-1765
40. Chung YK, Kim J-S, Lee SS, et al. Effect of daily chlorhexidine bathing on acquisition of carbapenem-resistant Acinetobacter baumannii (CRAB) in the medical intensive care unit with CRAB endemicity. American Journal of Infection Control 2015;43(11):1171-77. doi: 10.1016/j.ajic.2015.07.001
41. Čižman M, Plankar Srovin T, Blagus R, et al. The long-term effects of restrictive interventions on consumption and costs of antibiotics. Journal of Global Antimicrobial Resistance 2015;3(1):31-35. doi: 10.1016/j.jgar.2014.11.004
42. Corcoran P, Griffin E, Arensman E, et al. Impact of the economic recession and subsequent austerity on suicide and self-harm in Ireland: An interrupted time series analysis. International Journal of Epidemiology 2015;44(3):969-77. doi: 10.1093/ije/dyv058
43. Cunningham JK, Liu L-M, Callaghan RC. Essential ("Precursor") chemical control for heroin: Impact of acetic anhydride regulation on US heroin availability. Drug and Alcohol Dependence 2013;133(2):520-28. doi: 10.1016/j.drugalcdep.2013.07.014

44. Damiani G, Federico B, Anselmi A, et al. The impact of Regional co-payment and National reimbursement criteria on statins use in Italy: an interrupted time-series analysis. BMC Health Services Research 2014;14(1) doi: 10.1186/1472-6963-14-6
45. Denkel LA, Schwab F, Garten L, et al. Protective Effect of Dual-Strain Probiotics in Preterm Infants: A Multi-Center Time Series Analysis. PLOS ONE 2016;11(6):e0158136. doi: 10.1371/journal.pone.0158136
46. Desai SP, Lu B, Szent-Gyorgyi LE, et al. Increasing pneumococcal vaccination for immunosuppressed patients: A cluster quality improvement trial. Arthritis & Rheumatism 2012;65(1):39-47. doi: 10.1002/art.37716
47. Deslandes PN, Jenkins KSL, Haines KE, et al. A change in the trend in dosulepin usage following the introduction of a prescribing indicator but not after two national safety warnings. Journal of Clinical Pharmacy and Therapeutics 2016;41(2):224-28. doi: 10.1111/jcpt.12376
48. Dicks KV, Lofgren E, Lewis SS, et al. A Multicenter Pragmatic Interrupted Time Series Analysis of Chlorhexidine Gluconate Bathing in Community Hospital Intensive Care Units. Infection Control & Hospital Epidemiology 2016;37(7):791-97. doi: 10.1017/ice.2016.23
49. Dik J-WH, Hendrix R, Lo-Ten-Foe JR, et al. Automatic day-2 intervention by a multidisciplinary antimicrobial stewardship-team leads to multiple positive effects. Frontiers in Microbiology 2015;06 doi: 10.3389/fmicb.2015.00546
50. DiMaggio C, Chen Q, Muennig PA, et al. Timing and effect of a safe routes to school program on child pedestrian injury risk during school travel hours: Bayesian changepoint and difference-in-differences analysis. Injury Epidemiology 2014;1(1) doi: 10.1186/s40621-014-0017-0
51. Doernberg SB, Dudas V, Trivedi KK. Implementation of an antimicrobial stewardship program targeting residents with urinary tract infections in three community long-term care facilities: a quasi-experimental study using time-series analysis. Antimicrobial Resistance and Infection Control 2015;4(1) doi: 10.1186/s13756-015-0095-y
52. Dresden SM, Powell ES, Kang R, et al. Increased Emergency Department Use in Illinois After Implementation of the Patient Protection and Affordable Care Act. Annals of Emergency Medicine 2017;69(2):172-80. doi: 10.1016/j.annemergmed.2016.06.026
53. Druetz T, Fregonese F, Bado A, et al. Abolishing Fees at Health Centers in the Context of Community Case Management of Malaria: What Effects on Treatment-Seeking Practices for Febrile Children in Rural Burkina Faso? PLOS ONE 2015;10(10):e0141306. doi: 10.1371/journal.pone.0141306
54. Emmerick ICM, Campos MR, Luiza VL, et al. Retrospective interrupted time series examining hypertension and diabetes medicines usage following changes in patient cost sharing in the ‘Farmácia Popular’ programme in Brazil. BMJ Open 2017;7(11):e017308. doi: 10.1136/bmjopen-2017-017308
55. Faryar KA, Freeman CL, Persaud AK, et al. The Effects of Kentucky's Comprehensive Opioid Legislation on Patients Presenting with Prescription Opioid or Heroin Abuse to One Urban Emergency Department. The Journal of Emergency Medicine 2017;53(6):805-14. doi: 10.1016/j.jemermed.2017.08.066
56. Filippidis FT, Gerovasili V, Millett C, et al. Medium-term impact of the economic crisis on mortality, health-related behaviours and access to healthcare in Greece. Scientific Reports 2017;7(1) doi: 10.1038/srep46423
57. Finnell KJ, John R, Thompson DM. 1% low-fat milk has perks!: An evaluation of a social marketing intervention. Preventive Medicine Reports 2017;5:144-49. doi: 10.1016/j.pmedr.2016.11.017
58. Fisher D, Tambyah PA, Lin RTP, et al. Sustained meticillin-resistant Staphylococcus aureus control in a hyper-endemic tertiary acute care hospital with infrastructure challenges in Singapore. Journal of Hospital Infection 2013;85(2):141-48. doi: 10.1016/j.jhin.2013.07.005

59. Flett KB, Ozonoff A, Graham DA, et al. Impact of Mandatory Public Reporting of Central Line–Associated Bloodstream Infections on Blood Culture and Antibiotic Utilization in Pediatric and Neonatal Intensive Care Units. Infection Control & Hospital Epidemiology 2015;36(8):878-85. doi: 10.1017/ice.2015.100
60. Flynn D, Ford GA, Rodgers H, et al. A Time Series Evaluation of the FAST National Stroke Awareness Campaign in England. PLoS ONE 2014;9(8):e104289. doi: 10.1371/journal.pone.0104289
61. Fournier P, Dumont A, Tourigny C, et al. The Free Caesareans Policy in Low-Income Settings: An Interrupted Time Series Analysis in Mali (2003–2012). PLoS ONE 2014;9(8):e105130. doi: 10.1371/journal.pone.0105130
62. Gadzhanova SV, Roughead EE, Bartlett MJ. Improving cardiovascular disease management in Australia: NPS MedicineWise. The Medical Journal of Australia 2013;199(3):192-95. doi: 10.5694/mja12.11779
63. Gale M, Muscatello DJ, Dinh M, et al. Alcopops, taxation and harm: a segmented time series analysis of emergency department presentations. BMC Public Health 2015;15(1) doi: 10.1186/s12889-015-1769-3
64. Gallini A, Andrieu S, Donohue JM, et al. Trends in use of antipsychotics in elderly patients with dementia: Impact of national safety warnings. European Neuropsychopharmacology 2014;24(1):95-104. doi: 10.1016/j.euroneuro.2013.09.003
65. Gamble J-M, Johnson JA, Majumdar SR, et al. Evaluating the introduction of a computerized prior-authorization system on the completeness of drug exposure data. Pharmacoepidemiology and Drug Safety 2013;22(5):551-55. doi: 10.1002/pds.3427
66. Garnett M, Charyk Stewart T, Miller MR, et al. Did Amendments to the Ontario Highway Traffic Act in 2009-2010 Affect the Proportion of Alcohol-Related Motor Vehicle Collisions Seen at a Level I Trauma Centre over a 10-year Period? CJEM 2016;19(2):106-11. doi: 10.1017/cem.2016.343
67. Gaudreau K, Sanford CJ, Cheverie C, et al. The Effect of a Smoking Ban on Hospitalization Rates for Cardiovascular and Respiratory Conditions in Prince Edward Island, Canada. PLoS ONE 2013;8(3):e56102. doi: 10.1371/journal.pone.0056102
68. Gebrehiwot TG, San Sebastian M, Edin K, et al. The Health Extension Program and Its Association with Change in Utilization of Selected Maternal Health Services in Tigray Region, Ethiopia: A Segmented Linear Regression Analysis. PLOS ONE 2015;10(7):e0131195. doi: 10.1371/journal.pone.0131195
69. Gefenaite G, Bijlsma M, Bos H, et al. Did introduction of pneumococcal vaccines in the Netherlands decrease the need for respiratory antibiotics in children? Analysis of 2002 to 2013 data. Eurosurveillance 2014;19(44):20948. doi: 10.2807/1560-7917.es2014.19.44.20948
70. Gilbert C, Darlow B, Zin A, et al. Educating Neonatal Nurses in Brazil: A Before-and-After Study with Interrupted Time Series Analysis. Neonatology 2014;106(3):201-08. doi: 10.1159/000362532
71. Glantz SA, Gibbs E. Changes in Ambulance Calls After Implementation of a Smoke-Free Law and Its Extension to Casinos. Circulation 2013;128(8):811-13. doi: 10.1161/circulationaha.113.003455
72. Gobin M, Verlander N, Maurici C, et al. Do sexual health campaigns work? An outcome evaluation of a media campaign to increase chlamydia testing among young people aged 15–24 in England. BMC Public Health 2013;13(1) doi: 10.1186/1471-2458-13-484
73. Godman B, Persson M, Miranda J, et al. Changes in the Utilization of Venlafaxine after the Introduction of Generics in Sweden. Applied Health Economics and Health Policy 2013;11(4):383-93. doi: 10.1007/s40258-013-0037-x
74. Godman B, Wettermark B, Miranda J, et al. Influence of multiple initiatives in Sweden to enhance ARB prescribing efficiency following generic losartan; findings and implications

for other countries. International Journal of Clinical Practice 2013;67(9):853-62. doi: 10.1111/ijcp.12130

75. Gold R, Nelson C, Cowburn S, et al. Feasibility and impact of implementing a private care system's diabetes quality improvement intervention in the safety net: a cluster-randomized trial. Implementation Science 2015;10(1) doi: 10.1186/s13012-015-0259-4
76. Graves AJ, Kozhimannil KB, Kleinman KP, et al. The Association between High-Deductible Health Plan Transition and Contraception and Birth Rates. Health Services Research 2015;51(1):187-204. doi: 10.1111/1475-6773.12326
77. Guthrie B, Clark SA, Reynish EL, et al. Differential Impact of Two Risk Communications on Antipsychotic Prescribing to People with Dementia in Scotland: Segmented Regression Time Series Analysis 2001–2011. PLoS ONE 2013;8(7):e68976. doi: 10.1371/journal.pone.0068976
78. Haas JP, Menz J, Dusza S, et al. Implementation and impact of ultraviolet environmental disinfection in an acute care setting. American Journal of Infection Control 2014;42(6):586-90. doi: 10.1016/j.ajic.2013.12.013
79. Haggins A, Patrick S, Demonner S, et al. When Coverage Expands: Children's Health Insurance Program as a Natural Experiment in Use of Health Care Services. Academic Emergency Medicine 2013;20(10):1026-32. doi: 10.1111/acem.12236
80. Halim S, Jiang H. The effect of Operation 24 Hours on reducing collision in the City of Edmonton. Accident Analysis & Prevention 2013;58:106-14. doi: 10.1016/j.aap.2013.04.031
81. Hamilton I, Lloyd C, Bland JM, et al. The impact of assertive outreach teams on hospital admissions for psychosis: a time series analysis. Journal of Psychiatric and Mental Health Nursing 2015;22(7):484-90. doi: 10.1111/jpm.12239
82. Hanatani T, Sai K, Tohkin M, et al. Evaluation of two Japanese regulatory actions using medical information databases: a 'Dear Doctor' letter to restrict oseltamivir use in teenagers, and label change caution against co-administration of omeprazole with clopidogrel. Journal of Clinical Pharmacy and Therapeutics 2014;39(4):361-67. doi: 10.1111/jcpt.12153
83. Hansen BT, Østergaard SD, Sønderskov KM, et al. Increased Incidence Rate of Trauma- and Stressor-Related Disorders in Denmark After the September 11, 2001, Terrorist Attacks in the United States. American Journal of Epidemiology 2016;184(7):494-500. doi: 10.1093/aje/kww089
84. Hansen BT, Sønderskov KM, Hageman I, et al. Daylight Savings Time Transitions and the Incidence Rate of Unipolar Depressive Episodes. Epidemiology 2017;28(3):346-53. doi: 10.1097/ede.0000000000000580
85. Harper S, Bruckner TA. Did the Great Recession increase suicides in the USA? Evidence from an interrupted time-series analysis. Annals of Epidemiology 2017;27(7):409-14.e6. doi: 10.1016/j.annepidem.2017.05.017
86. Hartung DM, Middleton L, Markwardt S, et al. Changes in Long-acting β-agonist Utilization After the FDA's 2010 Drug Safety Communication. Clinical Therapeutics 2015;37(1):114-23.e1. doi: 10.1016/j.clinthera.2014.10.025
87. Hassanian-Moghaddam H, Ghorbani F, Rahimi A, et al. Federation Internationale de Football Association (FIFA) 2014 World Cup Impact on Hospital-Treated Suicide Attempt (Overdose) in Tehran. Suicide and Life-Threatening Behavior 2017;48(3):367-75. doi: 10.1111/sltb.12359
88. Hawton K, Bergen H, Geulayov G, et al. Impact of the recent recession on self-harm: Longitudinal ecological and patient-level investigation from the Multicentre Study of Self-harm in England. Journal of Affective Disorders 2016;191:132-38. doi: 10.1016/j.jad.2015.11.001

89. Hawton K, Bergen H, Simkin S, et al. Long term effect of reduced pack sizes of paracetamol on poisoning deaths and liver transplant activity in England and Wales: interrupted time series analyses. BMJ 2013;346(feb07 1):f403-f03. doi: 10.1136/bmj.f403
90. Hingwala J, Bhangoo S, Hiebert B, et al. Evaluating the Implementation Strategy for Estimated Glomerular Filtration Rate Reporting in Manitoba: The Effect on Referral Numbers, Wait Times, and Appropriateness of Consults. Canadian Journal of Kidney Health and Disease 2014;1:9. doi: 10.1186/2054-3581-1-9
91. Høgli JU, Garcia BH, Skjold F, et al. An audit and feedback intervention study increased adherence to antibiotic prescribing guidelines at a Norwegian hospital. BMC Infectious Diseases 2016;16(1) doi: 10.1186/s12879-016-1426-1
92. Honein-AbouHaidar GN, Rabeneck L, Paszat LF, et al. Evaluating the impact of public health initiatives on trends in fecal occult blood test participation in Ontario. BMC Cancer 2014;14(1) doi: 10.1186/1471-2407-14-537
93. Horton DB, Gerhard T, Davidow A, et al. Impact of the black triangle label on prescribing of new drugs in the United Kingdom: lessons for the United States at a time of deregulation. Pharmacoepidemiology and Drug Safety 2017;26(11):1307-13. doi: 10.1002/pds.4304
94. Hostenkamp G, Fischer KE, Borch-Johnsen K. Drug safety and the impact of drug warnings: An interrupted time series analysis of diabetes drug prescriptions in Germany and Denmark. Health Policy 2016;120(12):1404-11. doi: 10.1016/j.healthpol.2016.09.020
95. Hsu JC, Cheng C-L, Ross-Degnan D, et al. Effects of safety warnings and risk management plan for Thiazolidinediones in Taiwan. Pharmacoepidemiology and Drug Safety 2015;24(10):1026-35. doi: 10.1002/pds.3834
96. Hsu JC, Lu CY, Wagner AK, et al. Impacts of drug reimbursement reductions on utilization and expenditures of oral antidiabetic medications in Taiwan: An interrupted time series study. Health Policy 2014;116(2-3):196-205. doi: 10.1016/j.healthpol.2013.11.005
97. Huitema BE, Van Houten R, Manal H. Time-series intervention analysis of pedestrian countdown timer effects. Accident Analysis & Prevention 2014;72:23-31. doi: 10.1016/j.aap.2014.05.025
98. Humphreys DK, Gasparrini A, Wiebe DJ. Evaluating the Impact of Florida's “Stand Your Ground” Self-defense Law on Homicide and Suicide by Firearm. JAMA Internal Medicine 2017;177(1):44. doi: 10.1001/jamainternmed.2016.6811
99. Iams W, Heck J, Kapp M, et al. A Multidisciplinary Housestaff-Led Initiative to Safely Reduce Daily Laboratory Testing. Academic Medicine 2016;91(6):813-20. doi: 10.1097/acm.0000000000001149
100. Jenkins TC, Knepper BC, Shihadeh K, et al. Long-Term Outcomes of an Antimicrobial Stewardship Program Implemented in a Hospital with Low Baseline Antibiotic Use. Infection Control & Hospital Epidemiology 2015;36(6):664-72. doi: 10.1017/ice.2015.41
101. Jiao B, Kim S, Hagen J, et al. Cost-effectiveness of neighbourhood slow zones in New York City. Injury Prevention 2017;25(2):98-103. doi: 10.1136/injuryprev-2017-042499
102. Johri M, Ridde V, Heinmüller R, et al. Estimation of maternal and child mortality one year after user-fee elimination: an impact evaluation and modelling study in Burkina Faso. Bulletin of the World Health Organization 2014;92(10):706-15. doi: 10.2471/blt.13.130609
103. Katikireddi SV, Der G, Roberts C, et al. Has Childhood Smoking Reduced Following Smoke-Free Public Places Legislation? A Segmented Regression Analysis of Cross-Sectional UK School-Based Surveys. Nicotine & Tobacco Research 2016;18(7):1670-74. doi: 10.1093/ntr/ntw018
104. Kesselheim AS, Donneyong M, Dal Pan GJ, et al. Changes in prescribing and healthcare resource utilization after FDA Drug Safety Communications involving zolpidem-

containing medications. Pharmacoepidemiology and Drug Safety 2017;26(6):712-21. doi: 10.1002/pds.4215

105. Kim B, Kim K, Lee J, et al. Impact of bacteremia prediction rule in CAP: Before and after study. The American Journal of Emergency Medicine 2018;36(5):758-62. doi: 10.1016/j.ajem.2017.10.005

106. Kim J-S, Chung YK, Lee SS, et al. Effect of daily chlorhexidine bathing on the acquisition of methicillin-resistant Staphylococcus aureus in a medical intensive care unit with methicillin-resistant S aureus endemicity. American Journal of Infection Control 2016;44(12):1520-25. doi: 10.1016/j.ajic.2016.04.252

107. Kim SH, Cho BL, Shin DW, et al. The Effect of Asthma Clinical Guideline for Adults on Inhaled Corticosteroids PrescriptionTrend: A Quasi-Experimental Study. Journal of Korean Medical Science 2015;30(8):1048. doi: 10.3346/jkms.2015.30.8.1048

108. Kiran T, Wilton AS, Moineddin R, et al. Effect of Payment Incentives on Cancer Screening in Ontario Primary Care. The Annals of Family Medicine 2014;12(4):317-23. doi: 10.1370/afm.1664

109. Kisely S, Crowe E, Lawrence D, et al. A time series analysis of presentations to Queensland health facilities for alcohol-related conditions, following the increase in 'alcopops' tax. Australasian Psychiatry 2013;21(4):383-88. doi: 10.1177/1039856213486307

110. Klein EG, Forster JL, Toomey TL, et al. Did a local clean indoor air policy increase alcohol-related crime around bars and restaurants? Tobacco Control 2011;22(2):113-17. doi: 10.1136/tobaccocontrol-2011-050010

111. Kolhatkar A, Cheng L, Chan FKI, et al. The impact of medication reviews by community pharmacists. Journal of the American Pharmacists Association 2016;56(5):513-20.e1. doi: 10.1016/j.japh.2016.05.002

112. Kontopantelis E, Olier I, Planner C, et al. Primary care consultation rates among people with and without severe mental illness: a UK cohort study using the Clinical Practice Research Datalink. BMJ Open 2015;5(12):e008650. doi: 10.1136/bmjopen-2015-008650

113. Kontopantelis E, Reeves D, Valderas JM, et al. Recorded quality of primary care for patients with diabetes in England before and after the introduction of a financial incentive scheme: a longitudinal observational study. BMJ Quality & Safety 2012;22(1):53-64. doi: 10.1136/bmjqs-2012-001033

114. Kracalik I, Abdullayev R, Asadov K, et al. Changing Patterns of Human Anthrax in Azerbaijan during the Post-Soviet and Preemptive Livestock Vaccination Eras. PLoS Neglected Tropical Diseases 2014;8(7):e2985. doi: 10.1371/journal.pntd.0002985

115. Kracalik IT, Abdullayev R, Asadov K, et al. Human Brucellosis Trends: Re-emergence and Prospects for Control Using a One Health Approach in Azerbaijan (1983-2009). Zoonoses and Public Health 2015;63(4):294-302. doi: 10.1111/zph.12229

116. Kruik-Kollöffel WJ, van der Palen J, Kruik HJ, et al. Prescription behavior for gastroprotective drugs in new users as a result of communications regarding clopidogrel - proton pump inhibitor interaction. Pharmacology Research & Perspectives 2016;4(4):e00242. doi: 10.1002/prp2.242

117. Larney S, Lai W, Dolan K, et al. Monitoring a Prison Opioid Treatment Program Over a Period of Change to Clinical Governance Arrangements, 2007–2013. Journal of Substance Abuse Treatment 2016;70:58-63. doi: 10.1016/j.jsat.2016.08.001

118. Lavergne MR, Law MR, Peterson S, et al. Effect of incentive payments on chronic disease management and health services use in British Columbia, Canada: Interrupted time series analysis. Health Policy 2018;122(2):157-64. doi: 10.1016/j.healthpol.2017.11.001

119. Lee KR, Bagga B, Arnold SR. Reduction of Broad-Spectrum Antimicrobial Use in a Tertiary Children's Hospital Post Antimicrobial Stewardship Program Guideline

Implementation*. Pediatric Critical Care Medicine 2016;17(3):187-93. doi: 10.1097/pcc.0000000000000615

120. Lee TC, Frenette C, Jayaraman D, et al. Antibiotic Self-stewardship: Trainee-Led Structured Antibiotic Time-outs to Improve Antimicrobial Use. Annals of Internal Medicine 2014;161(10_Supplement):S53. doi: 10.7326/m13-3016

121. Lee Y-J, Chen J-Z, Lin H-C, et al. Impact of active screening for methicillin-resistant Staphylococcus aureus (MRSA) and decolonization on MRSA infections, mortality and medical cost: a quasi-experimental study in surgical intensive care unit. Critical Care 2015;19(1) doi: 10.1186/s13054-015-0876-y

122. Li Z, Li M, Fink G, et al. User–fee–removal improves equity of children’s health care utilization and reduces families’ financial burden: evidence from Jamaica. Journal of Global Health 2017;7(1) doi: 10.7189/jogh.07.010502

123. Lieberman DA, Polinski JM, Choudhry NK, et al. Unintended Consequences of a Medicaid Prescription Copayment Policy. Medical Care 2014;52(5):422-27. doi: 10.1097/mlr.0000000000000119

124. Lin CM, Liao CM. Inpatient expenditures on alcohol-attributed diseases and alcohol tax policy: a nationwide analysis in Taiwan from 1996 to 2010. Public Health 2014;128(11):977-84. doi: 10.1016/j.puhe.2014.09.004

125. Lopez Bernal J, Gasparrini A, Artundo C, et al. RE: The effect of the late 2000s financial crisis on suicides in Spain: an interrupted time-series analysis. The European Journal of Public Health 2014;24(2):183-84. doi: 10.1093/eurpub/ckt215

126. López-Ruiz M, Martínez JM, Pérez K, et al. Impact of road safety interventions on traffic-related occupational injuries in Spain, 2004–2010. Accident Analysis & Prevention 2014;66:114-19. doi: 10.1016/j.aap.2014.01.012

127. Lu CY, Zhang F, Lakoma MD, et al. Asthma Treatments and Mental Health Visits After a Food and Drug Administration Label Change for Leukotriene Inhibitors. Clinical Therapeutics 2015;37(6):1280-91. doi: 10.1016/j.clinthera.2015.03.027

128. Ma T, Byrne PA, Haya M, et al. Working in tandem: The contribution of remedial programs and roadside licence suspensions to drinking and driving deterrence in Ontario. Accident Analysis & Prevention 2015;85:248-56. doi: 10.1016/j.aap.2015.09.017

129. Maini R, Van den Bergh R, van Griensven J, et al. Picking up the bill - improving health-care utilisation in the Democratic Republic of Congo through user fee subsidisation: a before and after study. BMC Health Services Research 2014;14(1) doi: 10.1186/s12913-014-0504-6

130. Martin CL, Aldridge PJ, Harris AM, et al. Opening a New Level II Trauma Center Near an Established Level I Trauma Center. Journal of Orthopaedic Trauma 2016;30(10):517-23. doi: 10.1097/bot.0000000000000640

131. Marufu O, Desai N, Aldred D, et al. Analysis of interventions to reduce the incidence of Clostridium difficile infection at a London teaching hospital trust, 2003–2011. Journal of Hospital Infection 2015;89(1):38-45. doi: 10.1016/j.jhin.2014.10.003

132. Marwick CA, Guthrie B, Pringle JEC, et al. A multifaceted intervention to improve sepsis management in general hospital wards with evaluation using segmented regression of interrupted time series. BMJ Quality & Safety 2013;23(12):e2-e2. doi: 10.1136/bmjqs-2013-002176

133. McAlister FA, Bakal JA, Kaul P, et al. Changes in Heart Failure Outcomes After a Province-Wide Change in Health Service Provision A Natural Experiment in Alberta, Canada. Circulation: Heart Failure 2013;6(1):76-82. doi: 10.1161/circheartfailure.112.971119

134. McFarlane WR, Susser E, McCleary R, et al. Reduction in Incidence of Hospitalizations for Psychotic Episodes Through Early Identification and Intervention. Psychiatric Services 2014;65(10):1194-200. doi: 10.1176/appi.ps.201300336

135. McKirdy A, Imbuldeniya AM. The clinical and cost effectiveness of a virtual fracture clinic service. Bone & Joint Research 2017;6(5):259-69. doi: 10.1302/2046-3758.65.bjr-2017-0330.r1

136. McLeod A, Weir A, Aitken C, et al. Rise in testing and diagnosis associated with Scotland's Action Plan on Hepatitis C and introduction of dried blood spot testing. Journal of Epidemiology and Community Health 2014;68(12):1182-88. doi: 10.1136/jech-2014-204451

137. McLintock K, Russell AM, Alderson SL, et al. The effects of financial incentives for case finding for depression in patients with diabetes and coronary heart disease: interrupted time series analysis. BMJ Open 2014;4(8):e005178-e78. doi: 10.1136/bmjopen-2014-005178

138. McPhedran S, Mauser G. Lethal Firearm-Related Violence Against Canadian Women: Did Tightening Gun Laws Have an Impact on Women’s Health and Safety? Violence and Victims 2013;28(5):875-83. doi: 10.1891/0886-6708.vv-d-12-00145

139. Mead EL, Cruz-Cano R, Bernat D, et al. Association between Florida's smoke-free policy and acute myocardial infarction by race: A time series analysis, 2000–2013. Preventive Medicine 2016;92:169-75. doi: 10.1016/j.ypmed.2016.05.032

140. Meirambayeva A, Vingilis E, Zou G, et al. Evaluation of Deterrent Impact of Ontario's Street Racing and Stunt Driving Law on Extreme Speeding Convictions. Traffic Injury Prevention 2014;15(8):786-93. doi: 10.1080/15389588.2014.890721

141. Mellon L, Hickey A, Doyle F, et al. Can a media campaign change health service use in a population with stroke symptoms? Examination of the first Irish stroke awareness campaign. Emergency Medicine Journal 2013;31(7):536-40. doi: 10.1136/emermed-2012-202280

142. Melvin KE, Hart JC, Sorvig RD. Second-Generation Antipsychotic Prescribing Patterns for Pediatric Patients Enrolled in West Virginia Medicaid. Psychiatric Services 2017;68(10):1061-67. doi: 10.1176/appi.ps.201600489

143. Milder EA, Rizzi MD, Morales KH, et al. Impact of a New Practice Guideline on Antibiotic Use With Pediatric Tonsillectomy. JAMA Otolaryngology–Head & Neck Surgery 2015;141(5):410. doi: 10.1001/jamaoto.2015.95

144. Miller P, Curtis A, Palmer D, et al. Changes in injury-related hospital emergency department presentations associated with the imposition of regulatory versus voluntary licensing conditions on licensed venues in two cities. Drug and Alcohol Review 2014;33(3):314-22. doi: 10.1111/dar.12118

145. Miwa S, Visintainer P, Engelman R, et al. Effects of an Ambulation Orderly Program Among Cardiac Surgery Patients. The American Journal of Medicine 2017;130(11):1306-12. doi: 10.1016/j.amjmed.2017.04.044

146. Morse G. The Politics of Financing Evidence-Based Mental Health Treatments: Lessons From Assertive Community TreatmentThe Politics of Financing Evidence-Based Mental Health Treatments: Lessons From Assertive Community Treatment. PsycCRITIQUES 2011;5656(2222) doi: 10.1037/a0023502

147. Moyo P, Simoni-Wastila L, Griffin BA, et al. Impact of prescription drug monitoring programs (PDMPs) on opioid utilization among Medicare beneficiaries in 10 US States. Addiction 2017;112(10):1784-96. doi: 10.1111/add.13860

148. Muoto I, Darney BG, Lau B, et al. Shifting Patterns in Cesarean Delivery Scheduling and Timing in Oregon before and after a Statewide Hard Stop Policy. Health Services Research 2017;53:2839-57. doi: 10.1111/1475-6773.12797

149. Myung W, Lee G-H, Won H-H, et al. Paraquat Prohibition and Change in the Suicide Rate and Methods in South Korea. PLOS ONE 2015;10(6):e0128980. doi: 10.1371/journal.pone.0128980

150. Nakahara S, Ichikawa M, Nakajima Y. Effects of Increasing Child Restraint Use in Reducing Occupant Injuries Among Children Aged 0–5 Years in Japan. Traffic Injury Prevention 2014;16(1):55-61. doi: 10.1080/15389588.2014.897698

151. Narayan H, Thomas SHL, Eddleston M, et al. Disproportionate effect on child admissions of the change in Medicines and Healthcare Products Regulatory Agency guidance for management of paracetamol poisoning: an analysis of hospital admissions for paracetamol overdose in England and Scotland. British Journal of Clinical Pharmacology 2015;80(6):1458-63. doi: 10.1111/bcp.12779

152. Nazif-Munoz JI, Quesnel-Vallée A, van den Berg A. Did Chile's traffic law reform push police enforcement? Understanding Chile's traffic fatalities and injuries reduction. Injury Prevention 2014;21(3):159-65. doi: 10.1136/injuryprev-2014-041358

153. Newitt S, Myles PR, Birkin JA, et al. Impact of infection control interventions on rates of Staphylococcus aureus bacteraemia in National Health Service acute hospitals, East Midlands, UK, using interrupted time-series analysis. Journal of Hospital Infection 2015;90(1):28-37. doi: 10.1016/j.jhin.2014.12.016

154. Nistal-Nuño B. Segmented regression analysis of interrupted time series data to assess outcomes of a South American road traffic alcohol policy change. Public Health 2017;150:51-59. doi: 10.1016/j.puhe.2017.04.025

155. Norstrom T, Stickley A. Alcohol tax, consumption and mortality in tsarist Russia: is a public health perspective applicable? The European Journal of Public Health 2012;23(2):340-44. doi: 10.1093/eurpub/cks079

156. O'Brien NP, Foss RD, Goodwin AH, et al. Supervised hours requirements in graduated driver licensing: Effectiveness and parental awareness. Accident Analysis & Prevention 2013;50:330-35. doi: 10.1016/j.aap.2012.05.007

157. Okasha O, Rinta-Kokko H, Palmu AA, et al. Population-level impact of infant 10-valent pneumococcal conjugate vaccination on adult pneumonia hospitalisations in Finland. Thorax 2017;73(3):262-69. doi: 10.1136/thoraxjnl-2017-210440

158. Osman M, Parnell AC. Effect of the First World War on suicide rates in Ireland: an investigation of the 1864–1921 suicide trends. BJPsych Open 2015;1(2):164-65. doi: 10.1192/bjpo.bp.115.000539

159. Ostrowsky B, Sharma S, DeFino M, et al. Antimicrobial Stewardship and Automated Pharmacy Technology Improve Antibiotic Appropriateness for Community-Acquired Pneumonia. Infection Control & Hospital Epidemiology 2013;34(6):566-72. doi: 10.1086/670623

160. Owens CL, Peterson D, Kamineni A, et al. Effects of transitioning from conventional methods to liquid-based methods on unsatisfactory Papanicolaou tests. Cancer Cytopathology 2013;121(10):568-75. doi: 10.1002/cncy.21309

161. Pace LE, Dusetzina SB, Keating NL. Early Impact of the Affordable Care Act on Uptake of Long-acting Reversible Contraceptive Methods. Medical Care 2016;54(9):811-17. doi: 10.1097/mlr.0000000000000551

162. Pan SW, Chong HH, Kao H-C. Unintentional injury mortality among indigenous communities of Taiwan: trends from 2002 to 2013 and evaluation of a community-based intervention. Injury Prevention 2017;25(1):26-30. doi: 10.1136/injuryprev-2017-042321

163. Panagiotoglou D, Law MR, McGrail K. Effect of Hospital Closures on Acute Care Outcomes in British Columbia, Canada. Medical Care 2017;55(1):50-56. doi: 10.1097/mlr.0000000000000619

164. Panatto D, Domnich A, Gasparini R, et al. An eHealth Project on Invasive Pneumococcal Disease: Comprehensive Evaluation of a Promotional Campaign. Journal of Medical Internet Research 2016;18(12):e316. doi: 10.2196/jmir.6205

165. Parikh K, Hall M, Teach SJ. Bronchiolitis Management Before and After the AAP Guidelines. Pediatrics 2013;133(1):e1-e7. doi: 10.1542/peds.2013-2005

166. Patel PR, Yi SH, Booth S, et al. Bloodstream Infection Rates in Outpatient Hemodialysis Facilities Participating in a Collaborative Prevention Effort: A Quality Improvement Report. American Journal of Kidney Diseases 2013;62(2):322-30. doi: 10.1053/j.ajkd.2013.03.011

167. Pegues DA, Han J, Gilmar C, et al. Impact of Ultraviolet Germicidal Irradiation for No-Touch Terminal Room Disinfection on Clostridium difficile Infection Incidence Among Hematology-Oncology Patients. Infection Control & Hospital Epidemiology 2016;38(1):39-44. doi: 10.1017/ice.2016.222

168. Pellegrin KL, Krenk L, Oakes SJ, et al. Reductions in Medication-Related Hospitalizations in Older Adults with Medication Management by Hospital and Community Pharmacists: A Quasi-Experimental Study. Journal of the American Geriatrics Society 2016;65(1):212-19. doi: 10.1111/jgs.14518

169. Petereit D, Omidpanah A, Boylan A, et al. A Multi-faceted Approach to Improving Breast Cancer Outcomes in a Rural Population, and the Potential Impact of Patient Navigation. South Dakota medicine : the journal of the South Dakota State Medical Association 2016;69(6):268-73. [published Online First: 2016/07/23]

170. Petrou P. The Ariadne's thread in co-payment, primary health care usage and financial crisis: findings from Cyprus public health care sector. Public Health 2015;129(11):1503-09. doi: 10.1016/j.puhe.2015.07.032

171. Pinheiro SP, Kang EM, Kim CY, et al. Concomitant use of isotretinoin and contraceptives before and after iPledge in the United States. Pharmacoepidemiology and Drug Safety 2013;22(12):1251-57. doi: 10.1002/pds.3481

172. Poluzzi E, Veronese G, Piccinni C, et al. Switching among Equivalents in Chronic Cardiovascular Therapies: ‘Real World’ Data from Italy. Basic & Clinical Pharmacology & Toxicology 2015;118(1):63-69. doi: 10.1111/bcpt.12442

173. Pradhan A, Anasuya A, Pradhan MM, et al. Trends in Malaria in Odisha, India—An Analysis of the 2003–2013 Time-Series Data from the National Vector Borne Disease Control Program. PLOS ONE 2016;11(2):e0149126. doi: 10.1371/journal.pone.0149126

174. Pridemore WA, Chamlin MB, Andreev E. Reduction in Male Suicide Mortality Following the 2006 Russian Alcohol Policy: An Interrupted Time Series Analysis. American Journal of Public Health 2013;103(11):2021-26. doi: 10.2105/ajph.2013.301405

175. Prinja S, Kaur G, Gupta R, et al. Out-of-pocket expenditure for health care: District level estimates for Haryana state in India. The International Journal of Health Planning and Management 2018;34(1):277-93. doi: 10.1002/hpm.2628

176. Puig-Junoy J, Rodríguez-Feijoó S, Lopez-Valcarcel BG. Paying for Formerly Free Medicines in Spain After 1 Year of Co-Payment: Changes in the Number of Dispensed Prescriptions. Applied Health Economics and Health Policy 2014;12(3):279-87. doi: 10.1007/s40258-014-0097-6

177. Pun VC, Lin H, Kim JH, et al. Impacts of alcohol duty reductions on cardiovascular mortality among elderly Chinese: a 10-year time series analysis. Journal of Epidemiology and Community Health 2013;67(6):514-18. doi: 10.1136/jech-2012-201859

178. Rhodes D, Cheng AC, McLellan S, et al. Reducing Staphylococcus aureus bloodstream infections associated with peripheral intravenous cannulae: successful implementation of a care bundle at a large Australian health service. Journal of Hospital Infection 2016;94(1):86-91. doi: 10.1016/j.jhin.2016.05.020

179. Rooholamini SN, Clifton H, Haaland W, et al. Outcomes of a Clinical Pathway to Standardize Use of Maintenance Intravenous Fluids. Hospital Pediatrics 2017;7(12):703-09. doi: 10.1542/hpeds.2017-0099

180. Rosenthal MB, Friedberg MW, Singer SJ, et al. Effect of a Multipayer Patient-Centered Medical Home on Health Care Utilization and Quality. JAMA Internal Medicine 2013;173(20):1907. doi: 10.1001/jamainternmed.2013.10063

181. Rutman L, Wright DR, O'Callaghan J, et al. A Comprehensive Approach to Pediatric Pneumonia. Journal for Healthcare Quality 2017;39(4):e59-e69. doi: 10.1097/jhq.0000000000000048

182. Ryu S, Lau CL, Chun BC. The impact of Livestock Manure Control Policy on human leptospirosis in Republic of Korea using interrupted time series analysis. Epidemiology and Infection 2017;145(7):1320-25. doi: 10.1017/s0950268817000218

183. Sakai R, Wang W, Yamaguchi N, et al. The Impact of Japan's 2004 Postgraduate Training Program on Intra-Prefectural Distribution of Pediatricians in Japan. PLoS ONE 2013;8(10):e77045. doi: 10.1371/journal.pone.0077045

184. Santa-Ana-Tellez Y, Mantel-Teeuwisse AK, Dreser A, et al. Impact of Over-the-Counter Restrictions on Antibiotic Consumption in Brazil and Mexico. PLoS ONE 2013;8(10):e75550. doi: 10.1371/journal.pone.0075550

185. Santa-Ana-Tellez Y, Mantel-Teeuwisse AK, Leufkens HGM, et al. Seasonal Variation in Penicillin Use in Mexico and Brazil: Analysis of the Impact of Over-the-Counter Restrictions. Antimicrobial Agents and Chemotherapy 2014;59(1):105-10. doi: 10.1128/aac.03629-14

186. Scherb HH, Mori K, Hayashi K. Increases in perinatal mortality in prefectures contaminated by the Fukushima nuclear power plant accident in Japan. Medicine 2016;95(38):e4958. doi: 10.1097/md.0000000000004958

187. Sicsic J, Saint-Lary O, Rouveix E, et al. Impact of a primary care national policy on HIV screening in France: a longitudinal analysis between 2006 and 2013. British Journal of General Practice 2016;66(653):e920-e29. doi: 10.3399/bjgp16x687529

188. Singh K, Speizer I, Handa S, et al. Impact evaluation of a quality improvement intervention on maternal and child health outcomes in Northern Ghana: early assessment of a national scale-up project. International Journal for Quality in Health Care 2013;25(5):477-87. doi: 10.1093/intqhc/mzt054

189. Sinnott S-J, Normand C, Byrne S, et al. Copayments for prescription medicines on a public health insurance scheme in Ireland. Pharmacoepidemiology and Drug Safety 2015;25(6):695-704. doi: 10.1002/pds.3917

190. Slattery C, Freund M, Gillham K, et al. Increasing smoking cessation care across a network of hospitals: an implementation study. Implementation Science 2015;11(1) doi: 10.1186/s13012-016-0390-x

191. Smith RL, Hayashi VN, Lee YI, et al. The Medical Emergency Team Call. Critical Care Medicine 2014;42(2):322-27. doi: 10.1097/ccm.0b013e3182a27413

192. Staras SAS, Livingston MD, Christou AM, et al. Heterogeneous population effects of an alcohol excise tax increase on sexually transmitted infections morbidity. Addiction 2014;109(6):904-12. doi: 10.1111/add.12493

193. Stelfox HT, Bastos J, Niven DJ, et al. Critical care transition programs and the risk of readmission or death after discharge from ICU. Intensive Care Medicine 2016;42(3):401-10. doi: 10.1007/s00134-015-4173-7

194. Taber DJ, DuBay D, McGillicuddy JW, et al. Impact of the New Kidney Allocation System on Perioperative Outcomes and Costs in Kidney Transplantation. Journal of the American College of Surgeons 2017;224(4):585-92. doi: 10.1016/j.jamcollsurg.2016.12.009

195. Thijssen WAMH, Wijnen-van Houts M, Koetsenruijter J, et al. The Impact on Emergency Department Utilization and Patient Flows after Integrating with a General Practitioner Cooperative: An Observational Study. Emergency Medicine International 2013;2013:1-8. doi: 10.1155/2013/364659

196. Tiwari A, Osbert N, Matimelo SM, et al. Assessing the Impact of Leveraging Traditional Leadership on Access to Sanitation in Rural Zambia. The American Journal of Tropical Medicine and Hygiene 2017;97(5):1355-61. doi: 10.4269/ajtmh.16-0612

197. Troelstra S, Bosdriesz J, de Boer M, et al. Effect of tobacco control policies on information seeking for smoking cessation in the Netherlands: A Google Trends study. European Journal of Public Health 2014;24(suppl_2) doi: 10.1093/eurpub/cku164.043

198. Ullman M, Parlier G, Warren J, et al. The Economic Impact of Starting, Stopping, and Restarting an Antibiotic Stewardship Program: A 14-Year Experience. Antibiotics 2013;2(2):256-64. doi: 10.3390/antibiotics2020256

199. Yarnell CJ, Shadowitz S, Redelmeier DA. Hospital Readmissions Following Physician Call System Change: A Comparison of Concentrated and Distributed Schedules. The American Journal of Medicine 2016;129(7):706-14.e2. doi: 10.1016/j.amjmed.2016.02.022

200. Zhao A, Chen R, Qi Y, et al. Evaluating the Impact of Criminalizing Drunk Driving on Road-Traffic Injuries in Guangzhou, China: A Time-Series Study. Journal of Epidemiology 2016;26(8):433-39. doi: 10.2188/jea.je20140103

Appendix E. Additional file accompanying Chapter 4 – Computer code to demonstrate and create effective ITS graphs

“Creating effective interrupted time series graphs: review and recommendations”

Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Korevaar E, Cheng AC, Bero L, McKenzie JE.

The following Stata 15 computer code reproduces the graphs from the manuscript.

```
****************************************************************************************************
**********************************
// The following do file provides the code to generate the graphs in the paper:
// Turner, et al. "Creating effective interrupted time series graphs: review and recommendations".
Research Synthesis Methods. 2020.

// The do file is separated into the following sections:
// Part 1: code is provided to generate data used throughout the paper (lines 32 - 109);
// Part 2: code is provided to format the data in preparation for graphing (lines 111 - 224);
// Part 3: code is provided to create the graphs presented in the paper (lines 225 onwards).
// Please note that if you have your own dataset, you can start at lines 116 of this do file.

// Date created: 15 April 2020

// Note that this code will run using Stata version 15 or later.
// For those using Stata version 14 or earlier, some features, e.g. transparency, will not work.
version 15

// The circular package (by NJ Cox, University of Durham, UK) needs to be installed
// Note, if you have previously installed the circular package, please comment the next line of
code
// ssc install circular

// Begin by clearing any existing graphs from Stata's memory
graph drop _all

// Clear any data from Stata's memory
clear

// Save the name of the directory where the graph will be saved in a local macro
// Note that the local macro will be used throughout the do file
local graph_directory "<set directory>”

// Set the seed
set seed 000013

****************************************************************************************************
**********************************
// Part 1 - Create an interrupted time series dataset
// parameters are:
// beta_0 = intercept
// beta_1 = slope in first segment
// beta_2 = level change at first interruption
// beta_3 = slope change after first interruption
// sigma = normally distributed standard deviation
// rho = autocorrelation coefficient
// seasonality = maximum amplitude of sinusoidal seasonal effects
// interruption_time = time of interruption
// num_points = how many data points (time point)
// segment = the variable we will use to define each interruption time period
//          0 for pre-interruption, 1 for post-interruption
//          this could easily be extended to further interruption periods

// Assign values to local macros representing the parameters that were described above:
// Note, that the values below are those used in the paper
local beta_0 = 100
local beta_1 = 0
local beta_2 = -80
local beta_3 = 0.7
local sigma = 4
local rho = 0
local seasonality = 20
local interruption_time = 51
local num_points 100
set obs `num_points'

// Generate a new variable called time that contains the number of time points in the dataset
gen time = _n

// Create an indicator variable to represent the segments of the time series
// Note that 0 = pre-interruption period/segment and 1 = post-interruption period/segment
gen segment = 0
replace segment = 1 if time >= `interruption_time'

///////////////////////////////////////////////////////////////////////////////////////////////////
/
// Generate data according to a first order autoregressive model
// First, generate a random error based on a normal distribution
gen error_normal = rnormal(0,`sigma'^2)
gen error = 0

// Incorporate the autocorrelated component to the normal error to obtain an 'autocorrelated error'
component
// Note that the variance remains constant over time
replace error = sqrt(1/(1-(`rho'^2)))*error_normal in 1

// Add autocorrelated error to normally distributed error
replace error = `rho'*error[_n-1] + error_normal in 2/`num_points'

// Generate a new variable that represents seasonality
// Note that seasonality is incorporated as a sinusoidal curve
gen seasonality = `seasonality'*sin((c(pi)/6)*time)

// Generate the outcome variable as the sum of all the parameters specified above
gen outcome = `beta_0' + `beta_1'*time + `beta_2'*segment + `beta_3'*(time -
(`intervention_time'+1))*segment + error + seasonality

// Next, we generate a second set of variables similarly to above but this outcome can be thought
to represent a control series
gen error_normal_2 = rnormal(0,`sigma'^2)
gen error_2 = 0

// Incorporate the autocorrelated component to the normal error to obtain an 'autocorrelated error'
component
// Note that the variance remains constant over time
replace error_2 = sqrt(1/(1-(`rho'^2)))*error_normal_2 in 1

// Add autocorrelated error to normally distributed error
replace error_2 = `rho'*error_2[_n-1] + error_normal_2 in 2/`num_points'

// Note that for this outcome/control series we include the same seasonality component as above

// Generate the second outcome by starting at a different value for the intercept (beta_0) and do
not add a level change or slope change
gen outcome_2 = 170 + `beta_1'*time + error_2 + seasonality

// Keep the variables that we will use for the remaining do file
keep time segment outcome outcome_2

// save "`graph_directory'\ITS_data.dta", replace


***************************************************************************************************
**********************************
// Part 2:  Format the data in preparation for graphing
// use "`graph_directory'\ITS_data.dta", clear

// Note that if you are starting here and have your own dataset, then you will need the following
variables:
// 1. time - contains the number of time points in the dataset
// 2. segment - contains the segments in the dataset; 0 = pre-interruption and 1 = post-
interruption
// 3. outcome - contains the outcome dataset
// 4. outcome_2 - contains the outcome data for a second series


// To format the data, we need to:
//   i. Find the interruption time and the corresponding segments
//  ii. Generate additional ITS variables
// iii. Set up the model
//  iv. Fit the analysis model
// i. Find the interruption time and the corresponding segments
summ segment
local min_seg_num = r(min)
local num_segments = r(max)

// If the segments are numbered 1, 2, 3, ..., change the segments to 0, 1, 2, ...
if `min_seg_num' != 0 {
   replace segment = segment - `min_seg_num'
}

summ segment
local num_segments = r(max)
local min_seg_num = r(min)

// Find the timing of each segment
forvalues segment = 0/`num_segments' {
   qui: summ time if segment == `segment'
   local time_`segment'_start = r(min)
   local time_`segment'_end = r(max)
   local segment_`segment'_length = r(N)
   display "segment `segment' goes from `time_`segment'_start' to `time_`segment'_end' and is
`segment_`segment'_length' points long"
}

// ii. Generate additional ITS variables, which are needed to fit the segmented regression model as
in Huitema and McKean (2007)
// Huitema and McKean (2007). "Identifying Autocorrelation Generated by Various Error Processes in
Interrupted Time-Series Regression Designs." Educational and Psychological Measurement 67(3): 447-
459.

forvalues segment = 0/`num_segments' {
   gen interruption_`segment' = 0
   replace interruption_`segment' = 1 if segment >= `segment'
   gen level_change_`segment' = interruption_`segment'
   gen slope_change_`segment' = (time-`time_`segment'_start')*level_change_`segment'
}
//
// The data need to be declared as time series data
tsset time

// iii. Set up the model so that each segment has a level and slope change
// Note that here we include the first segment but when we fit the statistical analysis we let it
know we've already included an intercept

local variables = ""

forvalues segment = 0/`num_segments' {
      local variables = "`variables'" + " level_change_`segment' slope_change_`segment'"
}

display "variables: `variables'"

// iv. Fit the analysis model
// Please note the following:
// a. We have used the mixed command with the restricted maximum likelihood method for the variance
to account for autocorrelation.
// b. Other methods (e.g. Prais) could also be used.
// c. We specify the nocons option because we have already specified an intercept (i.e.
level_change_0)
// d. The mixed model that we fit includes an autocorrelation with lag-1 (as specified by res(ar 1,
t(time)))
// e. The number of iterations are set to 1000

mixed outcome `variables' , nocons res(ar 1, t(time)) var reml iter(1000)

// Save the t-statistic for estimating the confidence limits in a scalar
matrix table = r(table)
scalar tcrit = table[8,1]

// Generate variables corresponding to the estimated values and their standard errors from the
mixed model
predict estimates
predict SE_prediction , stdp

// Generate variables with the confidence limits of the estimates
gen cl_lower = estimates - tcrit*SE_prediction
gen cl_upper = estimates + tcrit*SE_prediction
```

```
// Save to local macros the estimates (95% confidence intervals) for the level and slope change
with appropriate formatting for display
local level_change : di %3.1f `=table[1,3]'
local slope_change : di %3.2f `=table[1,4]'
local level_change_ll : di %3.1f `=table[5,3]'
local slope_change_ll : di %3.2f `=table[5,4]'
local level_change_ul : di %3.1f `=table[6,3]'
local slope_change_ul : di %3.2f `=table[6,4]'

// Generate a counterfactual from the first period
gen first_period = .
replace first_period = outcome if segment == 0
mixed first_period `variables' , nocons res(ar 1, t(time)) var reml iter(1000)

// Save the t-statistic for estimating the confidence limits in a scalar
matrix table = r(table)
scalar tcrit = table[8,1]

// Generate variables corresponding to the of the estimated values and their standard errors from
the mixed model for the counterfactual
predict counterfactual
predict counterfactual_SE, stdp

// Generate variables with the confidence limits of the estimates for the counterfactual
gen counter_cl_lower = counterfactual - tcrit*counterfactual_SE
gen counter_cl_upper = counterfactual + tcrit*counterfactual_SE

// Fit a mixed model to the second series
mixed outcome_2 `variables' , nocons res(ar 1, t(time)) var reml iter(1000)

// Generate a variable of the estimated values for the second series from the mixed model
predict estimates_2

// Generate a counterfactual based on the estimates of the parameters
gen counterfactual_2 = _b[level_change_0] + _b[slope_change_0]*slope_change_0


****************************************************************************************************
**********************************
//Part 3: code is provided to create the graphs presented in the paper

/////////////////////////////////////////////////////////////////////////////////
// Graph in Figure 1a
// Simple graph depicting the data points, trend lines, counterfactual and interruption

// find some points of interest
egen graph_min = rowmin(outcome counterfactual)
egen graph_max = rowmax(outcome counterfactual)
summ graph_max
local graph_max = r(max)

// Save the values for the colour and other graph parameters to local macros:

// Save the options for the colour for the data points and trend lines to local macro
local outcome_colour = "0 114 178" // colour blind friendly blue

local point_size = "small"
local point_symbol = "plus"
local point_transparency = "%70" // this can be adjusted depending on number of points

local trend_pattern = "solid"
local trend_width = "medthick"

// Save the options for the counterfactual lines to local macros
local counterfactual_pattern = "dash"
local counterfactual_width = "medium"

// Options for interruption lines
local interruption_colour = "red"
local interruption_pattern = "dash"
local interruption_width = "medium"

// Save the options for the titles and labels to local macro
local x_title "months since start of project"
local y_title "outcome details"
local interruption_label "interruption"
local interruption_label_height = `graph_max' + 5

// Generate the text for the graphing of the trend lines
```

```
local num_line = 0
local trend_lines = ""

forvalues segment = 0/`num_segments' {
   local trend_lines = `"`trend_lines'"' + `" || lfit estimates time if segment == `segment',
lcol("`outcome_colour'") lpattern(`trend_pattern') lwidth(`trend_width')"'
   local num_lines = `num_lines' + 1
}

// Generate the graph using Stata's twoway command
local name = "simple_graph_1a"

graph twoway scatter outcome time, msize(`point_size') msym(`point_symbol')
mcol("`outcome_colour'`point_transparency'") /// outcomes
          || lfit counterfactual time if segment == 1, lcol("`outcome_colour'")
lpattern(`counterfactual_pattern') lwidth(`counterfactual_width') /// counterfactual
          `trend_lines' /// trend lines
          ylab(,angle(h)) /// the angle ensures numbers are printed horizontally
          title("1a", position(10) ring(1)) ///
          ytitle(`y_title') ///
          xtitle(`x_title') ///
          xline(`time_1_start', lcol(`interruption_colour') lpattern(`interruption_pattern')
lwidth(`interruption_width')) /// interruption line
          text(`interruption_label_height' `time_0_end' "`interruption_label'", placement(w)) ///
interruption text label
          legend(order(1 "outcome" 3 "trends" 2 "counterfactual") row(1) region(style(none)) ) ///
legend without border
          graphregion(col(white)) /// set background to white
          name(`name')

graph export "`graph_directory'\`name'.svg", replace
graph export "`graph_directory'\`name'.png", replace

// Drop variables that are no longer needed
drop graph_max
drop graph_min

////////////////////////////////////////////////////////////////////////////////
// Code to generate graph in Figure 1b
// Similar to graph in Figure 1a except a shaded area is used to depict the interruption

// Find mininmum and maximum points of the graph
egen graph_max = rowmax(outcome counterfactual)
egen graph_min = rowmin(outcome counterfactual)
summ graph_max
local graph_max = r(max)
replace graph_max = `graph_max'
summ graph_min
local graph_min = r(min)


// Save the values for the colour and other graph parameters to local macros:
local outcome_colour = "0 114 178" // colour blind friendly blue

local point_size = "small"
local point_symbol = "plus"
local point_transparency = "%70" // this can be adjusted depending on number of points

local trend_pattern = "solid"
local trend_width = "medthick"

// Save the options for the counterfactual lines to local macros
local counterfactual_pattern = "dash"
local counterfactual_width = "medium"

// Save the options for the shading colour to represent the interruption to local macro
local interruption_colour = "gray%15"

// Save the options for the titles and labels to local macro
local x_title "months since start of project"
local y_title "outcome details"
local interruption_label "interruption"
local interruption_label_height = `graph_max'
local interruption_label_x_position = `time_1_start'

// Generate the text for the graphing of the trend lines
local num_line = 0
local trend_lines = ""
```

```
forvalues segment = 0/`num_segments' {
   local trend_lines = `"`trend_lines'"' + `" || lfit estimates time if segment == `segment',
lcol("`outcome_colour'") lpattern(`trend_pattern') lwidth(`trend_width')"'
   local num_lines = `num_lines' + 1
}

// Generate the graph using Stata's twoway command
local name = "simple_graph_shaded_1b"

graph twoway area graph_max time if inrange(time, `time_1_start', `time_1_end'),
color(`interruption_colour') lwidth(none) base(`graph_min') /// shaded area
          || scatter outcome time, msize(`point_size') msym(`point_symbol')
mcol("`outcome_colour'`point_transparency'") /// outcomes
          || lfit counterfactual time if segment == 1, lcol("`outcome_colour'")
lpattern(`counterfactual_pattern') lwidth(`counterfactual_width') /// counterfactual
          `trend_lines' /// trend lines
          ylab(,angle(h)) /// the angle ensures numbers are printed horizontally
          title("1b", position(10) ring(1)) ///
          ytitle(`y_title') ///
          xtitle(`x_title') ///
          text(`graph_max' `interruption_label_x_position' "`interruption_label'", placement(e))
/// interruption text label
          legend(off) /// no legend - detail will need to be added to the caption
          graphregion(col(white)) /// set background to white
          name(`name')

graph export "`graph_directory'\`name'.svg", replace
graph export "`graph_directory'\`name'.png", replace

drop graph_max
drop graph_min

graph combine simple_graph_1a simple_graph_shaded_1b, col(1) ysize(8) iscale(0.7273)
graphregion(margin(zero) col(white)) name(combined_graph_1)
graph export "`graph_directory'\combined_graph_1.svg", replace
graph export "`graph_directory'\combined_graph_1.png", replace

////////////////////////////////////////////////////////////////////////////////
// Code to generate graph in Figure 2a
// To obtain the graph in Figure 2a, we add a second series to the graph in Figures 1a and b

// Find minimum and maximum points of the graph
egen graph_max = rowmax(outcome counterfactual outcome_2 counterfactual_2)
egen graph_min = rowmin(outcome counterfactual outcome_2 counterfactual_2)
summ graph_max
local graph_max = r(max)

// Save the values for the colour and other graph parameters to local macros:
local outcome_colour = "0 114 178" // colour blind friendly blue
local outcome_2_colour = "230 159 0" // colour blind friendly orange

local point_size = "small"
local point_symbol = "plus"
local point_transparency = "%70" // this can be adjusted depending on number of points
local point_2_symbol = "X"

local trend_pattern = "solid"
local trend_width = "medthick"

// Save the options for the counterfactual lines to local macro
local counterfactual_pattern = "dash"
local counterfactual_width = "medium"

// Save the options for the interruption lines to local macro
local interruption_colour = "red"
local interruption_pattern = "dash"
local interruption_width = "medium"

// Save the options for the titles and labels to local macro
local x_title "months since start of project"
local y_title "outcome details"
local interruption_label "interruption"
local interruption_label_height = `graph_max'

// Generate the text for the graphing of the trend lines
local num_line = 0
local trend_lines = ""

forvalues segment = 0/`num_segments' {
   local trend_lines = `"`trend_lines'"' + `" || lfit estimates time if segment == `segment',
lcol("`outcome_colour'") lpattern(`trend_pattern') lwidth(`trend_width')"'
      // here we add the second set of trend lines
   local trend_lines_2 = `"`trend_lines_2'"' + `" || lfit estimates_2 time if segment == `segment',
lcol("`outcome_2_colour'") lpattern(`trend_pattern') lwidth(`trend_width')"'
   local num_lines = `num_lines' + 1

}

// Generate the graph using Stata's twoway command
local name = "double_graph_2a"

graph twoway scatter outcome time, msize(`point_size') msym(`point_symbol')
mcol("`outcome_colour'`point_transparency'") /// outcomes
          || scatter outcome_2 time, msize(`point_size') msym(`point_2_symbol')
mcol("`outcome_2_colour'`point_transparency'") /// outcomes
          || lfit counterfactual time if segment == 1, lcol("`outcome_colour'")
lpattern(`counterfactual_pattern') lwidth(`counterfactual_width') /// counterfactual
          || lfit counterfactual_2 time if segment == 1, lcol("`outcome_2_colour'")
lpattern(`counterfactual_pattern') lwidth(`counterfactual_width') /// counterfactual
          `trend_lines' /// trend lines
          `trend_lines_2' /// second series trend lines
          ylab(,angle(h)) /// the angle ensures numbers are printed horizontally
          title("2a", position(10) ring(1)) ///
          ytitle(`y_title') ///
          xtitle(`x_title') ///
          xline(`time_1_start', lcol(`interruption_colour') lpattern(`interruption_pattern')
lwidth(`interruption_width')) /// interruption line
          text(`interruption_label_height' `time_0_end' "`interruption_label'", placement(w)) ///
interruption text label
          legend(order(1 "intervention" 2 "control") row(1) region(style(none)) ) /// legend
without border
          graphregion(col(white)) /// set background to white
          name(`name')

graph export "`graph_directory'\`name'.svg", replace
graph export "`graph_directory'\`name'.png", replace

drop graph_max
drop graph_min

////////////////////////////////////////////////////////////////////////////////
// Code to generate graph in Figure 2b
// To this graph, we add labels on the graph rather than the placing the information in the legend

// Find minimum and maximum points of the graph
egen graph_max = rowmax(outcome counterfactual outcome_2 counterfactual_2)
egen graph_min = rowmin(outcome counterfactual outcome_2 counterfactual_2)
summ graph_max
local graph_max = r(max)


// Save the values for the colour and other graph parameters to local macros:
local outcome_colour = "0 114 178" // colour blind friendly blue
local outcome_2_colour = "230 159 0" // colour blind friendly orange

local point_size = "small"
local point_symbol = "plus"
local point_transparency = "%70" // this can be adjusted depending on number of points
local point_2_symbol = "X"

local trend_pattern = "solid"
local trend_width = "medthick"

// Save the options for the counterfactual lines to local macro
local counterfactual_pattern = "dash"
local counterfactual_width = "medium"

// Save the options for the interruption lines to local macro
local interruption_colour = "red"
local interruption_pattern = "dash"
local interruption_width = "medium"

// Save the options for the titles and labels to local macro
local x_title "months since start of project"
local y_title "outcome details"
local interruption_label "interruption"
local interruption_label_height = `graph_max'
```

```
// Note that the y-coordinates specified below will need to be adjusted to ensure that data points
are not obstructed
local outcome_label "intervention"
local outcome_label_y_coord = 70
local outcome_2_label "control"
local outcome_2_label_y_coord = 200

// Generate the text for the graphing of the trend lines
local num_line = 0
local trend_lines = ""

forvalues segment = 0/`num_segments' {
   local trend_lines = `"`trend_lines'"' + `" || lfit estimates time if segment == `segment',
lcol("`outcome_colour'") lpattern(`trend_pattern') lwidth(`trend_width')"'
      // here we add the second set of trend lines
   local trend_lines_2 = `"`trend_lines_2'"' + `" || lfit estimates_2 time if segment == `segment',
lcol("`outcome_2_colour'") lpattern(`trend_pattern') lwidth(`trend_width')"'
   local num_lines = `num_lines' + 1
}

// Generate the graph using Stata's twoway command
local name = "double_graph_labelled_2b"

graph twoway scatter outcome time, msize(`point_size') msym(`point_symbol')
mcol("`outcome_colour'`point_transparency'") /// outcomes
          || scatter outcome_2 time, msize(`point_size') msym(`point_2_symbol')
mcol("`outcome_2_colour'`point_transparency'") /// outcomes
          || lfit counterfactual time if segment == 1, lcol("`outcome_colour'")
lpattern(`counterfactual_pattern') lwidth(`counterfactual_width') /// counterfactual
          || lfit counterfactual_2 time if segment == 1, lcol("`outcome_2_colour'")
lpattern(`counterfactual_pattern') lwidth(`counterfactual_width') /// counterfactual
          `trend_lines' /// trend lines
          `trend_lines_2' /// second series trend lines
          ylab(,angle(h)) /// the angle ensures numbers are printed horizontally
          title("2b", position(10) ring(1)) ///
          ytitle(`y_title') ///
          xtitle(`x_title') ///
          xline(`time_1_start', lcol(`interruption_colour') lpattern(`interruption_pattern')
lwidth(`interruption_width')) /// interruption line
          text(`interruption_label_height' `time_0_end' "`interruption_label'", placement(w)) ///
interruption text label
          text(`outcome_2_label_y_coord' 0 "`outcome_2_label'", placement(e)
col(`outcome_2_colour')) ///
          text(`outcome_label_y_coord' 0 "`outcome_label'", placement(e) col(`outcome_colour')) ///
          legend(off) /// legend off this time
          graphregion(col(white)) /// set background to white
          name(`name')

graph export "`graph_directory'\`name'.svg", replace
graph export "`graph_directory'\`name'.png", replace

drop graph_max
drop graph_min

graph combine double_graph_2a double_graph_labelled_2b, col(1) ysize(8) iscale(0.7273)
graphregion(margin(zero) col(white)) name(combined_graph_2)
graph export "`graph_directory'\combined_graph_2.svg", replace
graph export "`graph_directory'\combined_graph_2.png", replace

////////////////////////////////////////////////////////////////////////////////
// Code to generate graph in Figure 3a
// To this graph we add a line representing seasonality.
// Note that there are several ways to fit/graph seasonality, we have used the method specified by
Bhaskaran et al. (2013) "Time series regression studies in environmental epidemiology"
International Journal of Epidemiology 42: 1187-1195.

// Find the minimum and maximum points of the graph
egen graph_max = rowmax(outcome counterfactual)
egen graph_min = rowmin(outcome counterfactual)
summ graph_max
local graph_max = r(max)

// Define harmonics as specified in:
// Bhaskaran et al. (2013) "Time series regression studies in environmental epidemiology"
International Journal of Epidemiology 42: 1187-1195.

// (a) Generate sine and cosine functions of time with annual period, plus 2 harmonics
local num_harmonics = 2
```

```
gen degrees=(time/12)*360
fourier degrees, n(`num_harmonics')

// (b) Fit a mixed model
mixed outcome `variables' cos* sin* , nocons res(ar 1, t(time)) var reml iter(1000)
predict full_model

// Set up a list of the seasonal model outputs
local seasonal = ""
forvalues harmonic = 1/`num_harmonics' {
   if `harmonic' == 1 {
      local seasonal = `" `seasonal' "' + `" _b[cos_`harmonic']*cos_`harmonic' +
_b[sin_`harmonic']*sin_`harmonic' "'
   }
   else {
      local seasonal = `" `seasonal' "' + `" + _b[cos_`harmonic']*cos_`harmonic' +
_b[sin_`harmonic']*sin_`harmonic' "'
   }
}

// Use the model estimates to generate the seasonal component of the model
gen seasonal = `seasonal'

// Calculate the deseasonalised data
gen deseasonal = outcome - seasonal

// Fit the mixed model to obtain the model estimates based on de-seasonalised data
mixed deseasonal `variables' , nocons res(ar 1, t(time)) var reml iter(1000)
predict estimates_deseasonal

// Generate counterfactual
gen counterfactual_deseasonal = _b[level_change_0] + _b[slope_change_0]*slope_change_0

// Save the values for the colour and other graph parameters to local macros:
local outcome_colour = "0 114 178" // colour blind friendly blue

local point_size = "small"
local point_symbol = "plus"
local point_transparency = "%70" // this can be adjusted depending on number of points

local trend_pattern = "solid"
local trend_width = "medthick"

// Save the options for the counterfactual lines to local macro
local counterfactual_pattern = "dash"
local counterfactual_width = "medium"

// Save the options for the interruption lines to local macro
local interruption_colour = "red"
local interruption_pattern = "dash"
local interruption_width = "medium"

// Save the options for the titles and labels to local macro
local x_title "months since start of project"
local y_title "outcome details"
local interruption_label "interruption"
local interruption_label_height = `graph_max' + 5

forvalues segment = 0/`num_segments' {
   local fitted_lines_deseason = `"`fitted_lines_deseason'"' + `" || lfit estimates_deseasonal time
if segment == `segment', lcol("`outcome_colour'") lpattern(`trend_pattern') lwidth(`trend_width')"'
   local num_lines = `num_lines' + 1
}

// Generate the graph using Stata's twoway command
local name = "seasonal_graph_3a"
graph twoway scatter outcome time, msize(`point_size') msym(`point_symbol')
mcol("`outcome_colour'`point_transparency'") /// outcomes
          || line full_model time if segment == 0, lc(gray%50) /// sine wave plots
          || line full_model time if segment == 1, lc(gray%50) /// sine wave plots
          || lfit counterfactual_deseasonal time if segment == 1, lcol("`outcome_colour'")
lpattern(`counterfactual_pattern') lwidth(`counterfactual_width') /// counterfactual
          `fitted_lines_deseason' /// deseasonal fitted lines
          ylab(,angle(h)) ///
          title("3a", position(10) ring(1)) ///
          ytitle(`y_title') ///
          xtitle(`x_title') ///
          xline(`time_1_start', lcol(`interruption_colour') lpattern(`interruption_pattern')
lwidth(`interruption_width')) /// interruption line
```

```
          text(`interruption_label_height' `time_0_end' "`interruption_label'", placement(w)) ///
interruption text label
          legend(order(1 "outcome" 2 "estimated seasonality" 4 "counterfactual" 5 "seasonality
adjusted trends")  region(style(none)) ) ///
          graphregion(col(white)) ///
          name(`name')

graph export "`graph_directory'\`name'.svg", replace
graph export "`graph_directory'\`name'.png", replace

drop graph_max
drop graph_min

////////////////////////////////////////////////////////////////////////////////
// Code to generate graph in Figure 3b
// In this graph, we show the precision around the estimates using shading

// Find the minumum and maximum points on the graph
egen graph_max = rowmax(cl_upper counter_cl_upper )
egen graph_min = rowmin(cl_lower counter_cl_lower)
summ graph_max
local graph_max = r(max)

// Save the values for the colour and other graph parameters to local macros:
local outcome_colour = "0 114 178" // colour blind friendly blue

local point_size = "small"
local point_symbol = "plus"
local point_transparency = "%70" // this can be adjusted depending on number of points

local trend_pattern = "solid"
local trend_width = "medthick"

// Save the options for the counterfactual lines to local macro
local counterfactual_pattern = "dash"
local counterfactual_width = "medium"

// Save the options for the interruption lines to local macro
local interruption_colour = "red"
local interruption_pattern = "dash"
local interruption_width = "medium"

// Save the options for the titles and labels to local macro
local x_title "months since start of project"
local y_title "outcome details"
local interruption_label "interruption"
local interruption_label_height = `graph_max'

// Generate the text for the graphing of the trend lines
local num_line = 0
local trend_lines = ""

forvalues segment = 0/`num_segments' {
   local trend_lines = `"`trend_lines'"' + `" || lfit estimates time if segment == `segment',
lcol("`outcome_colour'") lpattern(`trend_pattern') lwidth(`trend_width')"'
   local num_lines = `num_lines' + 1
}

// Generate the graph using Stata's twoway command
local name = "graph_with_cls_3b"

graph twoway rarea counter_cl_lower counter_cl_upper time , fcolor(gray%10) lcolor(white%100)
lwidth(none) ///
          || rarea cl_lower cl_upper time if segment == 1, fcolor(gray%10) lcolor(white%100)
lwidth(none) ///
          || scatter outcome time, msize(`point_size') msym(`point_symbol')
mcol("`outcome_colour'`point_transparency'") /// outcomes
          || lfit counterfactual time if segment == 1, lcol("`outcome_colour'")
lpattern(`counterfactual_pattern') lwidth(`counterfactual_width') /// counterfactual
           `trend_lines' /// trend lines
          || line counter_cl_lower time , lc(gray%50) lpattern(dot) /// CL plots
          || line counter_cl_upper time , lc(gray%50) lpattern(dot) /// CL plots
          || line cl_lower time if segment == 1, lc(gray%50) lpattern(dot) /// CL plots
          || line cl_upper time if segment == 1, lc(gray%50) lpattern(dot) /// CL plots
          || scatteri `=`interruption_label_height'-5' 1 `=`interruption_label_height'-5' 6
`=`interruption_label_height'+5' 6 `=`interruption_label_height'+5' 1, recast(area)
fcolor("gray%10") lcol(gray%15) ///
          text(`interruption_label_height' 7 "95% confidence intervals", placement(e) size(vsmall))
///
```

```
          ylab(25(25)175,angle(h)) /// the angle ensures numbers are printed horizontally
          title("3b", position(10) ring(1)) ///
          ytitle(`y_title') ///
          xtitle(`x_title') ///
          xline(`time_1_start', lcol(`interruption_colour') lpattern(`interruption_pattern')
lwidth(`interruption_width')) /// interruption line
          text(`interruption_label_height' `time_0_end' "`interruption_label'", placement(w)) ///
interruption text label
          legend(off) /// no legend - detail will need to be added to the caption
          graphregion(col(white)) /// set background to white
          name(`name')

graph export "`graph_directory'\`name'.svg", replace
graph export "`graph_directory'\`name'.png", replace

drop graph_max
drop graph_min

graph combine seasonal_graph_3a graph_with_cls_3b, col(1) ysize(8) iscale(0.7273)
graphregion(margin(zero) col(white)) name(combined_graph_3)
graph export "`graph_directory'\combined_graph_3.svg", replace
graph export "`graph_directory'\combined_graph_3.png", replace

////////////////////////////////////////////////////////////////////////////////
// Code to generate graph in Figure 4a
// In this graph, we add the results of the analysis to the graph and add y-axis labels at points
of interest.
// Note that to ensure that no data points are obscured on the graph, the position of the
additional text may
// need some fine tuning for each graph

// Find minimum and maximum points on the graph
egen graph_max = rowmax(outcome counterfactual)
egen graph_min = rowmin(outcome counterfactual)
summ graph_max
local graph_max = r(max)
summ outcome
local outcome_max = r(max)
summ graph_min
local graph_min = r(min)

// Save the y-coordinates of the counterfactual and the estimate at the interruption to local
macros
local change_counter = counterfactual[51]
local estimate_counter = estimates[51]

// Save the values for the colour and other graph parameters to local macros:
local outcome_colour = "0 114 178" // colour blind friendly blue

local point_size = "small"
local point_symbol = "plus"
local point_transparency = "%70" // this can be adjusted depending on number of points

local trend_pattern = "solid"
local trend_width = "medthick"

// Save the options for the counterfactual lines to local macro
local counterfactual_pattern = "dash"
local counterfactual_width = "medium"

// Save the options for the interruption lines to local macro
local interruption_colour = "red"
local interruption_pattern = "dash"
local interruption_width = "medium"

// Save the options for the titles and labels to local macro
local x_title "months since start of project"
local y_title "outcome details"
local interruption_label "interruption"
local interruption_label_height = `graph_max' + 5

// Generate the text for the graphing of the trend lines
local num_line = 0
local trend_lines = ""

forvalues segment = 0/`num_segments' {
   local trend_lines = `"`trend_lines'"' + `" || lfit estimates time if segment == `segment',
lcol("`outcome_colour'") lpattern(`trend_pattern') lwidth(`trend_width')"'
   local num_lines = `num_lines' + 1
}

// Generate the graph using Stata's twoway command
local name = "simple_graph_added_text_4a"

graph twoway scatter outcome time, msize(`point_size') msym(`point_symbol')
mcol("`outcome_colour'`point_transparency'") /// outcomes
          || lfit counterfactual time if segment == 1, lcol("`outcome_colour'")
lpattern(`counterfactual_pattern') lwidth(`counterfactual_width') /// counterfactual
          `trend_lines' /// trend lines
          ylab(`graph_min' `outcome_max' `change_counter' `estimate_counter',angle(h)
format(%3.0fc)) ///
          title("4a", position(10) ring(1)) ///
          ytitle(`y_title') ///
          xtitle(`x_title') ///
          xline(`time_1_start', lcol(`interruption_colour') lpattern(`interruption_pattern')
lwidth(`interruption_width')) /// interruption line
          text(`interruption_label_height' `time_0_end' "`interruption_label'", placement(w)) ///
interruption text label
          text(55 -1 "level change: `level_change', 95% CI (`level_change_ll', `level_change_ul')",
placement(e) size(small)) /// level change details
          text(45 -1 "slope change: `slope_change', 95% CI (`slope_change_ll', `slope_change_ul')",
placement(e) size(small)) /// slope change details
          legend(off) /// legend without border
          graphregion(col(white)) /// set background to white
          name(`name')

graph export "`graph_directory'\`name'.svg", replace
graph export "`graph_directory'\`name'.png", replace

drop graph_max
drop graph_min

////////////////////////////////////////////////////////////////////////////////
// Code to generate graph in Figure 4b
// In this graph, there are features that go against recommendations.
// That is, the box around the legend unnecessarily clutters the image;
// the y-axis labels are vertical making them more difficult to read.

// Save the values for the colour and other graph parameters to local macros:
local outcome_colour = "0 114 178" // colour blind friendly blue

local point_size = "small"
local point_symbol = "plus"
local point_transparency = "%70" // this can be adjusted depending on number of points

local trend_pattern = "solid"
local trend_width = "medthick"

// Save the options for the counterfactual lines to local macro
local counterfactual_pattern = "dash"
local counterfactual_width = "medium"

// Save the options for the interruption lines to local macro
local interruption_colour = "red"
local interruption_pattern = "dash"
local interruption_width = "medium"

// Save the options for the titles and labels to local macro
local x_title "months since start of project"
local y_title "outcome details"
local interruption_label "interruption"
local interruption_label_height = `graph_max' + 5

// Generate the text for the graphing of the trend lines
local num_line = 0
local trend_lines = ""

forvalues segment = 0/`num_segments' {
   local trend_lines = `"`trend_lines'"' + `" || lfit estimates time if segment == `segment',
lcol("`outcome_colour'") lpattern(`trend_pattern') lwidth(`trend_width')"'
   local num_lines = `num_lines' + 1
}

// Generate the graph using Stata's twoway command
local name = "graph_too_simple_4b"

graph twoway scatter outcome time, msize(`point_size') msym(`point_symbol')
mcol("`outcome_colour'`point_transparency'") /// outcomes
```

```
          || lfit counterfactual time if segment == 1, lcol("`outcome_colour'")
lpattern(`counterfactual_pattern') lwidth(`counterfactual_width') /// counterfactual
          `trend_lines' /// trend lines
          title("4b", position(10) ring(1)) ///
          ytitle(`y_title') ///
          xtitle(`x_title') ///
          xline(`time_1_start', lcol(`interruption_colour') lpattern(`interruption_pattern')
lwidth(`interruption_width')) /// interruption line
          text(`interruption_label_height' `time_0_end' "`interruption_label'", placement(w)) ///
interruption text label
          legend(order(1 "outcome" 3 "pre-interruption trend" 4 "post-interruption trend" 2
"counterfactual")) /// legend
          graphregion(col(white)) /// set background to white
          name(`name')

graph export "`graph_directory'\`name'.svg", replace
graph export "`graph_directory'\`name'.png", replace

graph combine simple_graph_added_text_4a graph_too_simple_4b, col(1) ysize(8) iscale(0.7273)
graphregion(margin(zero) col(white)) name(combined_graph_4)

graph export "`graph_directory'\combined_graph_4.svg", replace
graph export "`graph_directory'\combined_graph_4.png", replace

exit
```

Appendix F. Supplementary file 1 accompanying Chapter 5 – Supplementary graphs

“Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study”

Turner SL, Forbes AB, Karahalios A, Taljaard M, McKenzie JE.

The following are the supplementary graphs referenced in the main manuscript.

Supplementary 1.1 Interrupted time series graphs

The following, Figure S1, shows an example of a simulated data set using the parameterisation shown in equation 1, section 3.1.

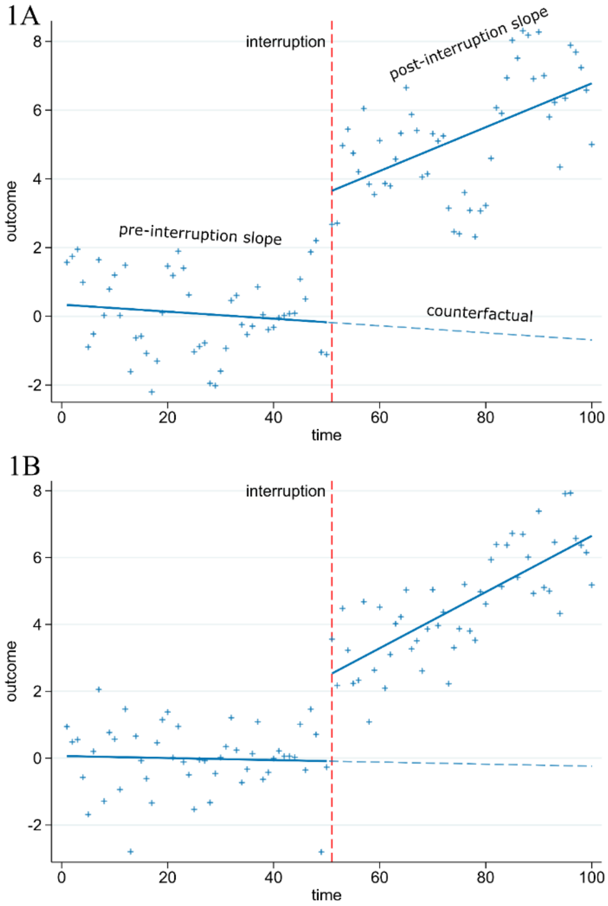


Figure S5: An example of two simulated data sets. This figure was created using the model ($Y_t = \beta_0 + \beta_1 t + \beta_2 D_t + \beta_3[t - T_I]D_t + \varepsilon_t$ where $\varepsilon_t = \rho\, \varepsilon_{t-1} + N(0,\sigma^2)$), with parameters $\beta_0 = 0$ (y-intercept), $\beta_1 = 0$ (pre-interruption slope), $\beta_2 = 2$ (level change at time of interruption), $\beta_3 = 0.1$ (slope change after interruption), and σ=1, D_t is an indicator variable that is 0 for pre-interruption times and 1 for post-interruption times. In Figure 1A the autocorrelation (ρ) is 0.8, in Figure 1B it is 0.

Supplementary 1.2 Slope change estimates

The following, Figure S2, shows the distribution of slope change estimates for parameters 2 for level change and 0.1 for slope change.

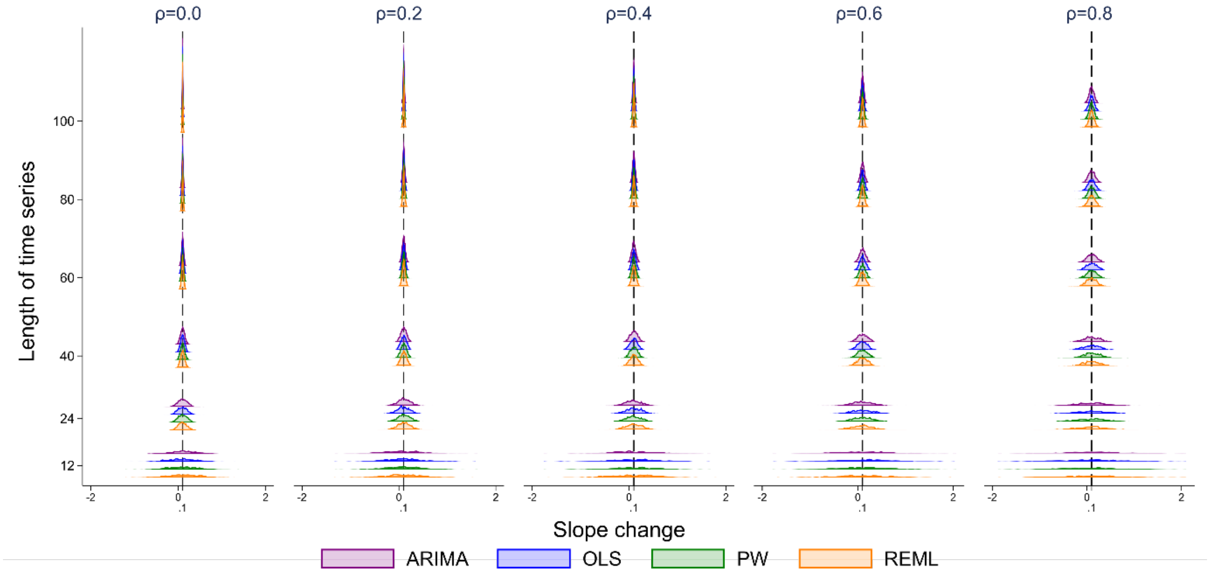


Figure S6: Distributions of slope change estimates calculated from four statistical methods, from top to bottom: autoregressive integrated moving average (ARIMA) (purple), ordinary least squares regression (OLS) (blue), Prais-Winsten (PW) (green) and restricted maximum likelihood (REML) (orange). The vertical axis shows the length of the time series. The five vertical columns display the results for different values of autocorrelation. The vertical black line represents the true parameter value (β_3). Each subset of four curves shows the distribution from a different analysis method for a given combination of time series length and autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. The Satterthwaite adjustment to the REML method and the Newey-West adjustment to the OLS method do not impact the estimate of level or slope change, hence these parameter estimates are not shown.

Supplementary 1.3 Differences between slope and level change parameters

The following nested loop plots (Rücker and Schwarzer 2014) show the similarities in estimates and performance measures across the eight different level and slope change parameters. Each statistical method is denoted using different coloured and shaped points across each combination of time series length, level change and slope change.

S 1.3.1 Level change bias

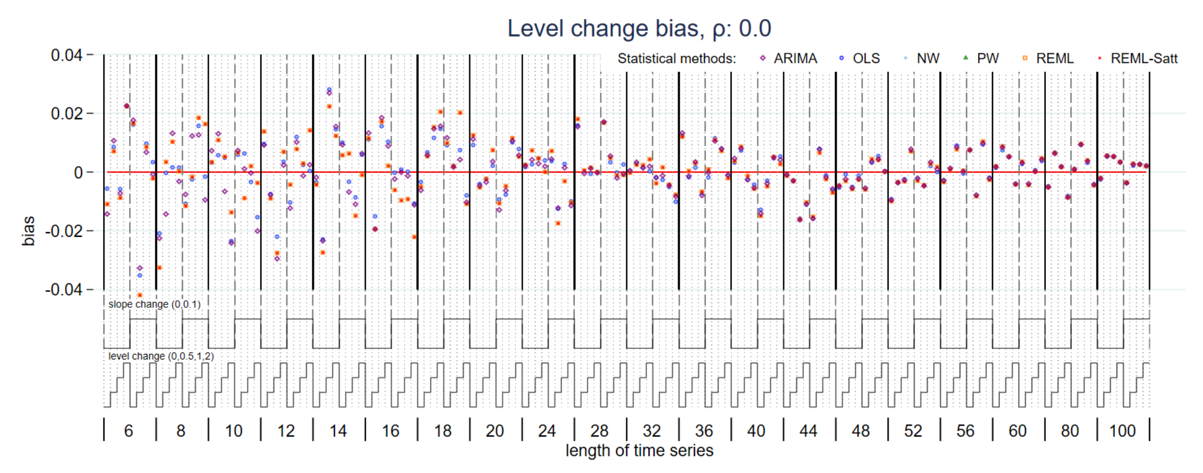


Figure S7: Bias in level change estimate for magnitude of autocorrelation 0. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

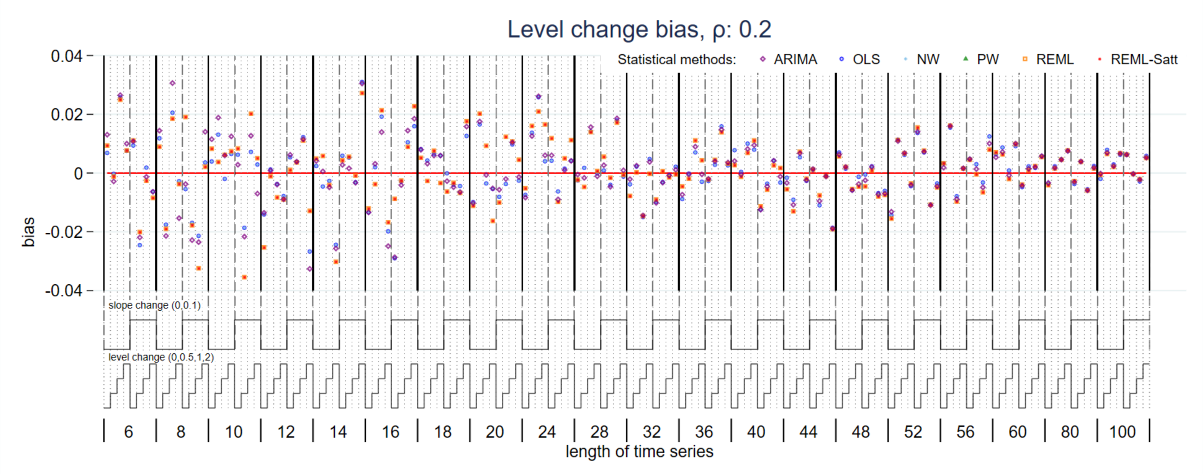


Figure S8: Bias in level change estimate for magnitude of autocorrelation 0.2. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

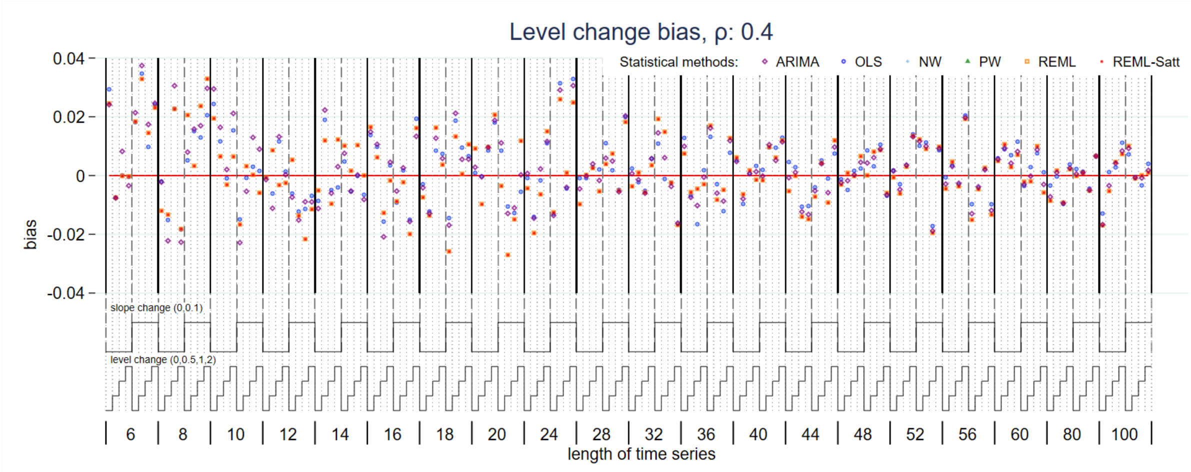


Figure S9: Bias in level change estimate for magnitude of autocorrelation 0.4. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

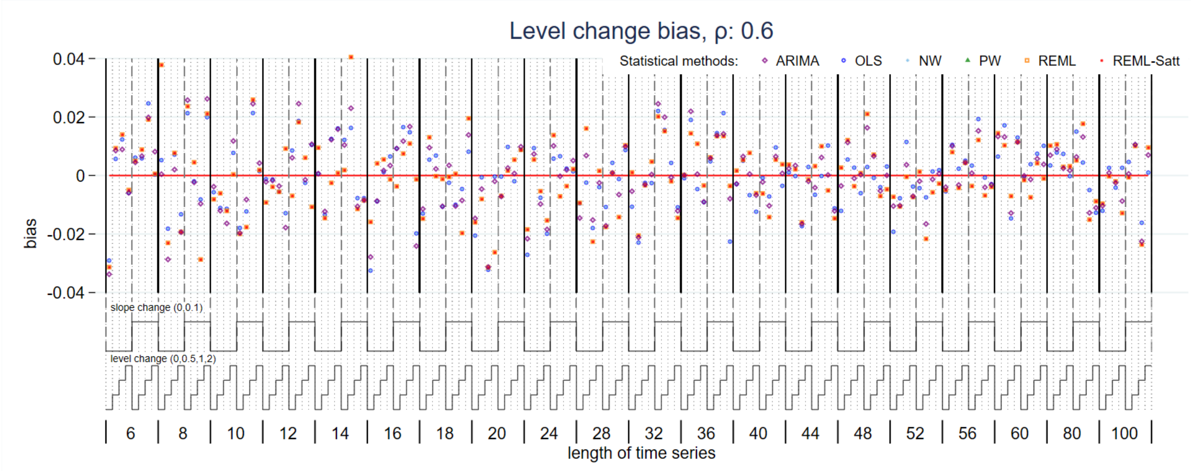


Figure S10: Bias in level change estimate for magnitude of autocorrelation 0.6. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

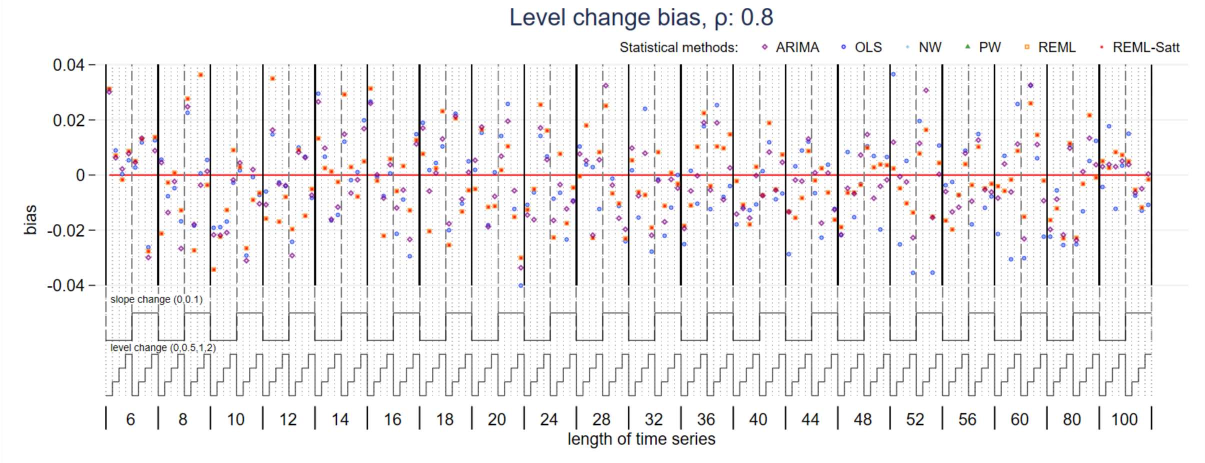


Figure S11: Bias in level change estimate for magnitude of autocorrelation 0.8. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

S 1.3.2 Slope change bias

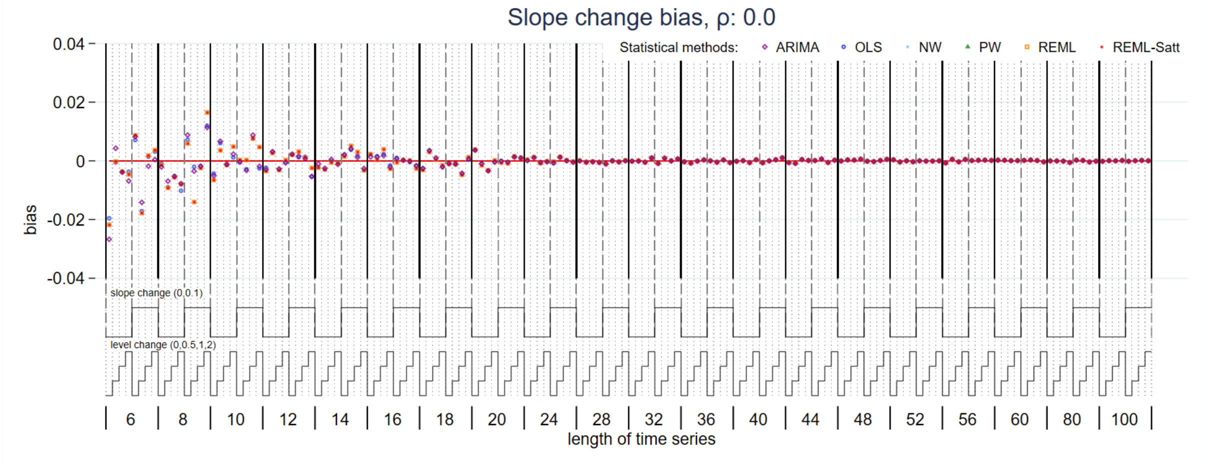


Figure S12: Bias in slope change estimate for magnitude of autocorrelation 0. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

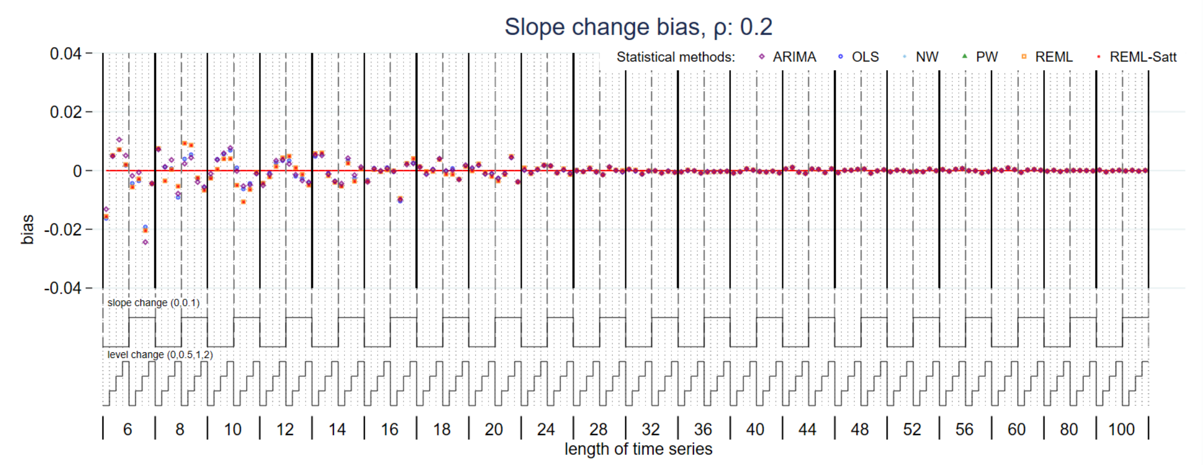


Figure S13: Bias in slope change estimate for magnitude of autocorrelation 0.2. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

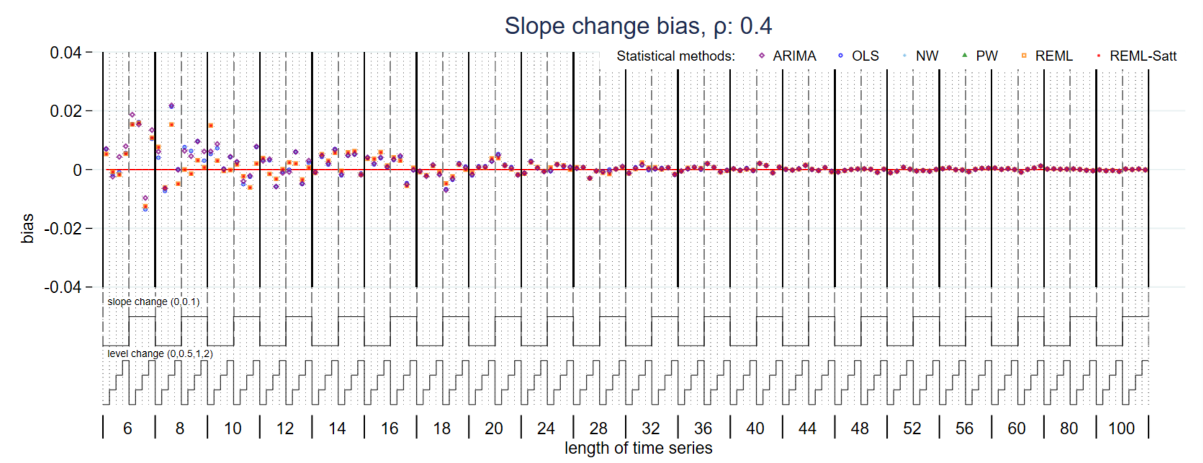


Figure S14: Bias in slope change estimate for magnitude of autocorrelation 0.4. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

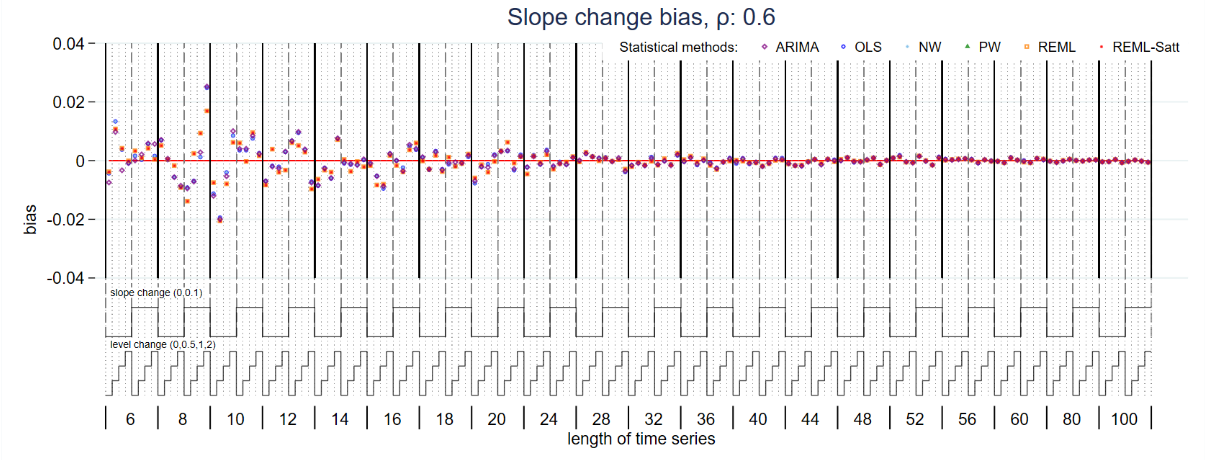


Figure S15: Bias in slope change estimate for magnitude of autocorrelation 0.6. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

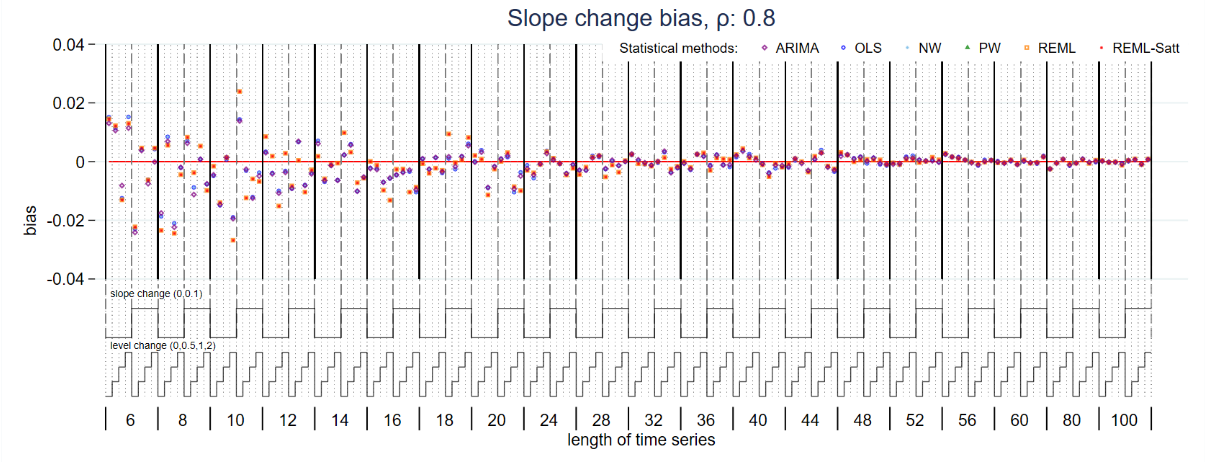


Figure S16: Bias in slope change estimate for magnitude of autocorrelation 0.8. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

S 1.3.3 Level change empirical standard error

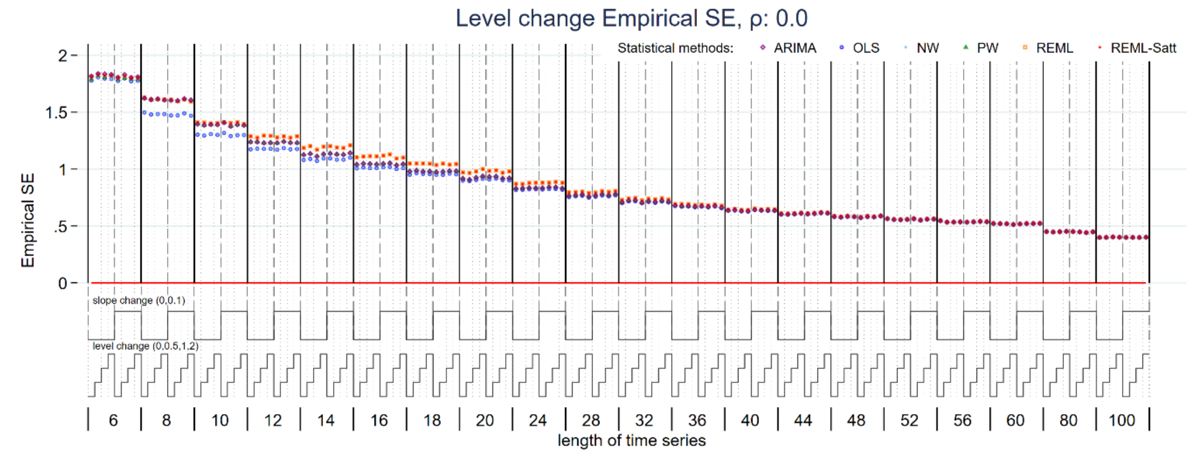


Figure S17: Empirical standard error of level change estimate for magnitude of autocorrelation 0. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

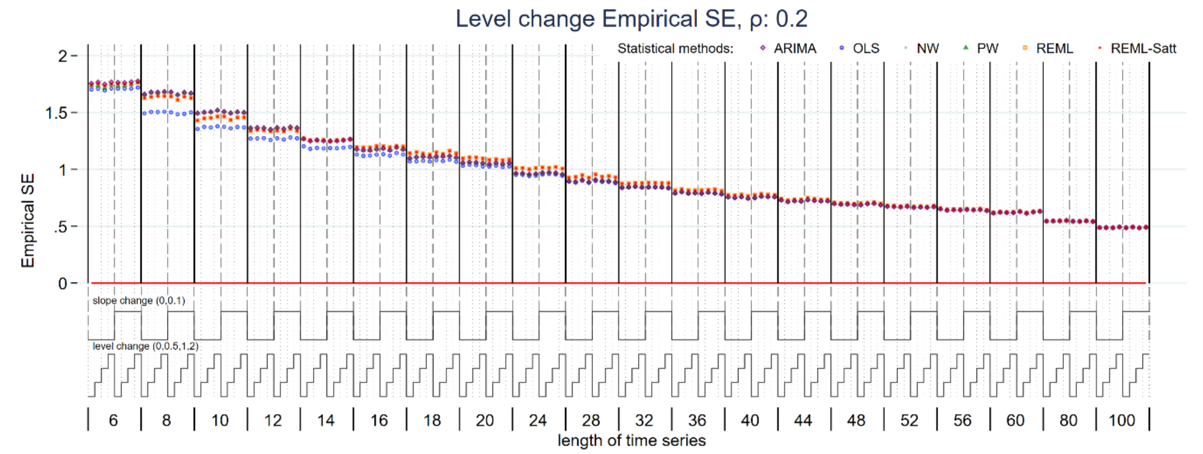


Figure S18: Empirical standard error of level change estimate for magnitude of autocorrelation 0.2. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

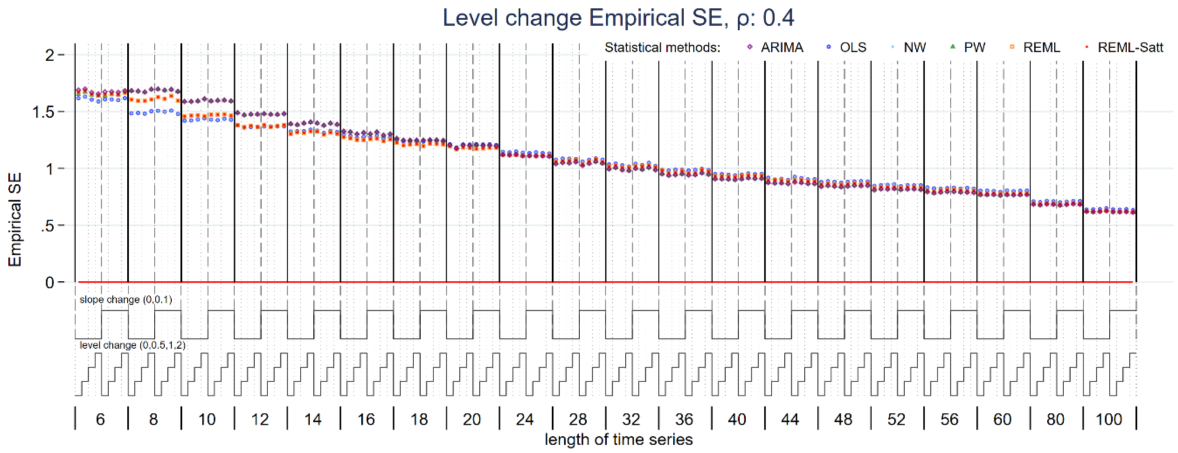


Figure S19: Empirical standard error of level change estimate for magnitude of autocorrelation 0.4. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

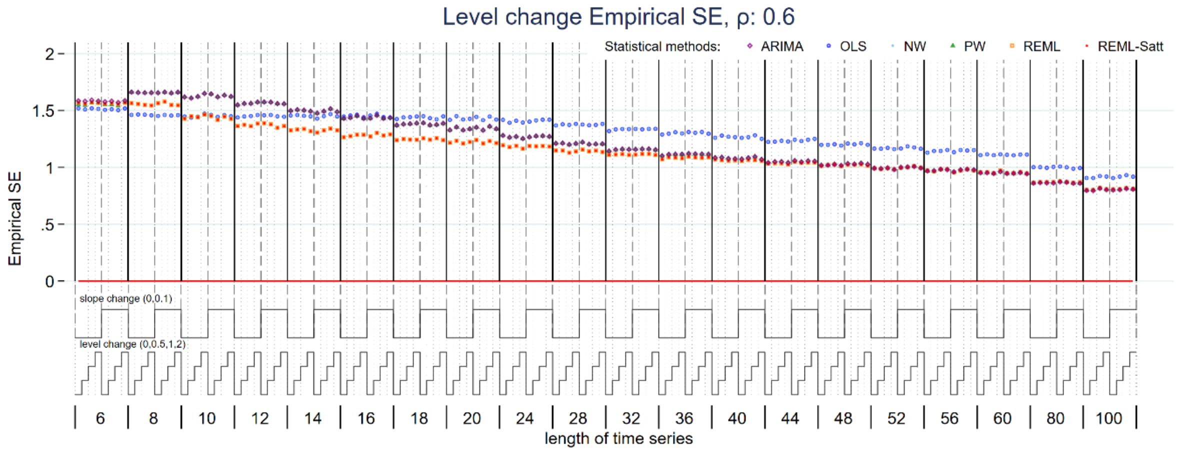


Figure S20: Empirical standard error of level change estimate for magnitude of autocorrelation 0.6. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

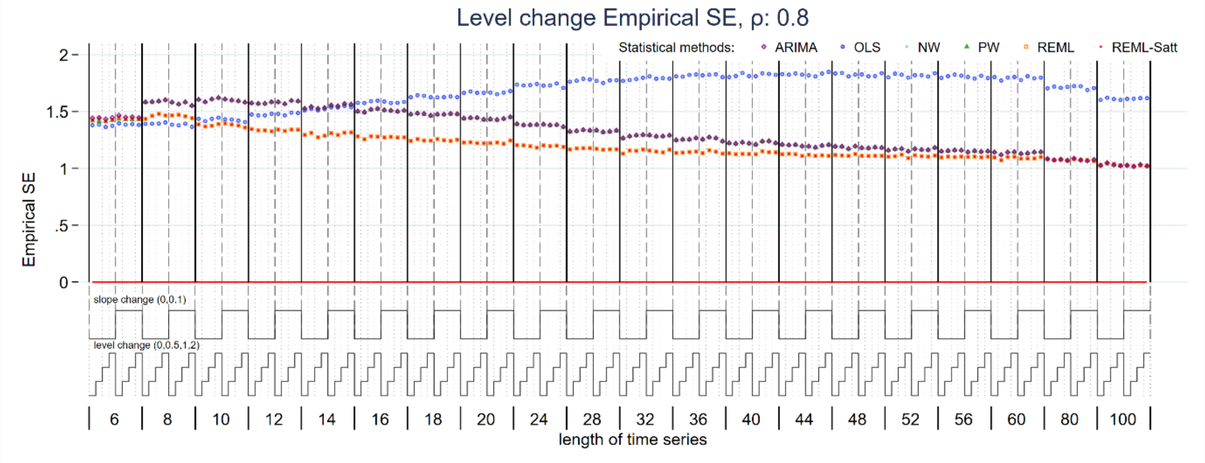


Figure S21: Empirical standard error of level change estimate for magnitude of autocorrelation 0.8. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

S 1.3.4 Slope change empirical standard error

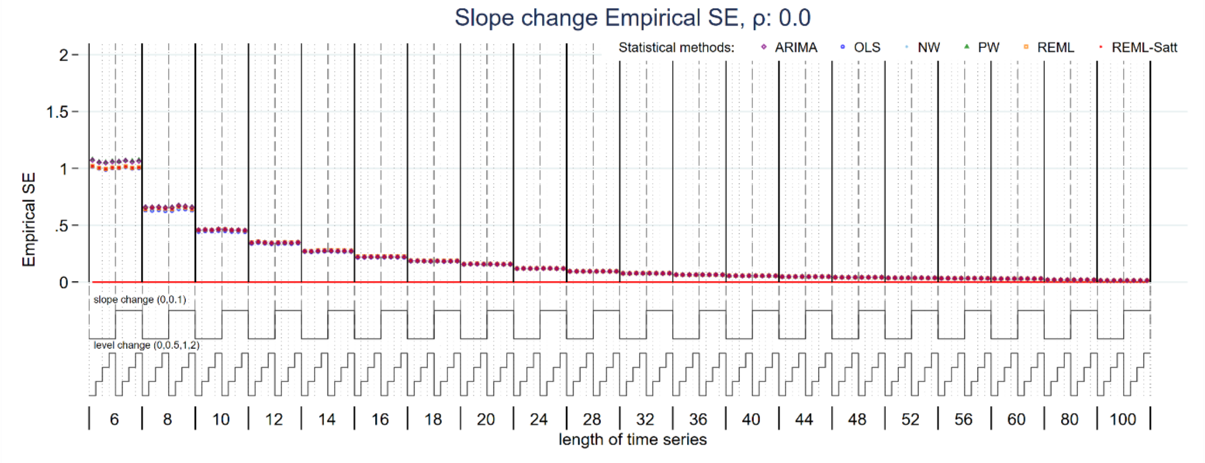


Figure S22: Empirical standard error of slope change estimate for magnitude of autocorrelation 0. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

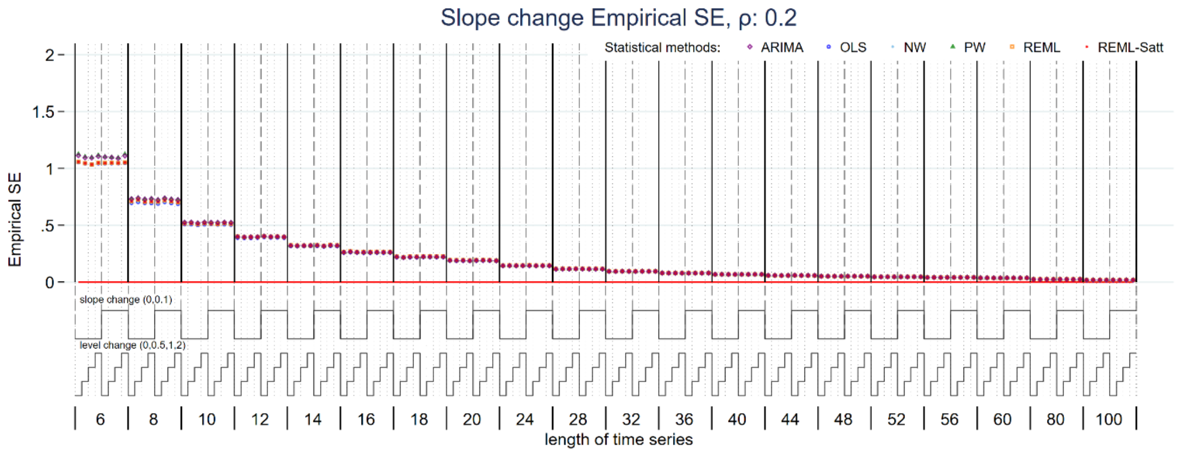


Figure S23: Empirical standard error of slope change estimate for magnitude of autocorrelation 0.2. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

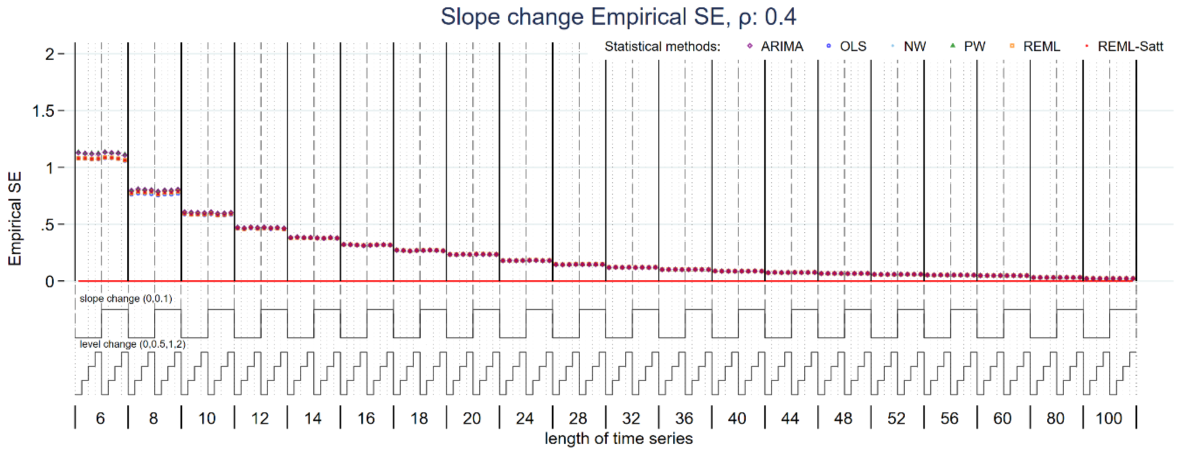


Figure S24: Empirical standard error of slope change estimate for magnitude of autocorrelation 0.4. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

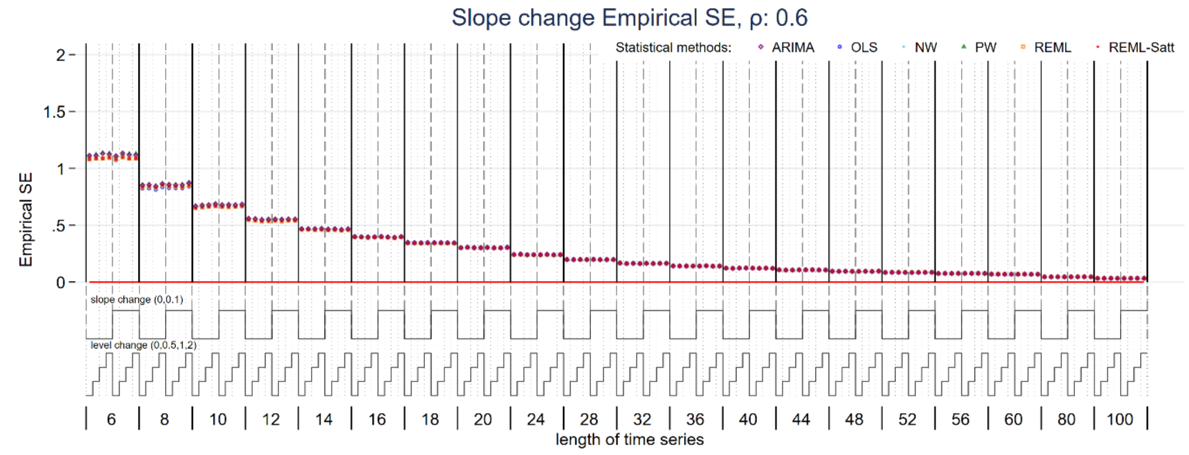


Figure S25: Empirical standard error of slope change estimate for magnitude of autocorrelation 0.6. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

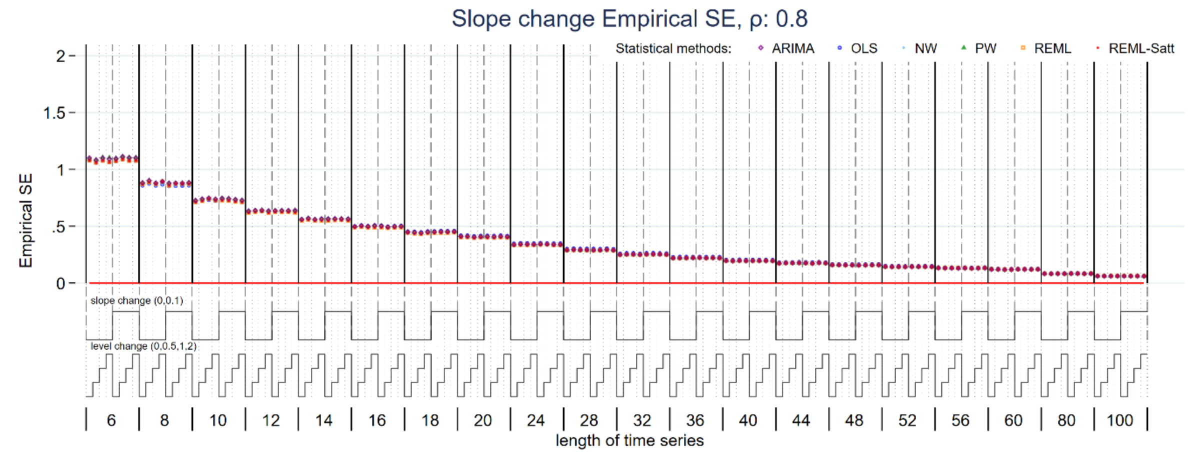


Figure S26: Empirical standard error of slope change estimate for magnitude of autocorrelation 0.8. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

S 1.3.5 Level change model-based standard error

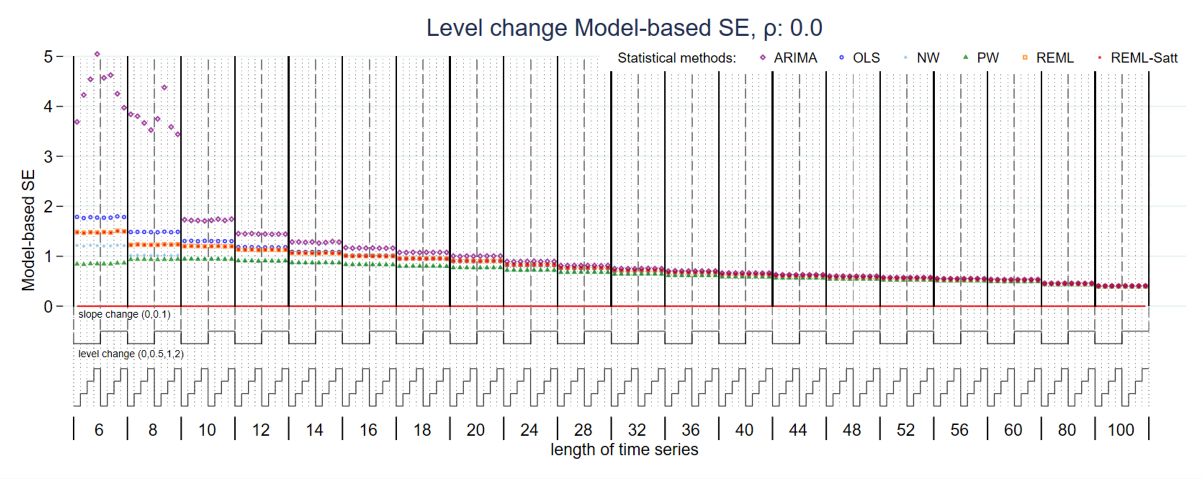


Figure S27: Model-based standard error of level change estimate for magnitude of autocorrelation 0. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

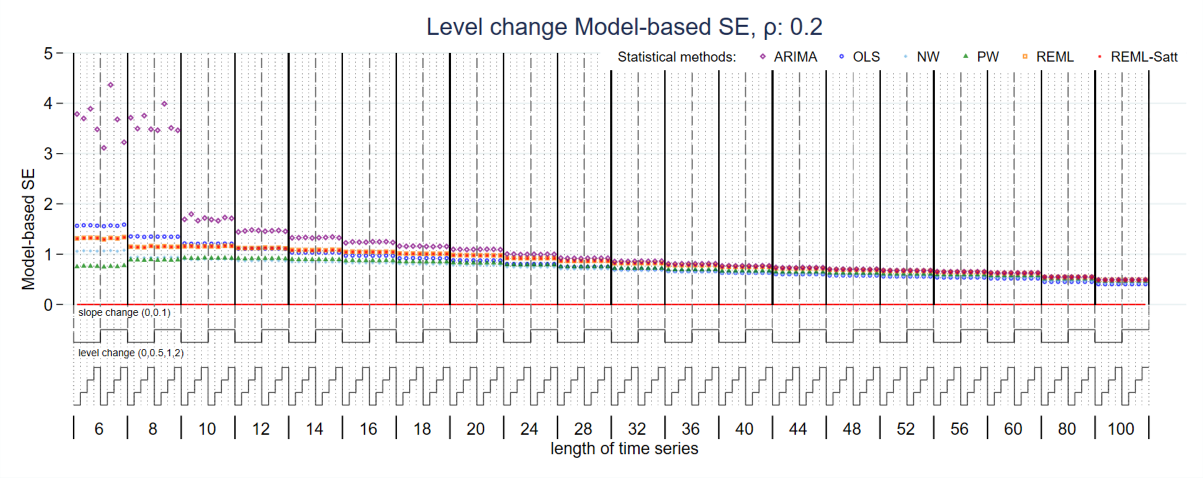


Figure S28: Model-based standard error of level change estimate for magnitude of autocorrelation 0.2. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

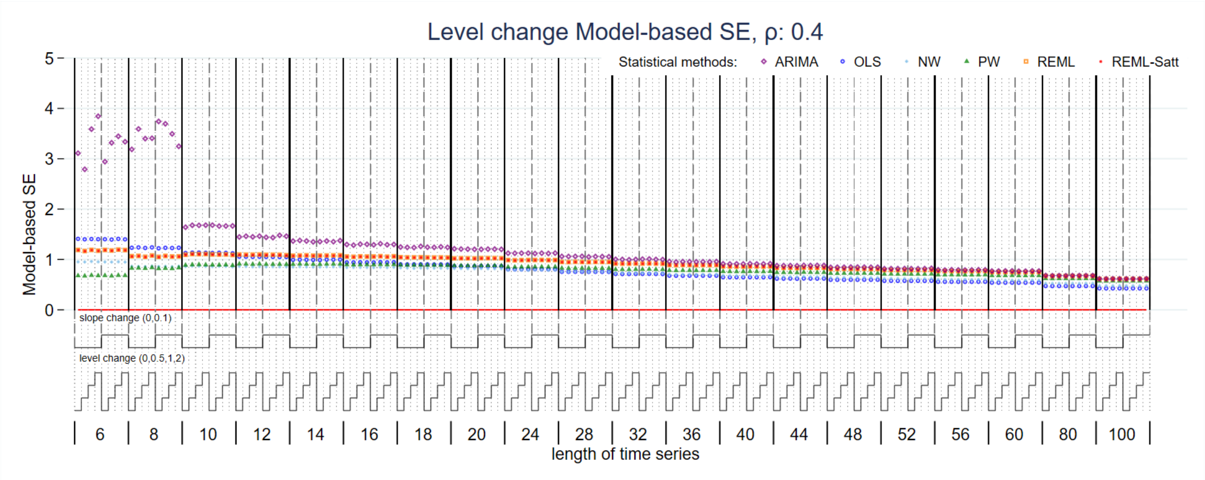


Figure S29: Model-based standard error of level change estimate for magnitude of autocorrelation 0.4. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

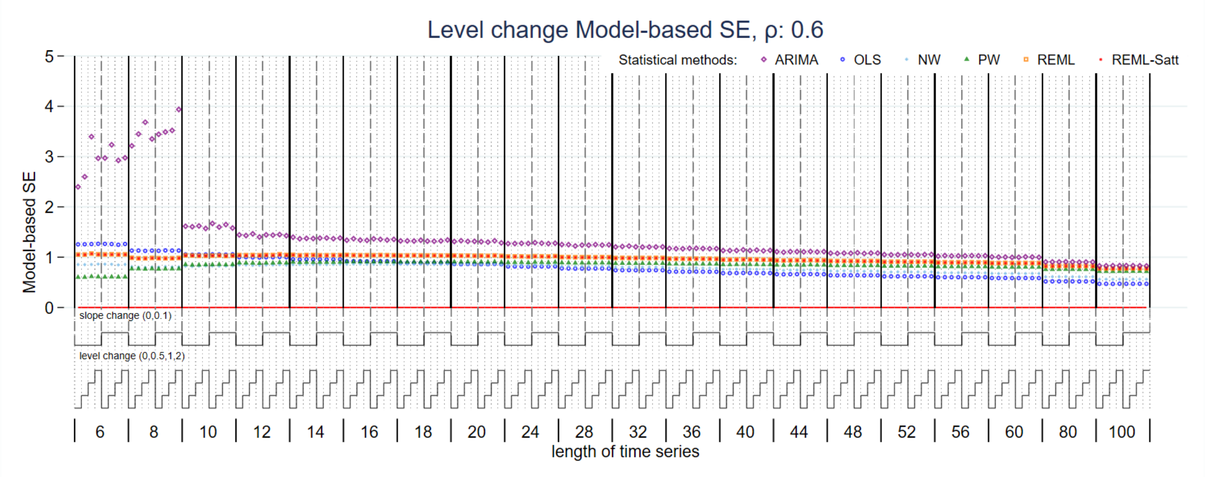


Figure S30: Model-based standard error of level change estimate for magnitude of autocorrelation 0.6. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

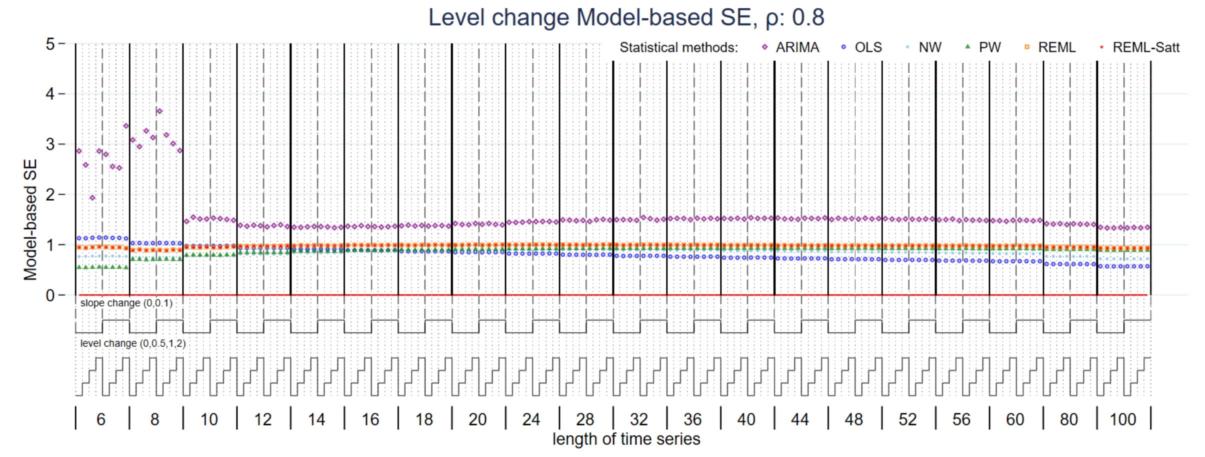


Figure S31: Model-based standard error of level change estimate for magnitude of autocorrelation 0.8. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

S 1.3.6 Slope change model-based standard error

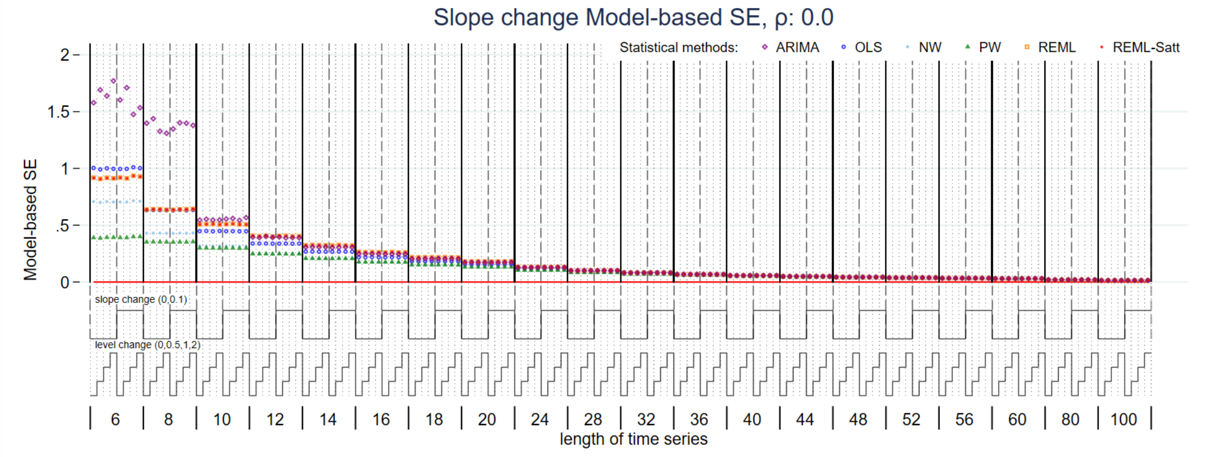


Figure S32: Model-based standard error of slope change estimate for magnitude of autocorrelation 0. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

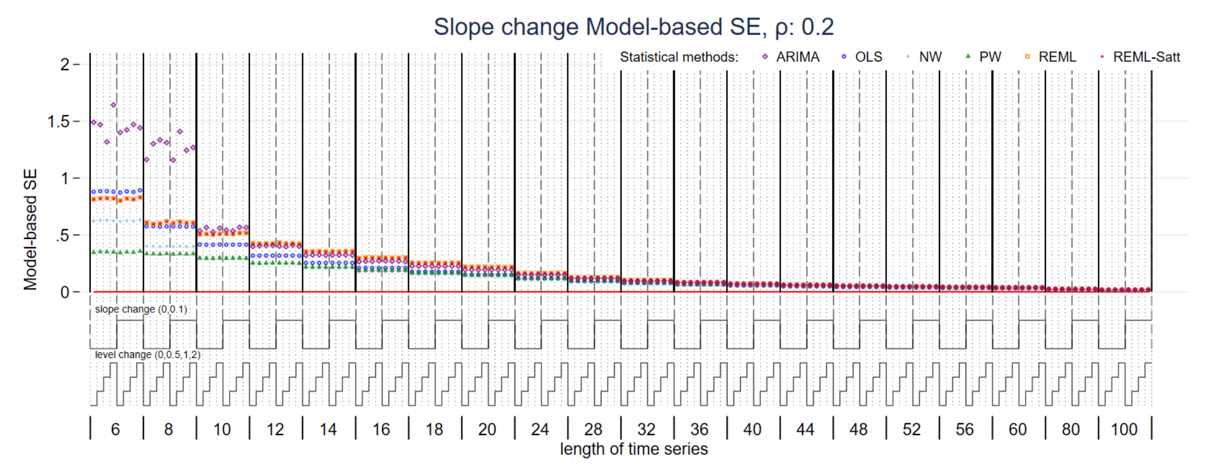


Figure S33: Model-based standard error of slope change estimate for magnitude of autocorrelation 0.2. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

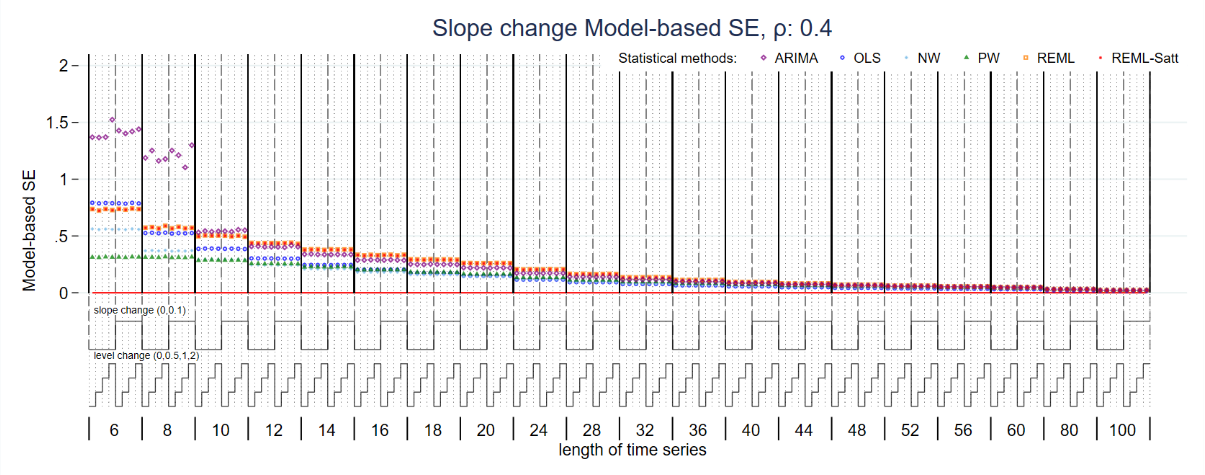


Figure S34: Model-based standard error of slope change estimate for magnitude of autocorrelation 0.4. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

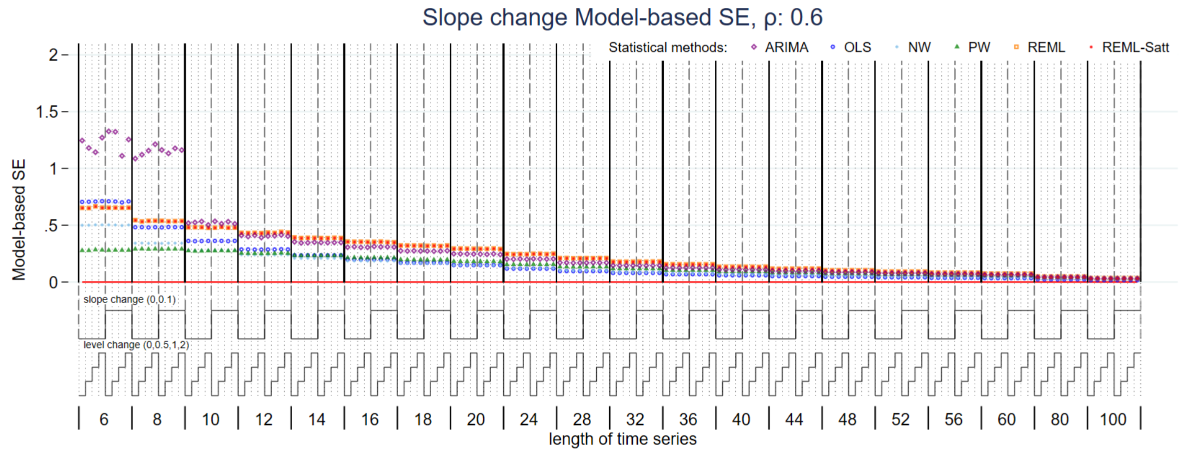


Figure S35: Model-based standard error of slope change estimate for magnitude of autocorrelation 0.6. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

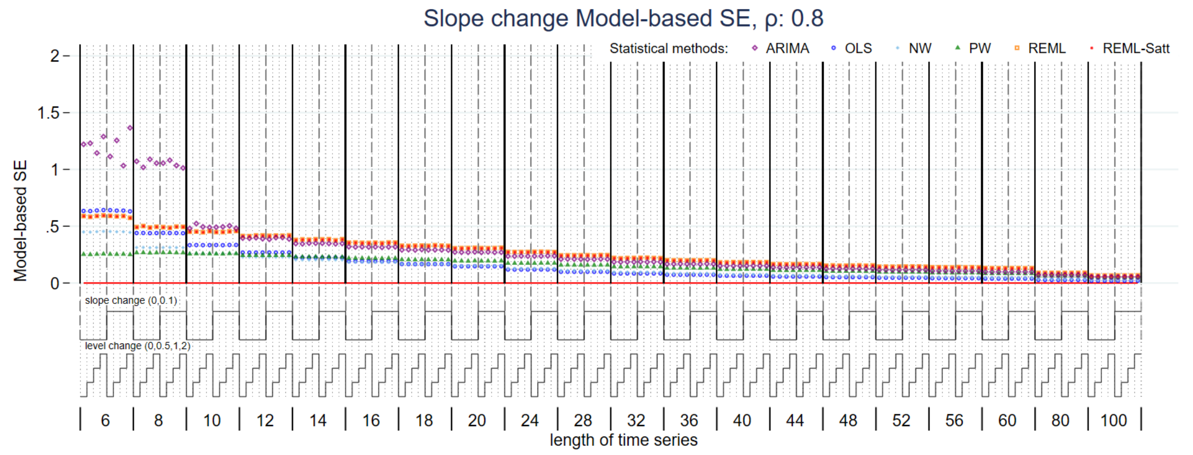


Figure S36: Model-based standard error of slope change estimate for magnitude of autocorrelation 0.8. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

S 1.3.7 Level change coverage

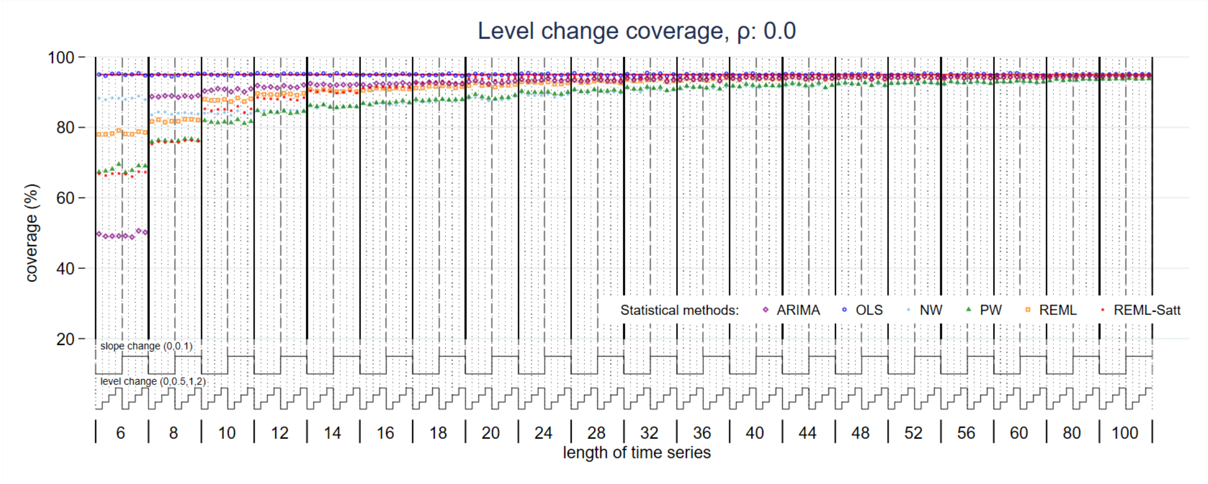


Figure S37: Coverage of level change estimate for magnitude of autocorrelation 0. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

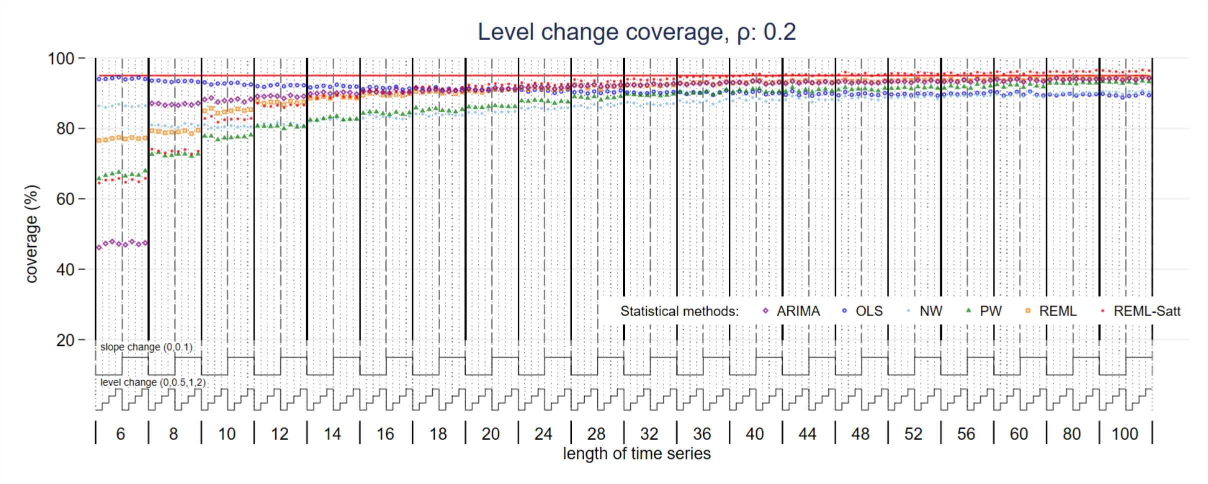


Figure S38: Coverage of level change estimate for magnitude of autocorrelation 0.2. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

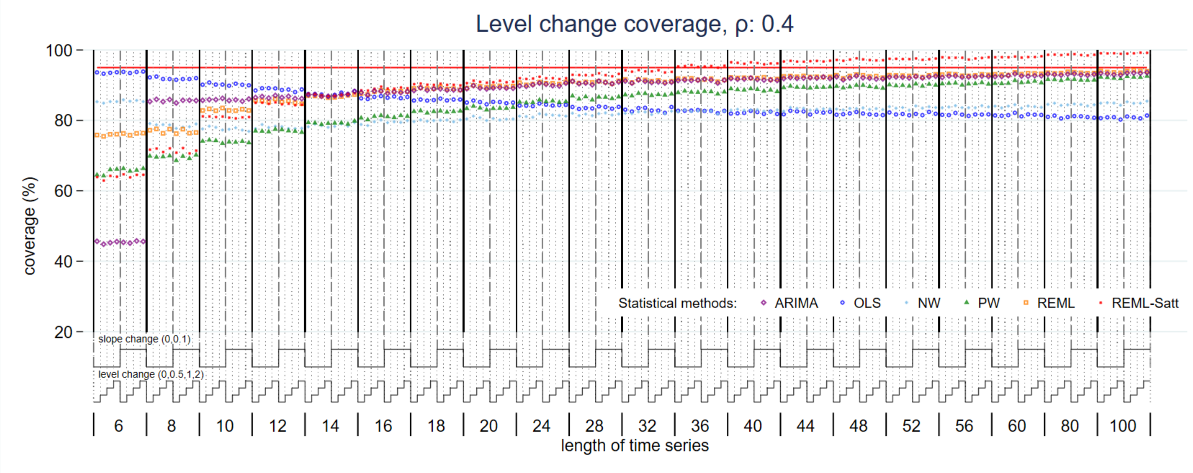


Figure S39: Coverage of level change estimate for magnitude of autocorrelation 0.4. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

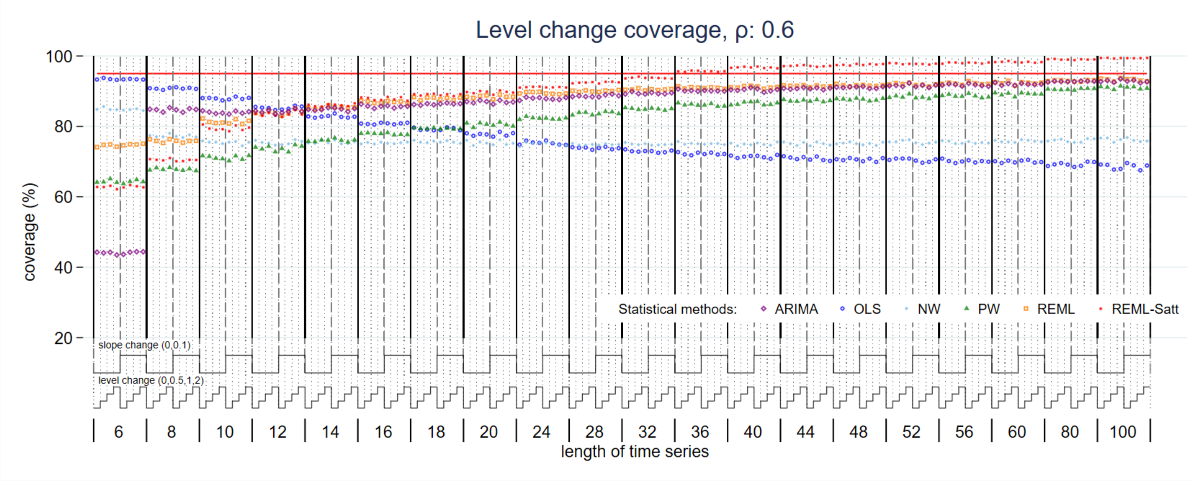


Figure S40: Coverage of level change estimate for magnitude of autocorrelation 0.6. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

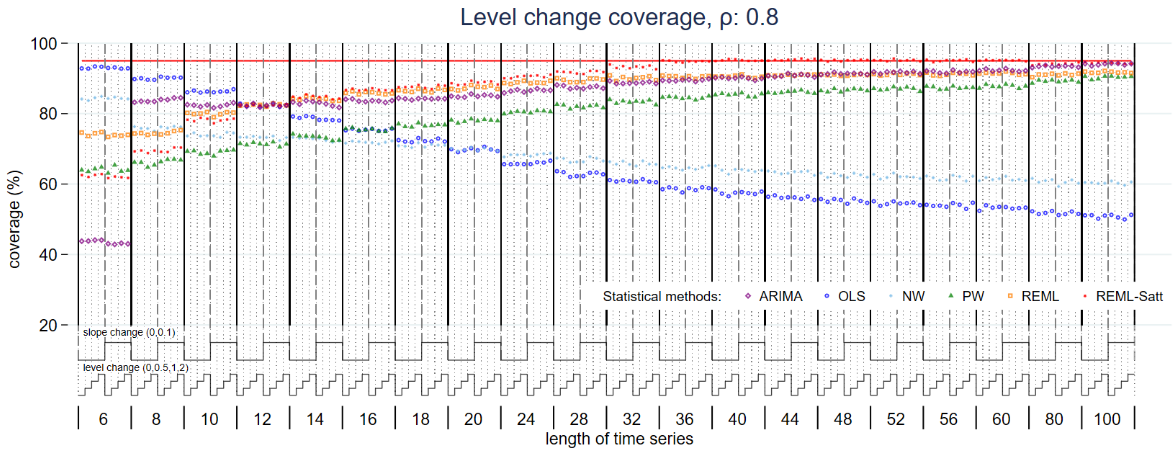


Figure S41: Coverage of level change estimate for magnitude of autocorrelation 0.8. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

S 1.3.8 Slope change coverage

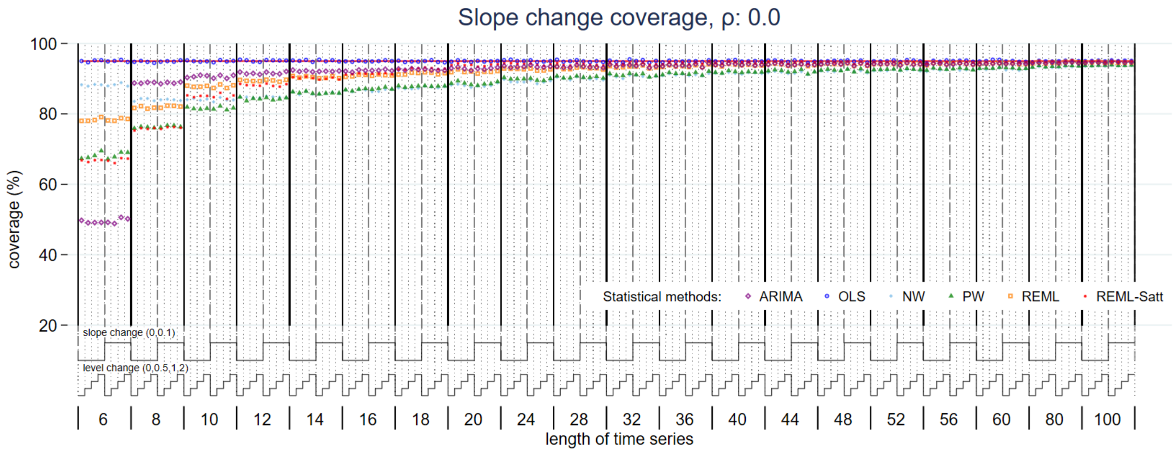


Figure S42: Coverage of slope change estimate for magnitude of autocorrelation 0. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

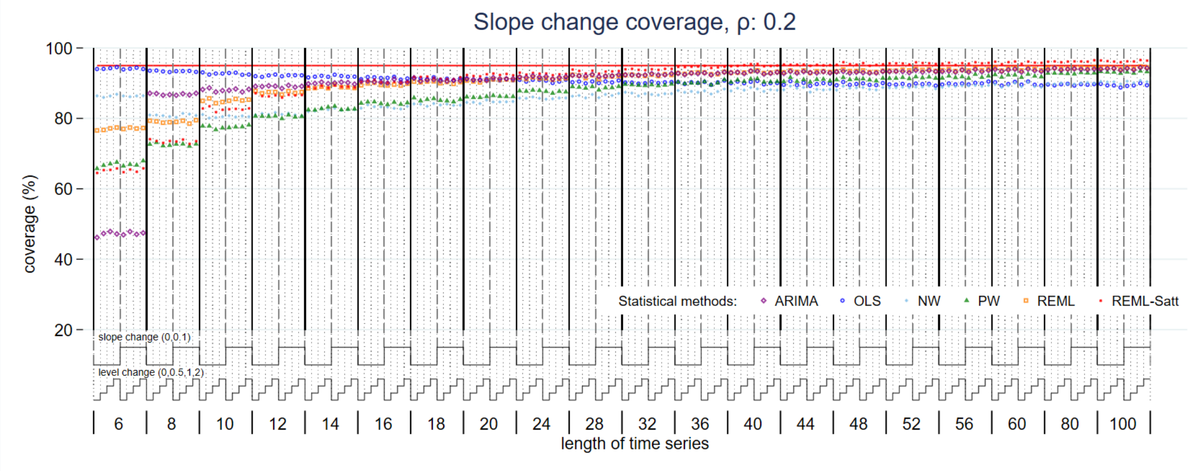


Figure S43: Coverage of slope change estimate for magnitude of autocorrelation 0.2. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

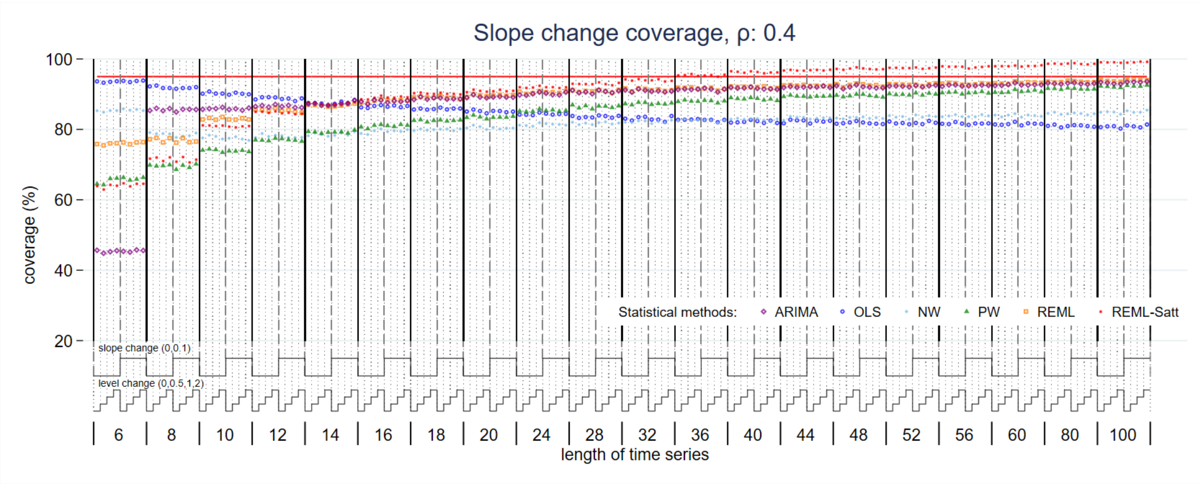


Figure S44: Coverage of slope change estimate for magnitude of autocorrelation 0.4. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

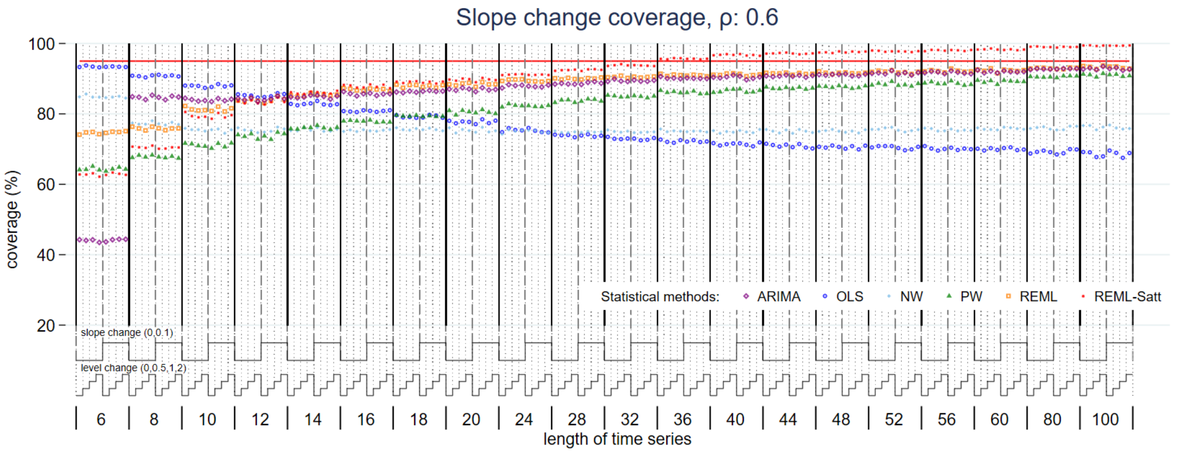


Figure S45: Coverage of slope change estimate for magnitude of autocorrelation 0.6. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

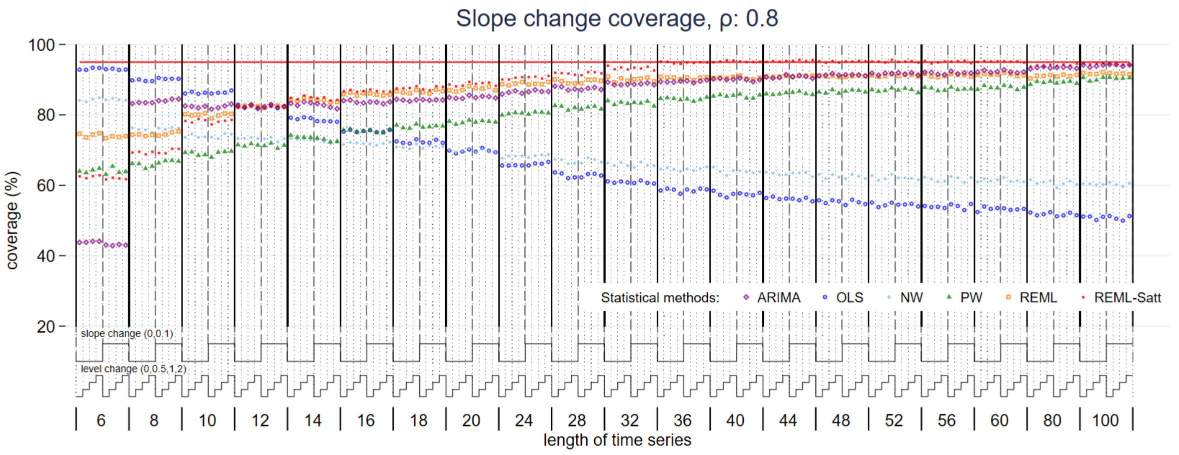


Figure S46: Coverage of slope change estimate for magnitude of autocorrelation 0.8. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

.

S 1.3.9 Estimate of autocorrelation

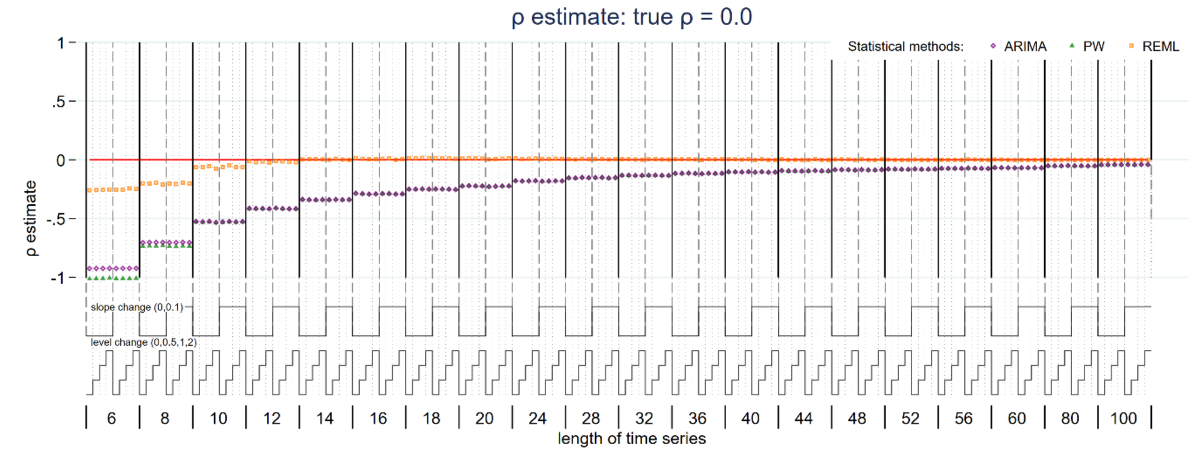


Figure S47: Autocorrelation estimate for true magnitude of autocorrelation 0. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; PW, Prais-Winsten; REML, restricted maximum likelihood.

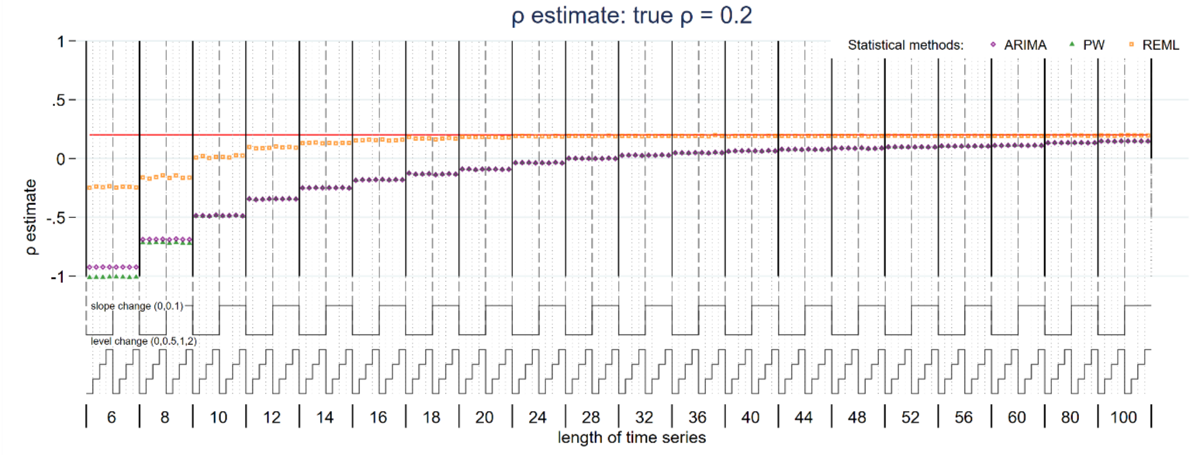


Figure S48: Autocorrelation estimate for true magnitude of autocorrelation 0.2. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; PW, Prais-Winsten; REML, restricted maximum likelihood.

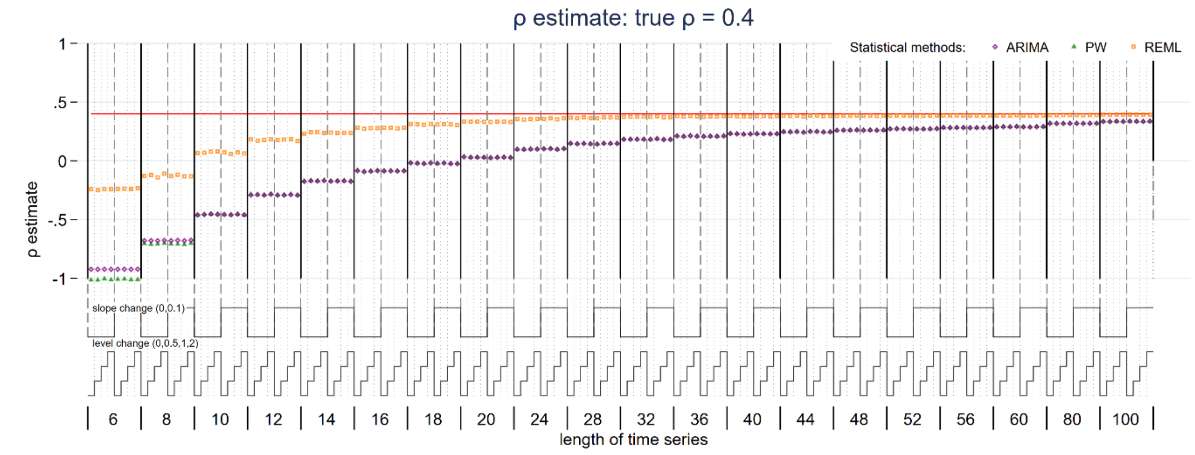


Figure S49: Autocorrelation estimate for true magnitude of autocorrelation 0.4. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; PW, Prais-Winsten; REML, restricted maximum likelihood.

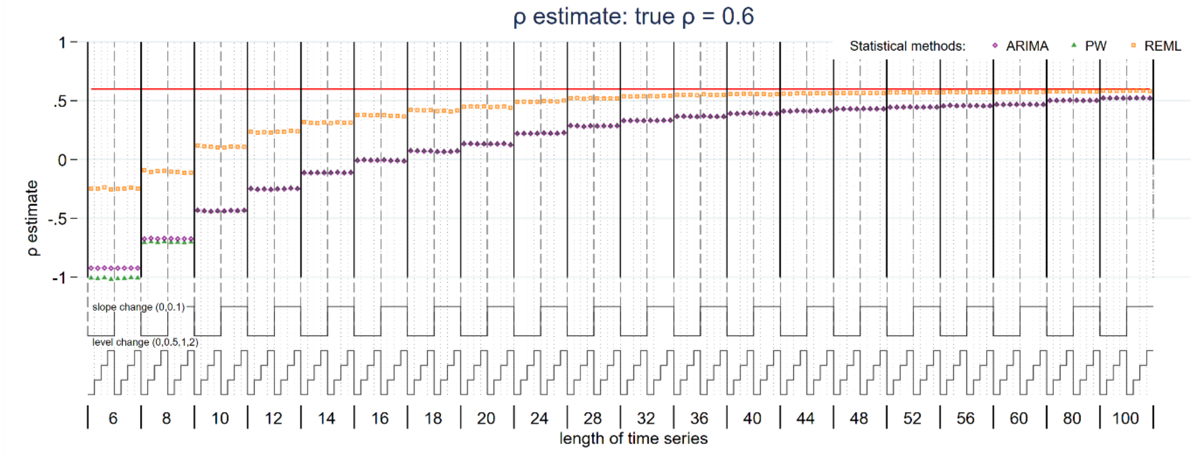


Figure S50: Autocorrelation estimate for true magnitude of autocorrelation 0.6. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; PW, Prais-Winsten; REML, restricted maximum likelihood.

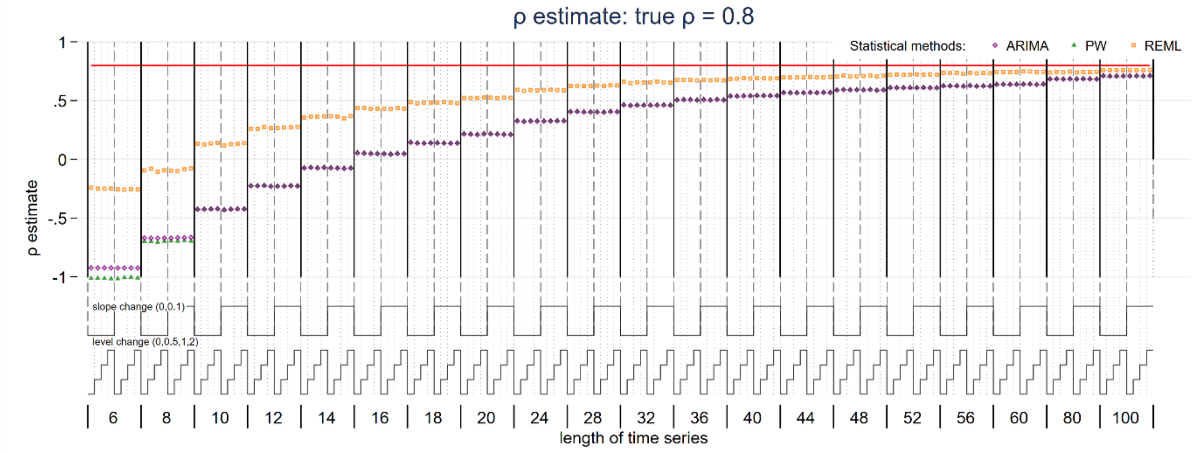


Figure S51: Autocorrelation estimate for true magnitude of autocorrelation 0.8. Each data point shows the mean value from 10,000 simulations for a given combination of slope change, level change and length of time series. Abbreviations: ARIMA, autoregressive integrated moving average; PW, Prais-Winsten; REML, restricted maximum likelihood.

Supplementary 1.4 Standard error of level change parameter for OLS

Additional simulations were undertaken to examine the behaviour of the OLS standard error estimator for the level change for additional autocorrelations (0.7 and 0.9). This showed a similar pattern of initially increasing SEs with an increasing number of points, which we had observed in the main simulation study for an underlying autocorrelation of 0.8.

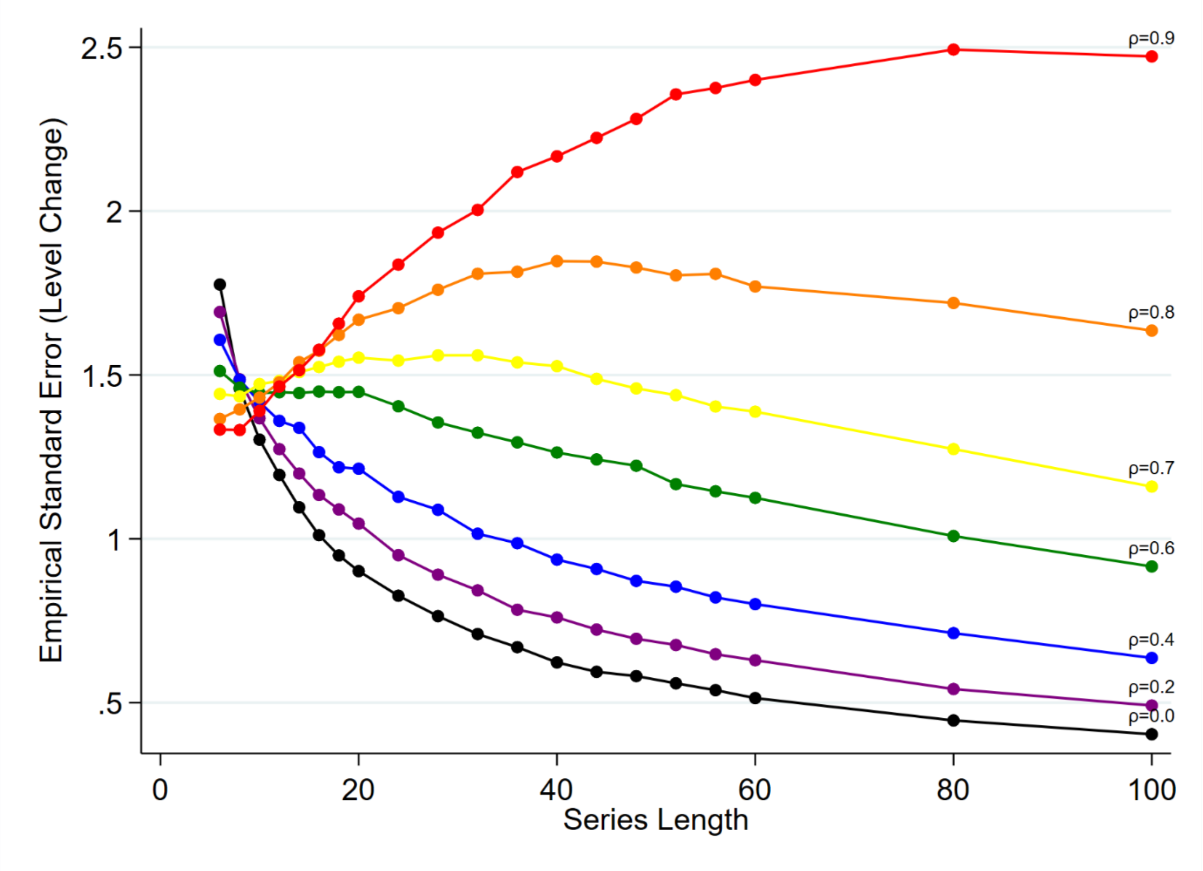


Figure S52: Empirical standard error (SE) of level change parameter for ordinary least squares regression. The horizontal axis shows the length of the time series, the vertical axis shows the empirical SE. Each coloured line shows the results for a different magnitude of autocorrelation ranging from 0 (bottom right, black) to 0.8 (top right, red). The simulation combination presented represents a level change of 2 and slope change of 0.1; however, other combinations give similar results.

Supplementary 1.5 Comparison between empirical and model-based standard errors: slope change

The ratio of model to empirical-based standard errors are shown in Figure S51. The accompanying text is in section 5.2.2.

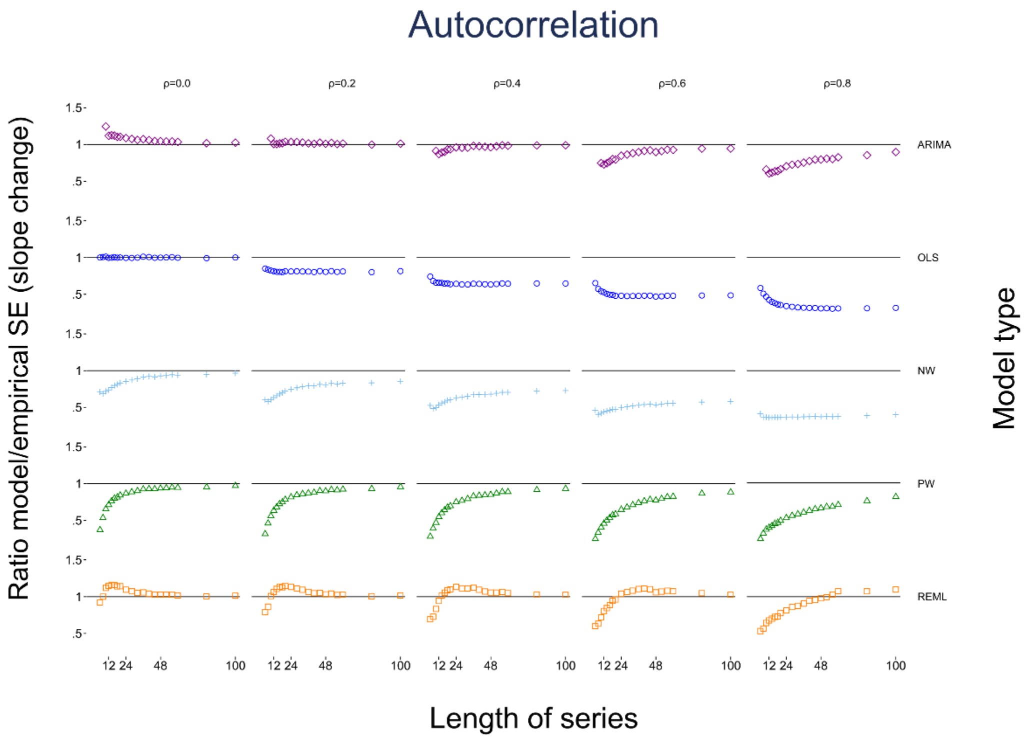


Figure S53: Scatter plots of the ratio of model-based standard error (SE) to the empirical SE for the slope change parameter with different levels of autocorrelation and series length. The horizontal axis represents the number of points in the time series, the vertical axis shows the ratio of model-based to empirical SE. The five vertical columns display the results for different values of autocorrelation. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. The first two series lengths are not shown for the ARIMA method due to extreme values. The Satterthwaite adjustment to the REML does not impact the estimate of SE, hence details of this method are not shown. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood.

Supplementary 1.6 Power

The following graphs show the estimated power for additional level and slope change parameters. The results presented below should be viewed as approximate power only and will generally be lower than the value observed if coverage was at least 95%.

S 1.6.1 Level change

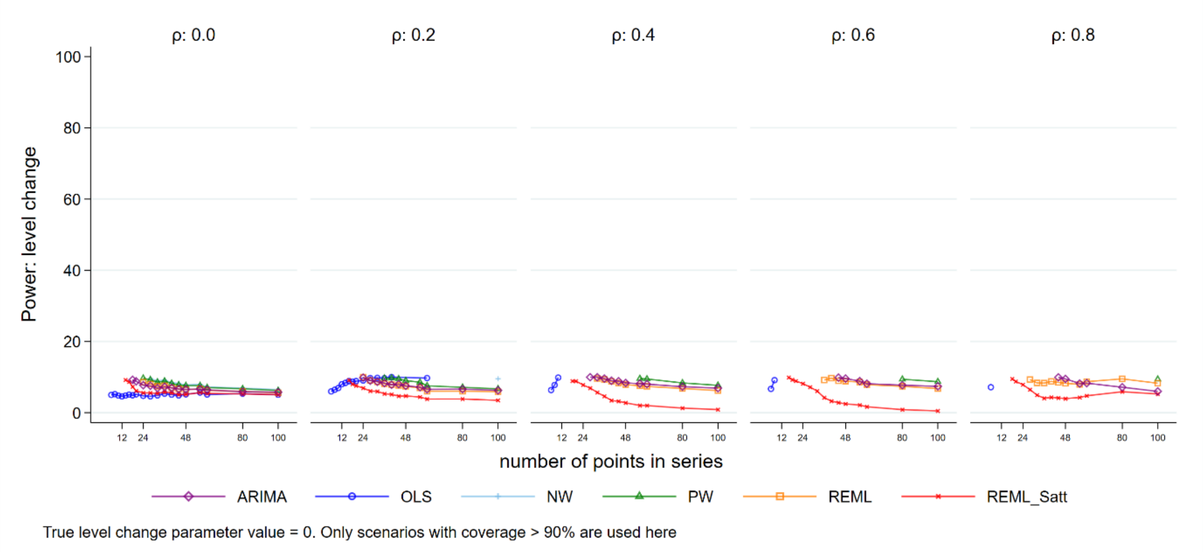


Figure S54: Power for level change, true value 0. Each point is the mean number of times the 95% confidence interval of the estimate did not include zero from 10,000 simulations. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite; NW, Newey-West.

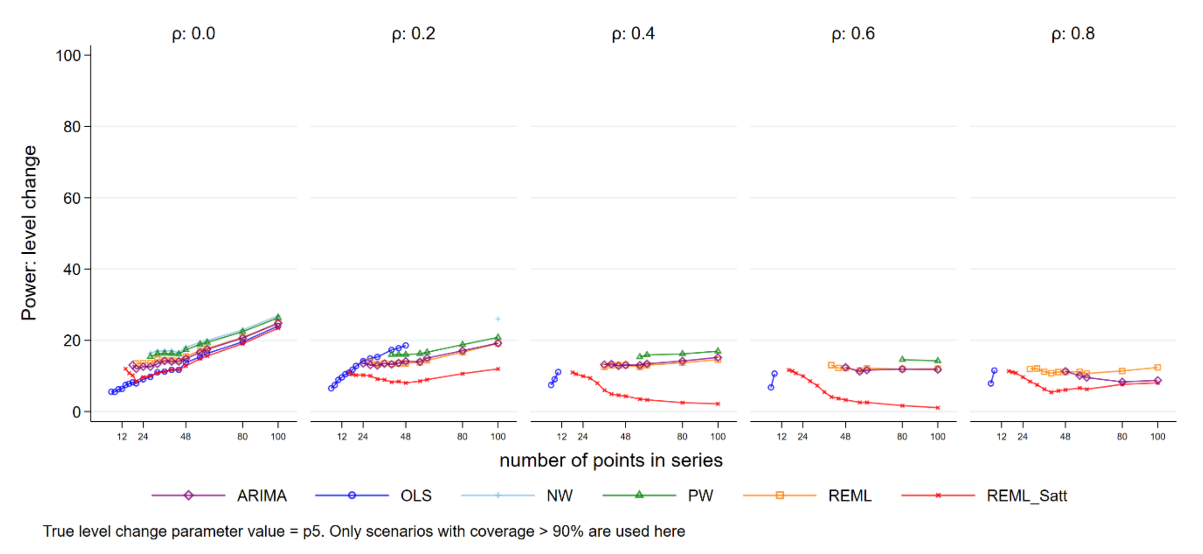


Figure S55: Power for level change, true value .5. Each point is the mean number of times the 95% confidence interval of the estimate did not include zero from 10,000 simulations. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite; NW, Newey-West.

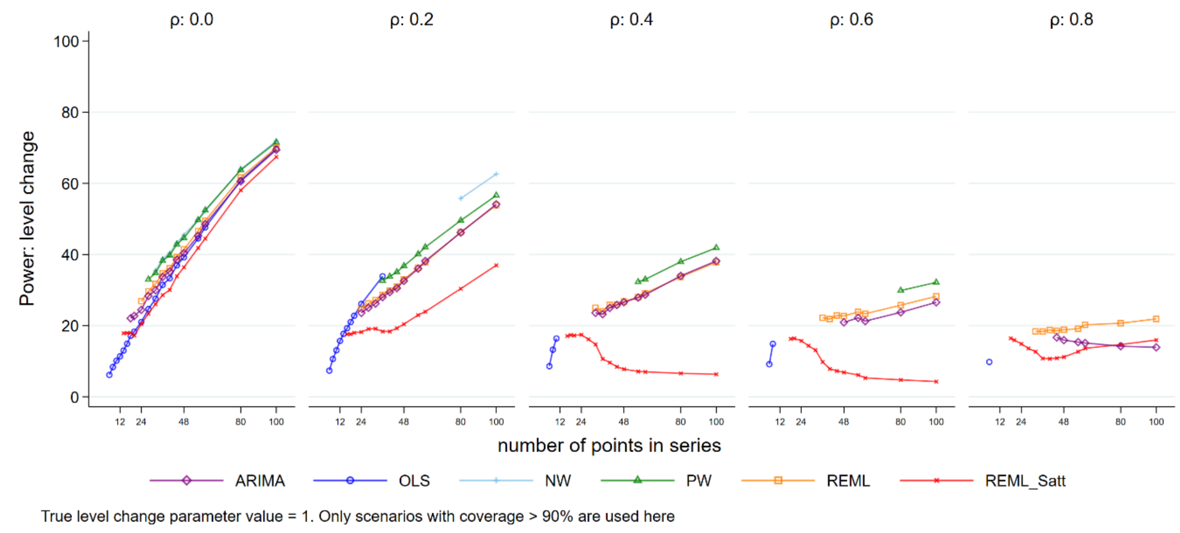


Figure S56: Power for level change, true value 1. Each point is the mean number of times the 95% confidence interval of the estimate did not include zero from 10,000 simulations. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite; NW, Newey-West.

S 1.6.2 Slope change

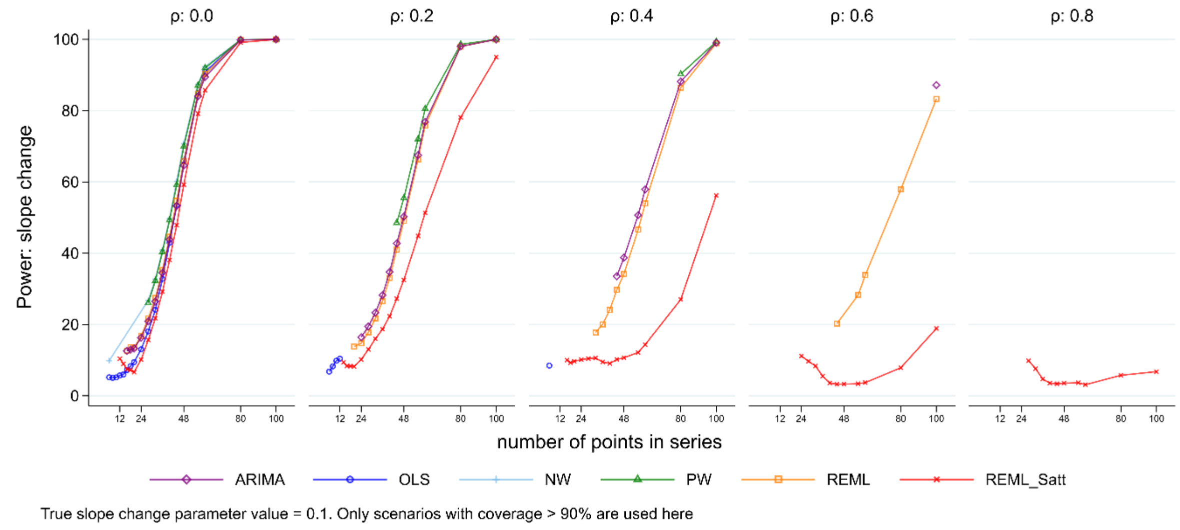


Figure S57: Power for slope change, true value 0.1. Each point is the mean number of times the 95% confidence interval of the estimate did not include zero from 10,000 simulations. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite; NW, Newey-West.

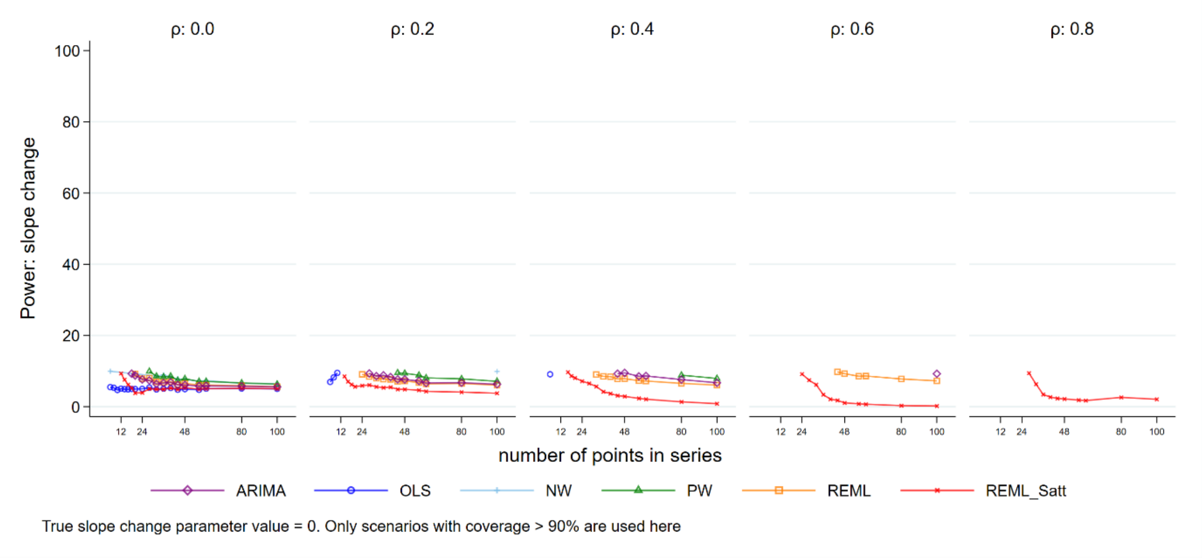


Figure S58: Power for slope change, true value 0. Each point is the mean number of times the 95% confidence interval of the estimate did not include zero from 10,000 simulations. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite; NW, Newey-West.

Supplementary 1.7 Standard error of autocorrelation coefficient estimates.

The following graph shows the empirical SE of the estimates of the magnitude of autocorrelation.

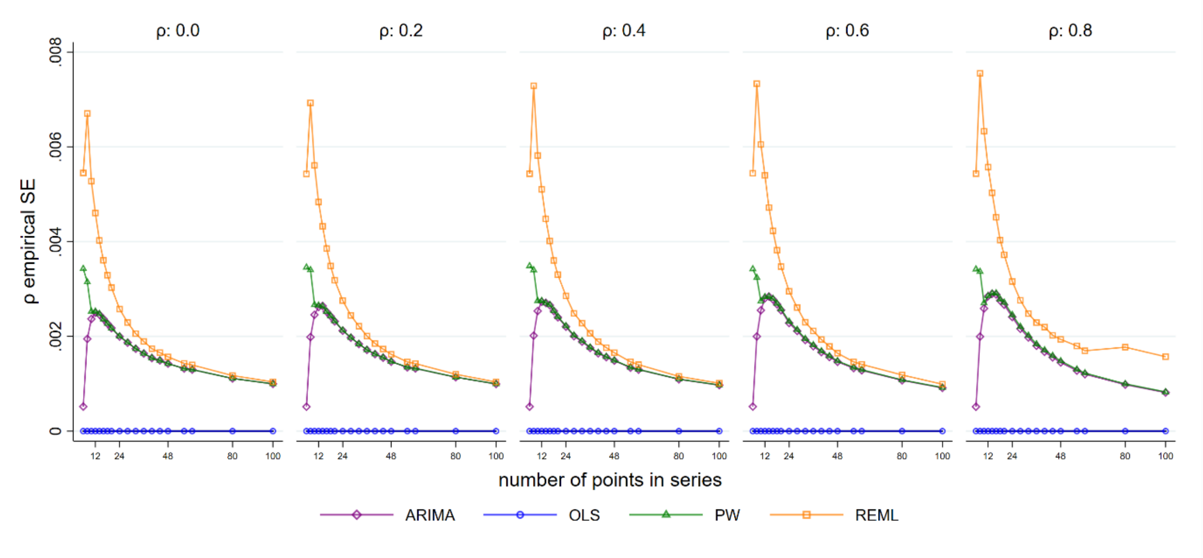


Figure S59: Empirical standard error (SE) for autocorrelation coefficient estimates. The horizontal axis shows the length of the time series. The vertical axis shows the empirical SE of the autocorrelation coefficient estimates. The five plots display the results for different values of autocorrelation ranging from 0 to 0.8. Each coloured point shows the mean value of the SE of the autocorrelation coefficient estimates from 10,000 simulations for a given combination of autocorrelation coefficient and number of points in the data series The simulation combinations presented represent a model structure with a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; PW, Prais-Winsten; REML, restricted maximum likelihood.

Supplementary 1.8 Convergence of estimation methods

Figure S54 shows the number of times the estimation methods converged. Accompanying text is in Section 5.6

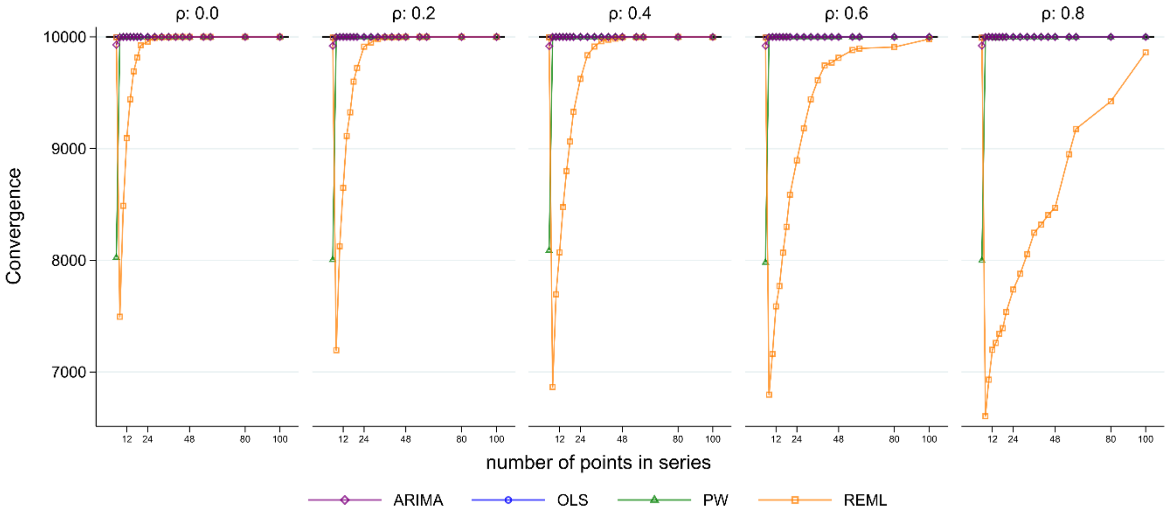


Figure S60: Model convergence. The horizontal axis shows the length of the time series. The vertical axis shows the number of times each method converged out of 10,000 simulation replications. The five plots display the results for different values of autocorrelation ranging from 0 to 0.8. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; PW, Prais-Winsten; REML, restricted maximum likelihood.

Supplementary 1.9 Coverage by autocorrelation bias: slope change

Figure S60 shows the coverage for slope change versus the bias in autocorrelation estimate. Accompanying text is in section 7.1.

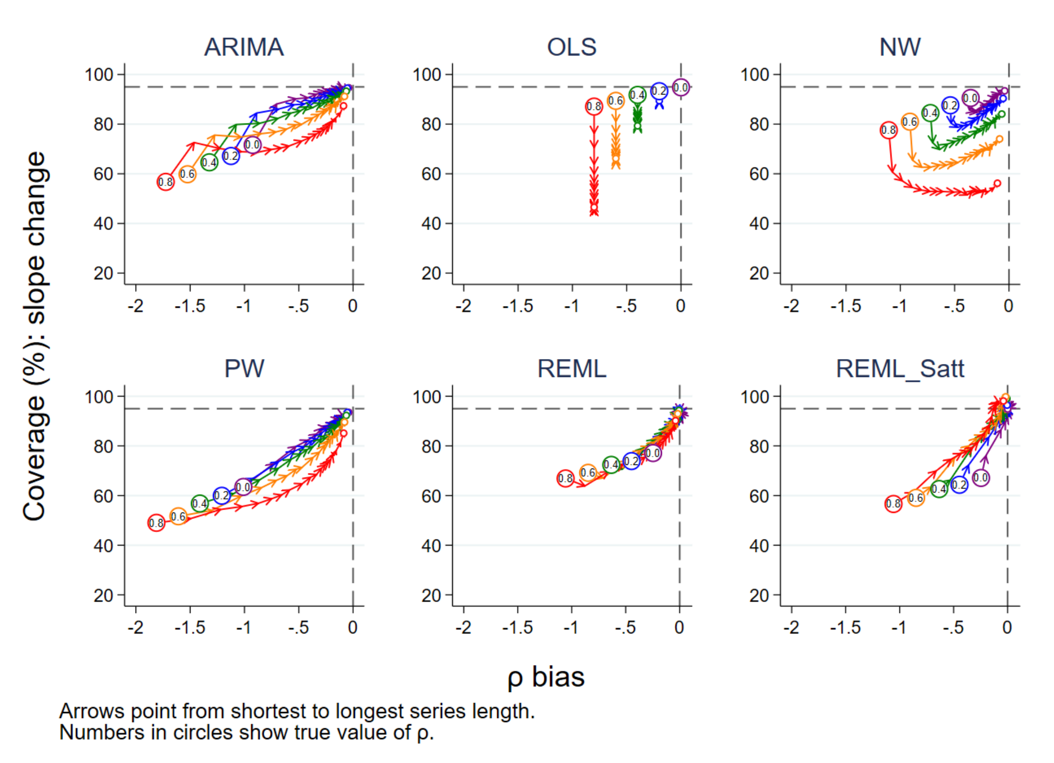


Figure S61: Bias in autocorrelation estimate versus coverage for slope change. The horizontal axis shows the bias in the autocorrelation estimate. The vertical axis shows the percentage coverage. The horizontal dashed line indicates 95% coverage, the vertical dashed line indicates no bias in the estimate of autocorrelation. Each colour represents a different value of underlying autocorrelation, ranging from zero (purple) to 0.8 (red), with each value displayed in a circle at the smallest series length (six points). The arrows point from shortest to longest series length, with the small circles at the end of each line showing coverage at a series length of 100 data points. Each data point shows the mean value from 10,000 simulations for a given combination of autocorrelation coefficient and number of points in the series. The simulation combination presented is for a level change of 2 and slope change of 0.1; however, other combinations give similar results. Abbreviations: ARIMA, autoregressive integrated moving average; OLS, ordinary least squares; NW, Newey-West; PW, Prais-Winsten; REML, restricted maximum likelihood; Satt, Satterthwaite.

Appendix G. Supplementary file 2 accompanying Chapter 5 – Computer code to create and analyse simulated data sets

“Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study”
Turner SL, Forbes AB, Karahalios A, Taljaard M, McKenzie JE.

The following sections contain the Stata 15 computer code used to create the simulated data and analyse it. Stata 15 do file versions can be found on the online repository figshare along with the data required to replicate the study and create the graphs in the publication and supplementary files: https://doi.org/10.26180/13284329 (57).

G.1 Data file 1: 01_S_Turner_ITS_Simulation_Master_File.do

```
////////////////////////////////////////////////////////////////////////////////
// The following do file provides the code to generate the simulations used in the paper:
// Turner, et al. "Evaluation of statistical methods used in the analysis of interrupted time
series studies: a simulation study".
//
// Master simulation do file
// Author: Simon Turner
// simon.turner@monash.edu
//
////////////////////////////////////////////////////////////////////////////////
// This code was tested using Stata IC version 15.1
// set Stata version
version 15.1
//
// This Master do file calls multiple other do files in order to create simulated data sets,
analyse them and summarise the results
// They should all be placed in the same directory.
//
//       The following do files are called:
//
//       to create the simulated data sets:
//       simulate_ITS_data.do
//
//       to analyse the datasets:
//       analyse_ITS_data.do
//
//       ARIMA code in Stata 15 has difficulties for large numbers of consecutive analyses
//      so the ARIMA analysis is performed in batches requiring the following do file:
//       get_simulations_yet_to_do.do
//
//       to append the results from the different methods together:
//       do append_results_methods.do
//
//       to append the results from the different autocorrelations together:
//       do append_results_rhos.do
//
//      finally, simsum (details below) is used to summarise the results and uses
//      simsum_final.do
//
//      Other files that are used for the remainder of the study include:
//      Durbin_Watson_tests.do for the durbin watson component of the study
//       And for the graphs:
//
//       Figure 3 gives parameter distributions for level change:
//      Figure_3_parameter_distributions.do
//       Figure 4 gives line graphs of empirical standard error for level change:
//      Figures_45789.do
//       Figure 5 gives line graphs of empirical standard error for level change:
//      Figures_45789.do
//       Figure 6 gives scatter plots of ratio of model/empirical SE vs series length for level
change
//      Figure_6.do
//       Figure 7 gives scatter plots of coverage for level change
//      Figures_45789.do
//       Figure 8 gives scatter plots of coverage for slope change
//      Figures_45789.do
//      Figure 9 gives line graphs of power for level change
//      Figures_45789.do
//      Figure 10 gives distributions of autocorrelation estimates
//      Figure_10.do
//       Figure 11 shows line graphs of autocorrelation estimates
//      Figure_11.do
//      Figure 12 shows Durbin-Watson tests by series length
//      Figure_12.do
//      Figure 13 shows bias in autocorrelation estimate versus coverage for level change
//      Figure_13.do
//      Supplementary 1.3 has all of the nested loop plots, these need some more
//      work to make them generalisable, currently they run separately
//      Supp_1_3_1_nested_loop_level_bias.do
//      Supp_1_3_2_nested_loop_slope_bias.do
//      Supp_1_3_3_nested_loop_level_empse.do
//      Supp_1_3_4_nested_loop_slope_empse.do
//      Supp_1_3_5_nested_loop_level_modelse.do
//      Supp_1_3_6_nested_loop_slope_modelse.do
//      Supp_1_3_7_nested_loop_level_coverage.do
//      Supp_1_3_8_nested_loop_slope_coverage.do
//      Supp_1_3_9_nested_loop_rho_estimate.do
//       Supplementary_1_5 gives scatter plots of ratio of model/empirical SE vs series length for
slope change
//      Supp_1_5.do
//      Supp 1.6 gives line graphs of power for slope change
//      Figures_45789.do
//      Supp 1.7 gives line graphs of rho empirical standard error
//      Supp_1_7.do
//      Supplementary 1.8 shows convergence
//      Figures_45789.do
//      Supp 1.9 shows bias in autocorrelation estimate versus coverage for slope change
//      Supp_1_9.do

////////////////////////////////////////////////////////////////////////////////
//
// Three further user written programs are required for this project, parallel and simsum for
simulations and analysis:
// grc1leg2 for the graphs
//
// the parallel user written code is required to run this version
// allowing a multi-core computer with the basic version of Stata to run
// much, much faster:
// parallel Stata module for parallel computing
// vers 1.19.0 26jul2017
// auth George G. Vega [cre,aut], Brian Quistorff [aut]

*ssc install parallel // remove the * to install parallel
//
// for the parallel software to work, set the number of clusters to the number of processor cores
on your computer
// it will then run an instance of Stata for each core
// the number (on my computer 6) can be changed to the number of cores on your computer (just
choose 1 if you're not sure)
parallel setclusters 5
//
// the simsum user written code is required to summarise the results
// and provide MCSE:
// version 0.17.1   Ian White   23nov2015
*ssc install simsum // remove the * to install simsum
//
// the grc1leg2 user written code is just to add a legend to the bottom of a graph combine.
// for the high quality images in the pape, normal combine was used and a legend was added using a
graphics package
//
*net install grc1leg2.pkg
////////////////////////////////////////////////////////////////////////////////
// A base directory needs to be supplied
// (the code creates any subdirectories for data files, graphs etc.)
// The other do files (listed above) should be placed in this directory also.
global base_directory `" "<set directory>" "'
//
// For reproducibility the random seed can be specified here:
set seed 150375
//
```

```
////////////////////////////
// if the program has crashed, or if you are only running one method at a time
// you will need to continue the random numbers from after the last one used
// to get seed back, you just need to know which simulation number to recover
// and then reset using
// local mystate = s1[num] + s2[num] + s3[num]
// set rngstate `mystate'
// put the last run simulations seedfile here filling in your last run details for b2, b3 and rho:
// e.g. if b2=0.5, b3=0.1 and rho=0.8: use
"`base_directory'\data_files\b2_p5_b3_p1\rho_8\simulations\seedfile.dta"
/*
use "`base_directory'\data_files\b2_p5_b3_p1\rho_8\simulations\seedfile.dta"
local mystate = s1[_N] + s2[_N] + s3[_N]
display "`mystate'"
set rngstate `mystate'
*/

////////////////////////////////////////////////////////////////////////////////
// Set up main parameters

   // how many simulations to run (this study used 10000)
global num_simulations = 100
   // simulation offset (drop this many observations to ensure independence from one series to the
next)
global simulation_offset = 300 // this study used 300
   // how many time points in the series
global num_points = 100 // this study used 100
   // the intervention time point
global intervention_time = int((${num_points}+1)/2) // halfway point is int((${num_points}+1)/2)
   // range of points to plot graphs, calculate results etc.
global min_num_points = 3 // this is the number of points per segment, any smaller than 3 is not an
interrupted time series by the EPOC definition
global max_num_points = $num_points
   // rather than create every single option, this can be restricted to looking at different
   // numbers of points pre-interruption
   // currently this is set to look at all of the points from 3 to 10, then every second point
   // up to 30, then every 10 after that. This keeps the resolution high for small numbers
   // but skips a lot of the longer time series where the results are stable anyway
global list_of_numbers "${min_num_points}(1)9 10(2)29 30(10)${intervention_time} "

// Model used is observation = beta0
//                        + beta_1*time
//                        + beta_2*Intervention
//                        + beta_3*(time - intervention_time)*intervention
//                        + error
//
// error is modelled in two parts
//       basic error of form N(0,sigma^2)
//       + autocorrelated error of form `rho'*error[_n-1] ie. a term based on the previous time
point error
// Huitema and McKean (2007). "Identifying Autocorrelation Generated by Various Error Processes in
Interrupted Time-Series Regression Designs." Educational and Psychological Measurement 67(3): 447-
459.

// beta_0 = intercept pre-intervention
global beta_0 = 0
// beta_1 = slope during pre-intervention period
global beta_1 = 0
// level changes, use p5 for 0.5
global b2_list "0 p5 1 2"
*global b2_list "2"
// slope changes, use p1 for 0.1
global b3_list "0 p1"
*global b3_list "p1"
// rho values, skip the decimal e.g. 0.2 -> 2
global rho_list "0 2 4 6 8"
// sigma = standard deviation of the error term of form N(0,sigma^2)
global sigma = 1.0
   // enter a text list of the models that you'd like to analyse
   // current options that are enabled:
   // model_type "regress" for OLS regression
   // model_type "newey" for Newey West with lag-1 autocorrelation
   // model_type "prais_raw" for basic Prais-Winsten
   // model_type "prais" for Prais-Winsten with optimal search
   // model_type "corc" for Cochrane-Orcutt
   // model_type "mixed" for REML, lag-1
   // model_type "mixed_satt" for REML with Satterthwaite approximation
   // model_type "arima" for ARIMA
   // new types can be added to the "analyse_ITS_data.do" file
```

```
global model_types "regress newey prais_raw mixed mixed_satt arima"

// for graphing, can use the same list as above (graphs for every set of parameters)
// use "${b2_list}" and ${b3_list}
// or choose to graph just one set e.g. "2" and "p1".
global b2_graph_list = "${b2_list}"
global b3_graph_list = "${b3_list}"

// this ends the section requiring any inputs
********************************************************************************
********************************************************************************
////////////////////////////////////////////////////////////////////////////////
// The structure of the outputs are as follows:
//
// A data_files directory will be created in the base folder
// within this directory will be a subdirectory for each level and slope change combination
// e.g. "b2_0_b3_0" for the zero cases for level and slope change
// a file "all_results_b2_x_b3_x.dta" holds the data for all of the analyses within that
combination
//
// Within each of these level and slope change combinations are subdirectories
// for the autocorrelation parameters "rho_0", "rho_2" and so on
// Within each of these subdirectories "simulations" for the simulations
// these include the "all_simulations.dta" file which includes all the simulations
// for that combination of level, slope, rho as well as "seedfile.dta" containing the random seeds
// Also within the autocorrelation parameter subdirectory is the "results" subdirectory
// Within "results" are separate folders for the results of each method type, e.g. arima,
regress...
// Within the "results" folder are combined datasets for each method and also
// "all_results.dta" holding all of the results for that combination of level, slope and rho
//
// In the data_files directory will be a subdirectory called "simsum"
// within this directory is the "all_simsum_long.dta" file which is the summary
// results of all of the analyses.
// There are also separate simsum files for each parameter combination and beta coefficient
//
// The Durbin-Watson analysis will use the first parameter combination and requires
// the simulation files as it re-runs regress and obtains the DW test parameters
// as such, the first parameter combination folder in the data_files directory
// will have a Durbin_Watson folder created with all the files for that analysis
// the final data required for the graph, "collapsed_dwatson.dta" will be stored
// in the data_file root directory.
//
// At the end of all the coding, the important files for analysing and graphing
// are placed in the root data_files directory. These are:
// "all_results.dta" which holds all of the analysed results and is a very big filemap
// "all_simsum_long.dta" which holds the simsum summary analysis results
// "collapsed_dwatson.dta" which hold the durbin watson summary results
//
// filemap of important files...
// .data_files
//      "all_results.dta"
//      "all_simsum_long.dta"
//      "collapsed_dwatson.dta"
//      -> b2_x_b3_x
//         "all_results_b2_x_b3_x.dta"
//         -> Durbin-Watson
//            "dwatson_bounds_n.dta"
//            "dwatson_checks.dta"
//            "rho_tests_n_points.dta"
//         -> rho_x
//            -> results
//               "results_method_x.dta"
//               "all_results_dta"
//               -> method_x
//                  "results_method_x_n.dta"
//            -> simulations
//               "all_simulations.dta"
//               "seedfile.dta"
//               "simulation_n_points.dta"
//      -> simsum
//         "simsum_b2_x_b3_x_beta_y.dta"
//         "simsum_b2_x_b3_x_beta_y_long.dta"
//
// .graphs
//      "all of the graphs required for the paper"
********************************************************************************
********************************************************************************
```

```
/////////////////////////////////////////////////////////////////////////////////
// The remainder of the do file carries out the simulations, analyses and summaries

/////////////////////////////////////////////////////////////////////////////////
// reset and clear memory
set more off
graph drop _all
matrix drop _all
log close _all
clear

/////////////////////////////////////////////////////////////////////////////////
// As simulations can be lengthy, timers are used to gauge where any long times
// are occurring, expect to see longer times for more complex analyses (e.g. ARIMA, REML)
// set up some timers for comparison of methods etc.
timer clear
timer clear 1
timer on 1

/////////////////////////////////////////////////////////////////////////////////
// Set up all the directories
// make the base directory if it doesn't already exist
cap mkdir $base_directory
cd $base_directory
   // create a directory to hold data files
global data_directory ".\data_files\\"
cap mkdir $data_directory
global graph_directory ".\graphs\\"
cap mkdir $graph_directory

////////////////////////////
// now loop over each level change and slope change combination

// create number list from rho list above
local rho_nums ""
foreach rho of global rho_list {
   local rho_nums = "`rho_nums'" + " 0.`rho'"
}

foreach b2 of global b2_list { // loop b2s
   foreach b3 of global b3_list { // loop b3s

      // beta_2 = immediate change at time of intervention
      global beta_2 = `=subinstr("`b2'","p",".",.)'
      global beta_3 = `=subinstr("`b3'","p",".",.)'

      display "beta_2 = $beta_2 and beta_3 = $beta_3"

      /////////////////////////////////////////////////////////////////////////////////
      // loop over values of autocorrelation

      foreach rho_num of numlist `rho_nums' {

         // rho = autocorrelation, lag 1, parameter between -1 and 1
         global rho = `rho_num'
            // to make a matching directory string need to remove the decimal place for the folder
name
            global rho_name = string(${rho})

            local decimal_place = strpos("${rho}",".")
            local no_decimal = substr("${rho}",`=`decimal_place'+1',.)

            global rho_name = `no_decimal'

            // make a new folder for this set of simulations
         global new_subfolder_name "b2_`b2'_b3_`b3'"
         cap mkdir ${data_directory}\${new_subfolder_name}

         global new_folder_name "${data_directory}\${new_subfolder_name}\rho_${rho_name}\"
         cap mkdir ${new_folder_name}

            // where the graph files are
         global graphs_directory "${new_folder_name}\graphs\\"
         cap mkdir ${graphs_directory}

            // where the simulation files are
         global simulation_directory   "${new_folder_name}\simulations\\"
         cap mkdir ${simulation_directory}
            // where the results files are
         global results_directory "${new_folder_name}\results\\"
         cap mkdir ${results_directory}

            // where the analysis files are
         global analysis_directory "${new_folder_name}\analysis\\"
         cap mkdir ${analysis_directory}

         ////////////////////////////////////////////////////////////////////////////////
         // Generate the simulations

         timer clear 2
         timer on 2
         display "simulations in rho ${rho} $S_TIME  $S_DATE"
               // run the simulations:
               // this do file creates the simulation data sets
               // it also creates a list of the random seeds used so that
               // any simulation can be re-created
               // and if the program crashes, the simulations can be re-run
               // from the point of crash using the appropriate random numbers
            do simulate_ITS_data.do

         timer off 2
         timer list 2

         ////////////////////////////////////////////////////////////////////////////////
         // Run models
         // the analyse_ITS_data.do file analyses the simulated datasets
         // and outputs the results
         // The following section is a series of "if" statements that runs through
         // each selected model in the "model_types" string input earlier
         // It then calls the do file, passing it the appropriate analysis type.
         // Timers are included as it can be valuable to see which methods are particularly slow
         // There is some extra code around the ARIMA method as we found that there were
         // difficulties running too many consecutive analyses and so have broken the
         // analysis up into batches for ARIMA. It's slower, but crashes less frequently.
         // The default value for model is standard regression if nothing was input
         //

         local store_model_types = "${model_types}"

         foreach model_type of global model_types {

            if "`model_type'" == "regress" {
               global model_types "regress"
               display "${model_types} in rho ${rho} $S_TIME  $S_DATE"
               timer clear 3
               timer on 3
                  use ${simulation_directory}all_simulations.dta, clear
               parallel do analyse_ITS_data.do, by(simulation)

               timer off 3
               timer list 3
            }
            else if "`model_type'" == "newey" {
               global model_types "newey"
               // rho obtained by regressing residual = B*residual[_n-1]
               display "${model_types} in rho ${rho} $S_TIME  $S_DATE"
               timer clear 4
               timer on 4
                  use ${simulation_directory}all_simulations.dta, clear
               parallel do analyse_ITS_data.do, by(simulation)
               timer off 4
               timer list 4
            }
            else if "`model_type'" == "prais_raw" {
               global model_types "prais_raw"
               display "${model_types} in rho ${rho} $S_TIME  $S_DATE"
               timer clear 5
               timer on 5
                  use ${simulation_directory}all_simulations.dta, clear
               parallel do analyse_ITS_data.do, by(simulation)
               timer off 5
               timer list 5
            }
            else if "`model_type'" == "prais" {
            global model_types "prais"
               // currently using ssesearch for minimising sum of squared errors searching for rho
               display "${model_types} in rho ${rho} $S_TIME  $S_DATE"
```

```
              timer clear 5
              timer on 5
                 use ${simulation_directory}all_simulations.dta, clear
              parallel do analyse_ITS_data.do, by(simulation)
              timer off 5
              timer list 5
           }
           else if "`model_type'" == "corc" {
              global model_types "corc"
              display "${model_types} in rho ${rho} $S_TIME  $S_DATE"
              timer clear 6
              timer on 6
                 use ${simulation_directory}all_simulations.dta, clear
              parallel do analyse_ITS_data.do, by(simulation)
              timer off 6
              timer list 6
           }
           else if "`model_type'" == "mixed" {
              global model_types "mixed"
              display "${model_types} in rho ${rho} $S_TIME  $S_DATE"
              timer clear 7
              timer on 7
                 use ${simulation_directory}all_simulations.dta, clear
              parallel do analyse_ITS_data.do, by(simulation)
              timer off 7
              timer list 7
           }
           else if "`model_type'" == "mixed_satt" {
              global model_types "mixed_satt"
              display "${model_types} in rho ${rho} $S_TIME  $S_DATE"
              timer clear 8
              timer on 8
                 use ${simulation_directory}all_simulations.dta, clear
              parallel do analyse_ITS_data.do, by(simulation)
              timer off 8
              timer list 8
           }
           else if "`model_type'" == "arima" {
              // ARIMA model seems to crash or hang after a while, so instead
              // of analysing all of the data at once, we do it in chunks
              // but because we're running "parallel"
              // we can't always tell exactly which simulations we've already
              // analysed, so we have to find that out first, then analyse the
              // next batch after that

              global model_types "arima"
              display "${model_types} in rho ${rho} $S_TIME  $S_DATE"
              timer clear 9
              timer on 9

              local arima_directory "${results_directory}\arima\"
              cap mkdir `arima_directory'

              // here we set how many batches to do things in
              local previous_num = 0
              // if we have small numbers of simulations, do in batches of 10, otherwise in
batches of 1000
              // this may need fine tuning for different computers
              if ${num_simulations} < 1000 {
                 local block_value = 10
              }
              else {
                 local block_value = 1000
              }
              forvalues current_num = `block_value'(`block_value')${num_simulations} {

                 // this do file works out which simulations we still need to do
                 do get_simulations_yet_to_do.do
                 // we then keep the simulations we have yet to do within the batch
                 keep if simulation > `previous_num' & simulation <= `current_num'

                 quietly summ simulation
                 local num_remaining = r(N)

                 if `num_remaining' != 0 {

                    display "`previous_num' to `current_num'"
                    sort simulation
                    // and finally analyse using ARIMA
```

```
                     parallel do analyse_ITS_data.do, by(simulation)

                  } // end 0 check

                  local previous_num = `current_num'

               } // end batch loop

               timer off 9
               timer list 9
            }
            // set the full list of model types back to what it should be
            global model_types = "`store_model_types'"

         } // end model type loop

         ////////////////////////////////////////////////////////////////////////////
         // combine all the results
         display "Results appending in rho ${rho} $S_TIME  $S_DATE"
         // append model results into master files
         display "Beginning model results appending"

            // within each value of autocorrelation, append all the methods
            do append_results_methods.do

      } // end rho loop
      // finally append all of the autocorrelations also
      global b2 = "`b2'"
      global b3 = "`b3'"
      do append_results_rhos.do

      /////////////////////////////////////////////////////////////////////////
      // make the satterthwaite small sample adjustment for unusually small degrees of freedom
      // doing it here in the code means that you could see the results otherwise by just
commenting this out
      use "${data_directory}\${new_subfolder_name}\all_results_b2_${b2}_b3_${b3}.dta", clear
      // we have chosen a minimum cut off of 2
      gen use_dof = 2
      // do this for each parameter
      forvalues beta_num = 0/3 {
         // back up the originals just in case we want to look at them later
         gen original_`beta_num'_cil = beta_`beta_num'_cil
         gen original_`beta_num'_ciu = beta_`beta_num'_ciu
         // create a new version that uses a dof of 2 if the dof < 2
         gen new_`beta_num'_cil = .
         replace new_`beta_num'_cil = beta_`beta_num'_cil
         replace new_`beta_num'_cil = beta_`beta_num' - invttail(use_dof,0.025)*beta_`beta_num'_se
if beta_`beta_num'_dof < 2 & model_type == "mixed_satt"
         gen new_`beta_num'_ciu = .
         replace new_`beta_num'_ciu = beta_`beta_num'_ciu
         replace new_`beta_num'_ciu = beta_`beta_num' + invttail(use_dof,0.025)*beta_`beta_num'_se
if beta_`beta_num'_dof < 2 & model_type == "mixed_satt"
         // replace the original with the new
         replace beta_`beta_num'_cil = new_`beta_num'_cil
         replace beta_`beta_num'_ciu = new_`beta_num'_ciu
         // back up the original dofs too
         gen original_`beta_num'_dof = beta_`beta_num'_dof
         replace beta_`beta_num'_dof = 2 if beta_`beta_num'_dof < 2 & model_type == "mixed_satt"
      } // end satterthwaite fix loop
      save "${data_directory}\${new_subfolder_name}\all_results_b2_${b2}_b3_${b3}.dta", replace
   } // end b3 loop
} // end b2 loop

// now create one master data file
local first = 1
foreach b2 of global b2_list { // loop b2s
   foreach b3 of global b3_list { // loop b3s
      if `first' == 1 {
         use "${data_directory}\b2_`b2'_b3_`b3'\all_results_b2_`b2'_b3_`b3'.dta", clear
         local first = 0
      }
      else {
         append using "${data_directory}\b2_`b2'_b3_`b3'\all_results_b2_`b2'_b3_`b3'.dta", force
      }
   } // end b3 loop
} // end b2 loop

save "${data_directory}\\all_results.dta", replace
```

```
// finally run the simsum summary statistics
// as mentioned above, the simsum package will need to be installed for this
do simsum_final.do

timer off 1
// list timers
timer list

////////////////////////////////////////////////////////////////////////////////
// Durbin-Watson section
do Durbin_Watson_tests.do


////////////////////////////////////////////////////////////////////////////////
// graphs

// Figure 3 gives parameter distributions for level change
do Figure_3_parameter_distributions.do

// Figure 4 gives line graphs of empirical standard error for level change
global figure "Figure_4"
do Figures_45789.do

// Figure 5 gives line graphs of empirical standard error for level change
global figure "Figure_5"
do Figures_45789.do

// Figure 6 gives scatter plots of ratio of model/empirical SE vs series length for level change
do Figure_6.do

// Figure 7 gives scatter plots of coverage for level change
global figure "Figure_7"
do Figures_45789.do

// Figure 8 gives scatter plots of coverage for slope change
global figure "Figure_8"
do Figures_45789.do

// Figure 9 gives line graphs of power for level change
global figure "Figure_9"
do Figures_45789.do

// Figure 10 gives distributions of autocorrelation estimates
do Figure_10.do

// Figure 11 shows line graphs of autocorrelation estimates
do Figure_11.do

// Figure 12 shows Durbin-Watson tests by series length
do Figure_12.do

// Figure 13 shows bias in autocorrelation estimate versus coverage for level change
do Figure_13.do

do Supp_1_3_1_nested_loop_level_bias.do
do Supp_1_3_2_nested_loop_slope_bias.do
do Supp_1_3_3_nested_loop_level_empse.do
do Supp_1_3_4_nested_loop_slope_empse.do
do Supp_1_3_5_nested_loop_level_modelse.do
do Supp_1_3_6_nested_loop_slope_modelse.do
do Supp_1_3_7_nested_loop_level_coverage.do
do Supp_1_3_8_nested_loop_slope_coverage.do
do Supp_1_3_9_nested_loop_rho_estimate.do

// Supplementary_1_5 gives scatter plots of ratio of model/empirical SE vs series length for slope
change
do Supp_1_5.do

// Supp 1.6 gives line graphs of power for slope change
global figure "Supp_1_6"
do Figures_45789.do

// Supp 1.7 gives line graphs of rho empirical standard error
do Supp_1_7.do

// Supplementary 1.8 shows convergence
global figure "Supp_1_8_1"
do Figures_45789.do
```

```
// Supp 1.9 shows bias in autocorrelation estimate versus coverage for slope change
do Supp_1_9.do
```

G.2 Data file 2: simulation_ITS_data.do

```
////////////////////////////////////////////////////////////////////////////////
// Simulation final file
// Simon Turner
//
// This program creates simulated ITS datasets
//
////////////////////////////////////////////////////////////////////////////////
//
// This is called from the Master file which supplies the parameters
//
////////////////////////////////////////////////////////////////////////////////
// set Stata version
version 15.1

///////////////////////////////
// set up the random seed file, keeps a track of the state of the random numbers so that each
simulation is reproducible
tempname postseed
postfile `postseed' simulation simulation_pre_num str2000 s1 str2000 s2 str1100 s3 using
${simulation_directory}\\seedfile.dta, replace
// to get seed back, you just need to know which simulation number to recover
// and then reset using
// local mystate = s1[num] + s2[num] + s3[num]
// set rngstate `mystate'
// you can find this code near the start of the Master do file

///////////////////////////////
// loop over each number in the global numlist list_of_numbers
// note that the full dataset is created anyway, as it doesn't take that much longer
// but the simulation_pre_num will be saved so that during analysis different
// numbers of pre_points can then be analysed
foreach simulation_pre_num of numlist $list_of_numbers {

   ////////////////////////////////////////////////////////////////////////////////
   // begin looping over multiple simulations
   // number of simulations from Master file
   forvalues simulation_num = 1/$num_simulations {

      // first post the random seed at this point
      post `postseed' (`simulation_num') ///
                  (`simulation_pre_num') ///
                  (substr(c(rngstate),1,2000)) ///
                  (substr(c(rngstate),2001,2000)) ///
                  (substr(c(rngstate),4001,.))

      display "Running simulation `simulation_num' out of ${num_simulations}"

      quietly {


/////////////////////////////////////////////////////////////////////////////////////////////////
      // generate simulation offset series (a way of ensuring independence from one series to the
next)
      clear
      set obs ${simulation_offset}

      // first order autoregressive model
         gen error_normal = rnormal(0,${sigma}^2) // start with just a normal error
         gen error = 0
         // add autocorrelated part and maintain constant variance by adding this factor to the
first observation
         replace error = sqrt(1/(1-(${rho}^2)))*error_normal in 1
         replace error = ${rho}*error[_n-1] +  error_normal in 2/${simulation_offset}

      ////////////////////////////////////////////////////////////////////////////////
      // begin proper model
      clear

      // create the empty dataset
      set obs ${num_points}

      // keep simulation number to keep track later
      gen simulation = `simulation_num'

      // time is just consecutive numbers
      gen time = _n

      // keep to keep track of
      gen simulation_pre_num = `simulation_pre_num'

      // set indicator variable for intervention defined previously e.g. 100 points, 50 is the
intervention time by this definition so 1-50 and 51-100 for the two segments
      gen intervention = 0
      replace intervention = 1 if time > ${intervention_time}

      ////////////////////////////////////////////////////////////////////////////////
      // first order autoregressive model
         gen error_normal = rnormal(0,${sigma}^2)

         gen error = 0
         // add autocorrelated part and maintain constant variance by using this factor to the
first observation: sqrt(1/(1-rho^2))
         replace error = sqrt(1/(1-(${rho}^2)))*error_normal in 1
         // now propogate the error through the data
         replace error = ${rho}*error[_n-1] +  error_normal in 2/${num_points}
         // and finally create the observation
         gen observation = ${beta_0} + ${beta_1}*time + ${beta_2}*intervention + ${beta_3}*(time -
(${intervention_time}+1))*intervention + error

      ////////////////////////////////////////////////////////////////////////////////
      // extra variables for analysis
      // Huitema and McKean (2007). "Identifying Autocorrelation Generated by Various Error
Processes in Interrupted Time-Series Regression Designs." Educational and Psychological Measurement
67(3): 447-459.

         // generate variables to indicate time of intervention
         gen level_change = intervention
         gen slope_change = (time-(${intervention_time}+1))*level_change
         gen slope_pre = time

      // save the simulated observations as a tempfile for later appending
      tempfile simulation_`simulation_num'
      save "`simulation_`simulation_num''"

      } // end quietly
   } // end loop over num simulations

   // now append tempfiles
   use "`simulation_1'"
   forvalues simulation_num = 2/$num_simulations {
      append using "`simulation_`simulation_num''"
   }
   save ${simulation_directory}\\simulation_`simulation_pre_num'_points.dta, replace

} // end loop over pre-num points

// now append all simulation pre_nums too
local iteration = 1
foreach simulation_pre_num of numlist $list_of_numbers {
   if `iteration' == 1 {
      use ${simulation_directory}\\simulation_`simulation_pre_num'_points.dta, clear
      local iteration = `iteration' + 1
   }
   else {
      append using ${simulation_directory}\\simulation_`simulation_pre_num'_points.dta
   }
}
sort simulation simulation_pre_num time
save ${simulation_directory}\\all_simulations.dta, replace

postclose `postseed'
```

G.3 Data file 3: analyse_ITS_data.do

```
////////////////////////////////////////////////////////////////////////////////
// Analysis final file
// Simon Turner
//
// This program analyses simulated ITS datasets
//
////////////////////////////////////////////////////////////////////////////////
//
// This is called from the Master file which supplies the parameters
//
////////////////////////////////////////////////////////////////////////////////

////////////////////////////////////////////////////////////////////////////////
// the basic estimator to find the autcorrelation coefficient for the Newey West:
cap program drop find_rho
program find_rho

syntax varlist [if]
marksample touse
quietly regress `varlist' l.`varlist' if `touse' , nocons
local rho=_b[L.`varlist']
scalar rho = `rho'

end

cap program drop find_rho_wrap
program find_rho_wrap

syntax varlist [if]

      marksample touse

      tempvar resid
      regress `varlist' if `touse'
      predict double `resid', resid
      find_rho `resid' if `touse'

end

///////////////////////////////////////////////////////////////////////////////
// set lags to 1 if not regress (may wish to look at more on this later)
// actually, now that we're using parallel processing  it's easier to do one
// model at a time, so this loop actually does nothing at present!
foreach model_type of global model_types {

   // set lags to 1 if not regress (may wish to look at more on this later)
   if "`model_type'" == "regress" {
      local lags = 0
   }
   else if "`model_type'" != "regress" {
      local lags = 1
   }

   local results_directory "${results_directory}\`model_type'\"
   cap mkdir `results_directory'

   // perhaps we'll look at multiple intervention times in the future and will need this...
   local min_intervention = ${intervention_time}
   local max_intervention = ${intervention_time}

   ////////////////////////////////////////////////////////////////////////////////
   // analysis


   tempfile all_sim_tempfile

   save `all_sim_tempfile'

   local first_sim = simulation[1]

   /////////////////////////////////////////
   //
         // save values we want
      local matrix_colnames "simulation intervention_time pre_num_points post_num_points"
      local matrix_colnames = "`matrix_colnames'" + " beta_0 beta_0_se beta_0_cil beta_0_ciu
beta_0_p beta_0_dof"
```

```
      local matrix_colnames = "`matrix_colnames'" + " beta_1 beta_1_se beta_1_cil beta_1_ciu
beta_1_p beta_1_dof"
      local matrix_colnames = "`matrix_colnames'" + " beta_2 beta_2_se beta_2_cil beta_2_ciu
beta_2_p beta_2_dof"
      local matrix_colnames = "`matrix_colnames'" + " beta_3 beta_3_se beta_3_cil beta_3_ciu
beta_3_p beta_3_dof"
      local matrix_colnames = "`matrix_colnames'" + " sigma str20 model_type lags beta_0_true
beta_1_true beta_2_true beta_3_true rho_true"
      local matrix_colnames = "`matrix_colnames'" + " converged num_iterations error_code run_time"
      local matrix_colnames = "`matrix_colnames'" + " rho_est rho_est_cil rho_est_ciu"
      local matrix_colnames = "`matrix_colnames'" + " mse rmse"

      // create postfile
      tempname results
      postfile `results' `matrix_colnames' using
`results_directory'results_`model_type'_`first_sim'.dta, replace

   /////////////////////////////////////////
   // To speed up this process as much as possible it's better to index the
   // simulation file and then just open a chunk of it at a time rather than
   // using preserve/restore or other such methods.
   // Significant speed improvements by doing it this way!

   local all = _N

   bysort simulation simulation_pre_num: gen long combinations_count = _N
   local index
   local pointer 1
   while `pointer' <= _N {
      local index `index' `pointer'
      local pointer = `pointer' + combinations_count[`pointer']
   }

   tokenize `"`index'"'

   while "`1'" != "" {
      // this bit uses the created indexes and pointers to open just one bit at a time
      cap use in `1'/`='`2'-1' using `all_sim_tempfile', clear
      if _rc != 0 {
         use in `1'/`all' using `all_sim_tempfile', clear
      }
      local simulation_num = simulation[1]
      local pre_num_points = simulation_pre_num[1]

      display "Analysing simulation `simulation_num' num points `pre_num_points'"

      // currently we only look at one intervention time, but maybe later we'll do something else
      forvalues intervention_time = `min_intervention'/`max_intervention' {
         // currently we're looking at just one set of points, but we may want to look at different
series lengths
         forvalues post_num_points = `pre_num_points'/`pre_num_points' { // set equal to
`pre_num_points' for equal pre- post- points


            ////////////////////
            // analysis
            tsset time

               // set up some defaults
                  local converged = .
                  local rho_est = .
                  local rho_est_cil = .
                  local rho_est_ciu = .
                  local num_iterations = .
                  local mse = .
                  local rmse = .
                  matrix dfs = J(1,6,.)

               timer clear 42
               timer on 42

               ///////////////////////////////////////////////////////////////
               // set up here any model name you wish and add the necessary details

               if "`model_type'" == "regress" {
                  regress observation slope_pre level_change slope_change if (time >
$intervention_time - `pre_num_points') & (time <= $intervention_time + `post_num_points')
                  matrix local_results = r(table)
                  local rho_est = 0
                  local num_iterations = 1
                  local error_code = _rc
                  local converged = 1
                  local rmse = e(rmse)
                  matrix dfs = J(1,6,`pre_num_points'*2 - 4)
               } // end regress check
               else if "`model_type'" == "regress_level" {
                  regress observation level_change if (time > $intervention_time - 
`pre_num_points') & (time <= $intervention_time + `post_num_points')
                  matrix local_results = r(table)
                  local rho_est = 0
                  local num_iterations = 1
                  local error_code = _rc
                  local converged = 1
                  local rmse = e(rmse)
                  matrix dfs = J(1,6,`pre_num_points'*2 - 2)
               } // end regress level check
               else if "`model_type'" == "newey" {
                  newey observation slope_pre level_change slope_change if (time > 
$intervention_time - `pre_num_points') & (time <= $intervention_time + `post_num_points'), 
lag(`lags')
                  matrix local_results = r(table)
                  local varlist = "observation slope_pre level_change slope_change"
                  quietly find_rho_wrap `varlist' if (time > $intervention_time - `pre_num_points') 
& (time <= $intervention_time + `post_num_points')
                  local rho_est = rho
                  local num_iterations = 1
                  local error_code = _rc
                  local converged = 1
                  local rmse = e(rmse)
                  matrix dfs = J(1,6,`pre_num_points'*2 - 4)
               } // end newey check
               else if "`model_type'" == "prais" {
                  cap prais observation slope_pre level_change slope_change if (time > 
$intervention_time - `pre_num_points') & (time <= $intervention_time + `post_num_points'), 
ssesearch
                  if _rc != 0 {
                     matrix local_results = J(6,5,.)
                     local rho_est = .
                     local converged = .
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     matrix local_results = r(table)
                     local rho_est = e(rho)
                     local num_iterations = e(ic)
                     local error_code = _rc
                     local converged = 1
                     local rmse = e(rmse)
                     matrix dfs = J(1,6,`pre_num_points'*2 - 4)
                  }
               } // end prais check
               else if "`model_type'" == "prais_raw" {
                  cap prais observation slope_pre level_change slope_change if (time > 
$intervention_time - `pre_num_points') & (time <= $intervention_time + `post_num_points')
                  if _rc != 0 {
                     matrix local_results = J(6,5,.)
                     local rho_est = .
                     local converged = .
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     matrix local_results = r(table)
                     local rho_est = e(rho)
                     local num_iterations = e(ic)
                     local error_code = _rc
                     local converged = 1
                     local rmse = e(rmse)
                     matrix dfs = J(1,6,`pre_num_points'*2 - 4)
                  }
               } // end prais_raw check
               else if "`model_type'" == "corc" {
                  cap prais observation slope_pre level_change slope_change if (time > 
$intervention_time - `pre_num_points') & (time <= $intervention_time + `post_num_points'), corc 
ssesearch
                  if _rc != 0 {
                     matrix local_results = J(6,5,.)
```

```
                     local rho_est = .
                     local converged = .
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     matrix local_results = r(table)
                     local rho_est = e(rho)
                     local num_iterations = e(ic)
                     local error_code = _rc
                     local converged = 1
                     local rmse = e(rmse)
                     matrix dfs = J(1,6,`pre_num_points'*2 - 4)
                  }
               } // end corc check
               else if "`model_type'" == "mixed" {
                  cap mixed observation slope_pre level_change slope_change if (time > 
$intervention_time - `pre_num_points') & (time <= $intervention_time + `post_num_points'),  res(ar 
1, t(time)) var reml iter(1000)
                  if _rc != 0 {
                     matrix local_results = J(6,5,.)
                     local rho_est = .
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     matrix local_results = r(table)
                     local num_cols = colsof(local_results)
                     local rho_est = tanh(local_results[1,`num_cols'])
                     local rho_est_cil = tanh(local_results[5,`num_cols'])
                     local rho_est_ciu = tanh(local_results[6,`num_cols'])
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc

                     predict double err, res
                     gen double sqerr = err^2
                     su sqerr, meanonly
                     local mse = r(sum)
                     local rmse = sqrt(`mse'/(e(N) - e(df_m) - 1))

                     matrix dfs = J(1,6,`pre_num_points'*2 - 4)
                  }
               } // end mixed check
               else if "`model_type'" == "mixed_kr" {
                  cap mixed observation slope_pre level_change slope_change if (time > 
$intervention_time - `pre_num_points') & (time <= $intervention_time + `post_num_points'),  res(ar 
1, t(time)) var reml iter(1000) dfmethod(kr)
                  if _rc != 0 {
                     matrix local_results = J(6,5,.)
                     local rho_est = .
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     matrix local_results = r(table)
                     local num_cols = colsof(local_results)
                     local rho_est = tanh(local_results[1,`num_cols'])
                     local rho_est_cil = tanh(local_results[5,`num_cols'])
                     local rho_est_ciu = tanh(local_results[6,`num_cols'])
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc

                     mat dfs = e(df)

                     predict double err, res
                     gen double sqerr = err^2
                     su sqerr, meanonly
                     local mse = r(sum)
                     local rmse = sqrt(`mse'/(e(N) - e(df_m) - 1))
                  }
               } // end mixed_kr check
               else if "`model_type'" == "mixed_satt" {
                  cap mixed observation slope_pre level_change slope_change if (time > 
$intervention_time - `pre_num_points') & (time <= $intervention_time + `post_num_points'),  res(ar 
1, t(time)) var reml iter(1000) dfmethod(satt)
```

```
                  if _rc != 0 {
                     matrix local_results = J(6,5,.)
                     local rho_est = .
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     matrix local_results = r(table)
                     local num_cols = colsof(local_results)
                     local rho_est = tanh(local_results[1,`num_cols'])
                     local rho_est_cil = tanh(local_results[5,`num_cols'])
                     local rho_est_ciu = tanh(local_results[6,`num_cols'])
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc

                     mat dfs = e(df)

                     predict double err, res
                     gen double sqerr = err^2
                     su sqerr, meanonly
                     local mse = r(sum)
                     local rmse = sqrt(`mse'/(e(N) - e(df_m) - 1))
                  }
               } // end mixed_satt check
               else if "`model_type'" == "arima" {
                  cap arima observation slope_pre level_change slope_change if (time >
$intervention_time - `pre_num_points') & (time <= $intervention_time + `post_num_points'),  ar(1)
iter(1000)
                  if _rc != 0 {
                  // error catch - sometimes this model mucks up when moving too quickly, so slow
it down a little
                     local iteration = 1
                     while `iteration' < 10 & _rc != 0 {
                        cap matrix local_results = J(6,5,.)
                        local rho_est = .
                        local converged = e(converged)
                        local num_iterations = e(ic)
                        local error_code = _rc
                        sleep 10
                        local iteration = `iteration' + 1
                     }
                  }
                  else {
                     matrix local_results = r(table)
                     local num_cols = colsof(local_results)
                     local rho_col = `num_cols' - 1
                     local rho_est = (local_results[1,`rho_col'])
                     local rho_est_cil = (local_results[5,`rho_col'])
                     local rho_est_ciu = (local_results[6,`rho_col'])
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc

                     predict double err, res
                     gen double sqerr = err^2
                     su sqerr, meanonly
                     local mse = r(sum)
                     local rmse = sqrt(`mse'/(e(N) - e(df_m) - 1))

                     matrix dfs = J(1,6,e(df_m))
                  }
               } // end arima check

               else {
                  display "Model type `model_type' is not recognised"
                  stop
               }

               timer off 42
               quietly timer list 42
               local current_time = r(t42)

               display "Posting simulation `simulation_num' num points `pre_num_points'"

               post `results' (`simulation_num') (`intervention_time') (`pre_num_points')
(`post_num_points') ///
```

```
                                             (local_results[1,4]) (local_results[2,4])
(local_results[5,4]) (local_results[6,4]) (local_results[4,4]) (dfs[1,4]) ///
                                             (local_results[1,1]) (local_results[2,1])
(local_results[5,1]) (local_results[6,1]) (local_results[4,1]) (dfs[1,1]) ///
                                             (local_results[1,2]) (local_results[2,2])
(local_results[5,2]) (local_results[6,2]) (local_results[4,2]) (dfs[1,2]) ///
                                             (local_results[1,3]) (local_results[2,3])
(local_results[5,3]) (local_results[6,3]) (local_results[4,3]) (dfs[1,3]) ///
                                             (${sigma}) ("`model_type'") (`lags') ///
                                             (${beta_0}) (${beta_1}) (${beta_2}) (${beta_3})
(${rho}) ///
                                             (`converged') (`num_iterations') (`error_code')
(`current_time') ///
                                             (`rho_est') (`rho_est_cil') (`rho_est_ciu') ///
                                             (`mse') (`rmse')

         } // end post_num_points loop
      } // end intervention_time loop



      macro shift // move on to next one
   }

   postclose `results'

} // end model type loop
```

G.4 Data file 4: get_simulations_yet_to_do.do

```
////////////////////////////////////////////////////////////////////////////////
// Subprogram to help with ARIMA difficulties
// Simon Turner
//
// This program works out what simulations have already been analysed and
// generates a simulation file containing what's left to do
//
////////////////////////////////////////////////////////////////////////////////
//
// This is called from the Master file which supplies the parameters
//
////////////////////////////////////////////////////////////////////////////////

// append files done so far
local dir "${results_directory}${model_types}\\"
local files: dir "`dir'" files "*.dta"

display `"`files'"'

// append all together
if `"`files'"' == "" {
        use "${simulation_directory}all_simulations.dta", clear
}
else {
local counter = 1

foreach file of local files {

        if `counter' == 1 {
                use "`dir'`file'", clear
        }
        else {
                append using "`dir'`file'", force
        }
        local counter = `counter'+1
}

// strip back to just simulation number
bysort simulation: gen dup = _n
keep if dup == 1
keep simulation
// merge with full simulation dataset
        merge 1:m simulation using "${simulation_directory}all_simulations.dta"
// keep unmerged
keep if _merge != 3
}
```

G.5 Data file 5: append_results_methods.do

```
////////////////////////////////////////////////////////////////////////////////
// Appending results final file
// Simon Turner
//
// This program appends all the results from the analysed ITS datasets
//
////////////////////////////////////////////////////////////////////////////////
//
// This is called from the Master file which supplies the parameters
//
////////////////////////////////////////////////////////////////////////////////


// first we append within each model type...
        foreach model_type of global model_types {

                local model_directory "${results_directory}\`model_type'\"

                local files: dir "`model_directory'" files "*.dta"

                display `"`files'"'

                // append all together

                local counter = 1

                        foreach file of local files {

                                if `counter' == 1 {
                                        use "`model_directory'\`file'", clear
                                }
                                else {
                                        append using "`model_directory'\`file'", force
                                }
                                local counter = `counter'+1
                        }

                sort simulation pre_num_points

                save ${results_directory}results_`model_type'.dta, replace

        }

// then we can append all the different model types together
        display "Now appending the different model types together"

        local counter = 1

        foreach model_type of global model_types {


                display "appending `model_type'"
                if `counter' == 1 {
                        use ${results_directory}results_`model_type'.dta, clear
                        local counter = `counter' + 1
                }
                else {

                        append using ${results_directory}results_`model_type'.dta
                }
        }

        sort model_type simulation pre_num_points

        save ${results_directory}all_results.dta, replace
```

G.6 Data file 6: append_results_rhos.do

```
////////////////////////////////////////////////////////////////////////////////
// Appending results final file - combines all rho values
// Simon Turner
//
// This program appends all the results from the analysed ITS datasets including
// different values of autocorrelation
//
////////////////////////////////////////////////////////////////////////////////
//
// This is called from the Master file which supplies the parameters
//
////////////////////////////////////////////////////////////////////////////////

local counter = 1
display "rhos ${rho_list}"
local rhos = "${rho_list}"
display "rhos `rhos'"
foreach rho of global rho_list {
        display "rho `rho'"
        if `counter' == 1 {
                use "${data_directory}\${new_subfolder_name}\rho_`rho'\results\all_results.dta",
clear
        }
        else {
                append using
"${data_directory}\${new_subfolder_name}\rho_`rho'\results\all_results.dta", force
        }

        local counter = `counter'+1


}
display "b2 $b2 b3 $b3"
save "${data_directory}\${new_subfolder_name}\all_results_b2_${b2}_b3_${b3}.dta", replace
```

G.7 Data file 7: simsum_final.do

```
////////////////////////////////////////////////////////////////////////////////
// Simsum file, this runs the simsum command to get summaries and MCSEs of results
// Simon Turner
//
// It also combines these results from all the options into a single file
//
////////////////////////////////////////////////////////////////////////////////
//
// first we need to actually run the simsum program
//
// we're going to set up a new directory for this part
global simsum_directory "${data_directory}\simsum\"
cap mkdir ${simsum_directory}

////////////////////////////////////////////////////////////////////////////////
// Now we loop over the b2s and b3s and betas
foreach b2 of global b2_list {
   foreach b3 of global b3_list {


      forvalues beta_num = 0/3 {

         display "b2: `b2' b3: `b2' beta: `beta_num'"
         display "Date and time: $S_TIME $S_DATE"

         // get the correct results
         use "${data_directory}\b2_`b2'_b3_`b3'\all_results_b2_`b2'_b3_`b3'.dta", clear

         // run the simsum program, using each beta number, giving simsum:
         // the standard errors, true values, which variable has the methods etc.
         // OLS regression is chosen as the baseline reference method
         simsum beta_`beta_num', se(beta_`beta_num'_se) true(`=beta_`beta_num'_true[1]')
methodvar(model_type) ///
               id(simulation) by(pre_num_points rho_true) df(beta_`beta_num'_dof) ///
               dropbig mcse ref("regress") clear
         // pop variables in for b2 and b3 so that we can tell them apart later!
         gen b2 = "`b2'"
         gen b3 = "`b3'"
         save "${simsum_directory}\simsum_b2_`b2'_b3_`b3'_beta_`beta_num'.dta", replace

         display "b2: `b2' b3: `b2' beta: `beta_num' complete"
      }
   }
}

////////////////////////////////////////////////////////////////////////////////
// it's a bit easier to see all the different methods in long format
// so this converts the wide to long...

forvalues beta_num = 0/3 {
   foreach b2 of global b2_list {
      foreach b3 of global b3_list {

         // grab the appropriate simsum file
         use "${simsum_directory}\simsum_b2_`b2'_b3_`b3'_beta_`beta_num'.dta", clear

         //
         gen big_rho = rho_true*10
         gen int_rho = int(big_rho)
         // the variable names are a bit long at the moment, so shorten them
         foreach model of global model_types {
            rename beta_`beta_num'`model' summary`model'
            rename beta_`beta_num'`model'_mcse mcse`model'
         }
         // reshape to long...
         reshape long summary@ mcse@, i(pre_num_points rho_true perfmeasnum) j(model) string
         // drop bits we don't need
         keep pre_num_points rho_true model summary mcse perfmeascode int_rho
         // now reshape the summary and mcse wide
         reshape wide summary mcse , i(pre_num_points int_rho model) j(perfmeascode) string

         rename model model_type
         // add these bits to tell this file apart from others
         gen b2 = "`b2'"
         gen b3 = "`b3'"
         gen beta_num = "`beta_num'"

         save "${simsum_directory}\simsum_b2_`b2'_b3_`b3'_beta_`beta_num'_long.dta", replace

      }
   }
}

////////////////////////////////////////////////////////////////////////////////
// Finally combine into one big file
local first_file = 1
forvalues beta_num = 0/3 {
   foreach b2 of global b2_list {
      foreach b3 of global b3_list {

         if `first_file' == 1 {
            use "${simsum_directory}\simsum_b2_`b2'_b3_`b3'_beta_`beta_num'_long.dta", clear
            local first_file = 0
         }
         else {
            append using "${simsum_directory}\simsum_b2_`b2'_b3_`b3'_beta_`beta_num'_long.dta"
         }

      }
   }
}

save "${data_directory}\all_simsum_long.dta", replace
```

Appendix H. Additional file 1 accompanying Chapter 6 – Computer code to analyse data sets for the empirical evaluation

“Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series”

Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, McKenzie JE.

The following sections contain the Stata 15 computer code used to analyse the data from the empirical evaluation study. The published and digitally extracted data can be found on the online repository figshare: https://doi.org/10.6084/m9.figshare.13297136 (59).

H.1 Data file 1: 001_master_empirical_evaluation.do

```
////////////////////////////////////////////////////////////////////////////////
// Master Empirical Evaluation final file
// Simon Turner
// For the study:
// "Comparison of six statistical methods for interrupted time series studies: empirical evaluation
of 190 published series"
// Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, McKenzie JE.
//
// This do file calls multiple other do files to:
//
// analyse each dataset using multiple statistical methods (empirical_all_methods_published.do)
// save level changes, slope changes, SEs, p-values and autocorrelation estimates
// combine all the results from the different methods (combine_methods_published.do)
// and choose the estimate wanted for analysis (get_wanted_empirical_estimates_published.do)
// the final file used for the manuscript (including graphs) is empirical_estimates.dta
//
// required data are found in the two excel files:
// STurner_Empirical_study_information.xls
// which contains information about the studies
// also
// STurner_Empirical_time_series.xls
// which contains the actual time series data
//
// further details about the variables can be found in:
// STurner_Empirical_Data_Dictionary.xls

////////////////////////////////////////////////////////////////////////////////
// firstly set the working directory...
global dir  "<set directory>"
cd "$dir"

version 15

////////////////////////////////////////////////////////////////////////////////
// analyse the datasets
// now use different methods and capture the outputs
do ${dir}empirical_all_methods_published

// combine the results from the different methods
do ${dir}combine_methods_published

// for the empirical study we only want to use a subset of the total datasets
// this next section obtains the wanted datasets
do ${dir}get_wanted_empirical_estimates_published

////////////////////////////////////////////////////////////////////////////////
// The final file we can use for graphing, tables etc. is called:
// empirical_estimates_published.dta
////////////////////////////////////////////////////////////////////////////////
```

H.2 Data file 2: empirical_all_methods_published.do

```
////////////////////////////////////////////////////////////////////////////////
// This is the analysis file for the empirical study:
// "Comparison of six statistical methods for interrupted time series studies: empirical evaluation
of 190 published series"
// loads the datasets,
// analyses each using a variety of statistical methods
// finally it saves the analysed data in "estimates_published_`model_type'"
// where `model_type' is one of the statistical methods used

// load the data that is ready for analysis
import excel "STurner_Empirical_time_series.xls", sheet("Sheet1") firstrow clear

// ensure that there is a proper study id
levelsof series_id, local(series_ids)

// there are several alternatives for scaling
// no scaling (just leave this blank)
// scaling by the rmse for the first segment only, use "rmse_on"
// or scaling by the rmse of the whole series, use "rmse_full"
// some of the datasets are very short (three points in pre-series) so first segment scaling does
not work
// scaling by the full rmse was used for the analysis
local rmse_full = "rmse_full"

////////////////////////////////////////////////////////////////////////////////
// this short program estimates autocorrelation from the residuals after a
// simple linear regression
// this was not used in the analysis, but served as an interesting reference

cap program drop find_rho
program find_rho

syntax varlist [if]
marksample touse
quietly regress `varlist' l.`varlist' if `touse' , nocons
local rho=_b[L.`varlist']
scalar rho = `rho'

end

cap program drop find_rho_wrap
program find_rho_wrap

syntax varlist [if]

      marksample touse

      tempvar resid
      regress `varlist' if `touse'
      predict double `resid', resid
      find_rho `resid' if `touse'

end

////////////////////////////////////////////////////////////////////////////////
// we investigated a range of statistical methods
// this string links to the sections below to identify which methods
// are going to be used here
// regress - OLS
// newey - OLS with newey-west standard errors
// prais - prais-winsten with an iterative search
// prais_raw - simple prais-winsten
// corc - cochrane-orcutt
// mixed - REML
// mixed_satt - REML with the Satterthwaite small series adjustment
// arima - ARIMA with lag-1

// for the final study we restricted these to the following...
local model_types "regress newey prais_raw mixed mixed_satt arima"

// loop over each statistical method
foreach model_type of local model_types {

   // set the end file name
   local save_name "estimates_published_`model_type'.dta"
   // set up the temporary file used to store the data
   // we are saving data per segment
   // the level change and slope change with CIs, SEs and p-values
   // autocorrelation estimates, degrees of freedom etc.
   tempname post_values_`model_type'
   postfile `post_values_`model_type''  series_id ///
                     str20 model_type  ///
                     segment analysis_autocorr analysis_effects segment_num_points total_num_points
rmse ///
                     level level_ll level_ul level_se level_p ///
                     slope slope_ll slope_ul slope_se slope_p ///
                     rho_est rho_cil rho_ciu ///
                     num_iterations error_code converged ///
                     lincom_level_dof lincom_slope_dof ///
                     using "`save_name'" , replace

   // now for each of the data sets...
   foreach series_id of local series_ids {

      // going to just use one at a time
      preserve
      keep if series_id == `series_id'

      local series_id = series_id[1]

      display "working through `series_id'"

      ////////////////////////////////////////////////////////////////////////////
      // set for program

      *keep outcome time segment segment_in_analysis
      sort time
      drop if time == .

      ////////////////////////////////////////////////////////////////////////////////
      // find times programatically
      // this goes through and works out the timing of each segment
      summ segment
      local num_segments = r(max)
      local min_seg_num = r(min)

      if `min_seg_num' != 0 {
         replace segment = segment - `min_seg_num'
      }

      summ segment
      local num_segments = r(max)
      local min_seg_num = r(min)

      forvalues segment = 0/`num_segments' {
         summ time if segment == `segment'
         local time_`segment'_start = r(min)
         local time_`segment'_end = r(max)
         display "segment `segment' goes from `time_`segment'_start' to `time_`segment'_end'"
      }

      ////////////////////////////////////////////////////////////////////////////////
      // extra variables for analysis
      // generate variables to indicate time of intervention

         forvalues segment = 0/`num_segments' {
            gen intervention_`segment' = 0
            replace intervention_`segment' = 1 if segment >= `segment'
            gen level_change_`segment' = intervention_`segment'
            gen slope_change_`segment' = (time-`time_`segment'_start')*level_change_`segment'
         }

      ////////////////////////////////////////////////////////////////////////////////
      // model
      ////////////////////////////////////////////////////////////////////////////////


      // create the variables to use (the various level and slope changes)
      // starting with level_0 and slope_0, then incrementing for each segment
      // e.g. regular segmented regression will have
      // level_0 (intercept) slope_0 (pre-interruption slope) level_1 (level change at
interruption) slope_1 (slope change post interruption)
      local variables = ""
      forvalues segment = 0/`num_segments' {
            local variables = "`variables'" + " level_change_`segment' slope_change_`segment'"
      }

      display "variables: `variables'"
      display "study: `year' number: `number_in_year' multiple: `multiple'  data_type: `data_type'"

      // first find rmse from all segments (for scaling)
      regress outcome `variables', nocons
      local rmse = e(rmse)
      local total_num_points = e(N)

      ///////////////////////////////////////////
      // now apply the correct statistical method according to model_type...

         tsset time

               local rho_est = .
               local rho_est_cil = .
               local rho_est_ciu = .

               local model_error = 0


               if "`model_type'" == "regress" { // basic OLS regression
                  regress outcome `variables', nocons
                  matrix local_results = r(table)
                  local rho_est = 0
                  local num_iterations = 1
                  local error_code = _rc
                  local converged = 1
               } // end regress check
               else if "`model_type'" == "newey" { // OLS regression with newey-west standard
errors
                  newey outcome `variables', nocons lag(1) force // need force option to ignore
missing values otherwise get time not equally spaced errors
                  matrix local_results = r(table)
                  local varlist = "outcome `variables'"
                  quietly find_rho_wrap `varlist'
                  local rho_est = rho
                  local num_iterations = 1
                  local error_code = _rc
                  local converged = 1
                  newey outcome `variables', nocons lag(1) force // need force option to ignore
missing values otherwise get time not equally spaced errors
               } // end newey check
               else if "`model_type'" == "prais" { // Prais-Winsten with iterative search
                  cap prais outcome `variables', nocons  ssesearch
                  if _rc != 0 {
                     local model_error = _rc
                     matrix local_results = J(6,5,.)
                     local rho_est = .
                     local converged = .
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     matrix local_results = r(table)
                     local rho_est = e(rho)
                     local num_iterations = e(ic)
                     local error_code = _rc
                     local converged = 1
                  }
               } // end prais check
               else if "`model_type'" == "prais_raw" { // standard Prais-Winsten
                  cap prais outcome `variables', nocons
                  if _rc != 0 {
                     local model_error = _rc
                     matrix local_results = J(6,5,.)
                     local rho_est = .
                     local converged = .
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     matrix local_results = r(table)
                     local rho_est = e(rho)
                     local num_iterations = e(ic)
                     local error_code = _rc
                     local converged = 1
                  }
               } // end prais_raw check
               else if "`model_type'" == "corc" { // Cochrane-Orcutt
                  cap prais outcome `variables', nocons  corc ssesearch
                  if _rc != 0 {
                     local model_error = _rc
                     matrix local_results = J(6,5,.)
                     local rho_est = .
                     local converged = .
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     matrix local_results = r(table)
                     local rho_est = e(rho)
                     local num_iterations = e(ic)
                     local error_code = _rc
                     local converged = 1
                  }
               } // end corc check
               else if "`model_type'" == "mixed" { // REML with maximum iterations set to 1000 to
stop really long non-convergence
                  cap mixed outcome `variables', nocons   res(ar 1, t(time)) var reml iter(1000)
                  if _rc != 0 {
                     local model_error = _rc
                     matrix local_results = J(6,5,.)
                     local rho_est = .
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     matrix local_results = r(table)
                     local num_cols = colsof(local_results)
                     local rho_est = tanh(local_results[1,`num_cols'])
                     local rho_est_cil = tanh(local_results[5,`num_cols'])
                     local rho_est_ciu = tanh(local_results[6,`num_cols'])
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
               } // end mixed check
               else if "`model_type'" == "mixed_kr" { // REML with KR adjustment
                  cap mixed outcome `variables', nocons  res(ar 1, t(time)) var reml iter(1000)
dfmethod(kr)
                  if _rc != 0 {
                     local model_error = _rc
                     matrix local_results = J(6,5,.)
                     local rho_est = .
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     matrix local_results = r(table)
                     local num_cols = colsof(local_results)
                     local rho_est = tanh(local_results[1,`num_cols'])
                     local rho_est_cil = tanh(local_results[5,`num_cols'])
                     local rho_est_ciu = tanh(local_results[6,`num_cols'])
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc
                     mat dfs = e(df)
                  }
               } // end mixed_kr check
               else if "`model_type'" == "mixed_satt" { // REML with Satt adjustment
                  cap mixed outcome `variables', nocons   res(ar 1, t(time)) var reml iter(1000)
dfmethod(satt)
                  if _rc != 0 {
                     local model_error = _rc
                     matrix local_results = J(6,5,.)
                     local rho_est = .
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc
                  }
                  else {
                     display "`model_type' model ran"
                     matrix local_results = r(table)
                     local num_cols = colsof(local_results)
                     local rho_est = tanh(local_results[1,`num_cols'])
                     local rho_est_cil = tanh(local_results[5,`num_cols'])
                     local rho_est_ciu = tanh(local_results[6,`num_cols'])
                     local converged = e(converged)
                     local num_iterations = e(ic)
                     local error_code = _rc
                     mat dfs = e(df)
                  }
               } // end mixed_satt check
               else if "`model_type'" == "arima" {
                     cap arima outcome `variables', nocons  collinear ar(1) iter(1000)
                     if _rc != 0 {
                     // new section for error catch includes a "slow down" as for some reason
                     // Stata sometimes crashes if it went too quickly here (Stata 15.0)
                     local model_error = _rc
                        local iteration = 1
                        while `iteration' < 10 & _rc != 0 {
                           cap matrix local_results = J(6,5,.)
                           local rho_est = .
                           local converged = e(converged)
                           local num_iterations = e(ic)
                           local error_code = _rc
                           sleep 10
                           local iteration = `iteration' + 1
                        }
                     }
                     else {
                        matrix local_results = r(table)
                        local num_cols = colsof(local_results)
                        local rho_col = `num_cols' - 1
                        local rho_est = (local_results[1,`rho_col'])
                        local rho_est_cil = (local_results[5,`rho_col'])
                        local rho_est_ciu = (local_results[6,`rho_col'])
                        local converged = e(converged)
                        local num_iterations = e(ic)
                        local error_code = _rc
                     }
               } // end arima check
               else {
                  display "Model type `model_type' is not recognised"
                  stop
               }

      display "study: `year' number: `number_in_year' multiple: `multiple'  data_type: `data_type'
model `model_type' ran with error code `model_error'"

      // if there was an error just set the output to missing values
      if `model_error' != 0 {
         forvalues segment = 0/`num_segments' {
            local level_change_counter_`segment' = .
            local level_change_counter_`segment'_ll = .
            local level_change_counter_`segment'_ul = .
            local level_change_counter_`segment'_se = .
            local level_change_counter_`segment'_p = .
            local slope_change_counter_`segment' = .
            local slope_change_counter_`segment'_ll = .
            local slope_change_counter_`segment'_ul = .
            local slope_change_counter_`segment'_se = .
            local slope_change_counter_`segment'_p = .
         }
      }
      else {
         // use lincom to find the CIs that go with the various level and slope changes

         // predict the estimates to find the counterfactual, level change etc.
         // base estimates of level change compared to first segment
         predict estimates
         gen counterfactual = _b[level_change_0] + _b[slope_change_0]*slope_change_0

         // if satterthwaite is used need to add ",small" to the options after lincom
         if "`model_type'" == "mixed_satt" {
            local small " , small"
         }
         else {
            local small ""
         }

            ///////////////////////////////////////////////////////////////////////////////
            // level changes from counterfactual
            // for each segment...
            forvalues segment = 0/`num_segments' {

               // for the first segment just initialise everything
               if `segment' == 0 {
                  local level_change_counter_`segment' = .
                  local level_change_counter_`segment'_ll = .
                  local level_change_counter_`segment'_ul = .
                  local level_change_counter_`segment'_se = .
                  local level_change_counter_`segment'_p = .
               }
               else {
                  // for each subsequent segment
                  // set up a local macro that holds the names of all the lincom variables we want
                  // this is going to be
                  // level_change_1 for the first level change
                  // and then adding on the subsequent level and slope change values for any
subsequent segments
                  forvalues segment_sub = 1/`segment' {
                     if `segment_sub' == 1 {
                        local L`segment' = `" level_change_`segment_sub' "'
                     }
                     else {
                        local segment_multiplier = slope_change_`=`segment_sub'-
1'[`time_`segment'_start']
                        local L`segment' = `" `L`segment'' "' + `" + slope_change_`=`segment_sub'-
1'*`segment_multiplier' + level_change_`segment_sub' "'
                     }
                     *display "L`segment' `L`segment''"

                  }
                  // now lincom those variables (adding the small option if using REML-Satt)
                  lincom `L`segment'' `small'
                  // lincom gives slightly different responses in different situations...
                  if "`model_type'" == "arima" | "`model_type'" == "mixed" {
                     local lincom_level_dof = e(N) - e(df_m)
                  }
                  else {
                     local lincom_level_dof = r(df)
                  }

                  *return list

                  display "level lincom dof = `lincom_level_dof'"

                  // we are using a cut-off of 2 for the degrees of freedom for the REML-Satt
method
                  if `lincom_level_dof' < 2 & "`model_type'" == "mixed_satt" {
                     lincom `L`segment'', df(2)
                  }

                  // now save those values
                  local level_change_counter_`segment' = r(estimate)
                  local level_change_counter_`segment'_ll = r(lb)
                  local level_change_counter_`segment'_ul = r(ub)
                  local level_change_counter_`segment'_se = r(se)
                  local level_change_counter_`segment'_p = r(p)
               } // end if loop

            } // end segment loop

         ////////////////////////////////////////////////////////////////////////////////
         // slope changes from counterfactual
         // this runs as the level change above, but for slope change values instead
         forvalues segment = 0/`num_segments' {
            if `segment' == 0 {
               local slope_change_counter_`segment' = .
               local slope_change_counter_`segment'_ll = .
               local slope_change_counter_`segment'_ul = .
               local slope_change_counter_`segment'_se = .
               local slope_change_counter_`segment'_p = .
            }
            else {
               forvalues segment_sub = 1/`segment' {
                  if `segment_sub' == 1 {
                     local SC_`segment' = "slope_change_`segment_sub'"
                  }
                  else {
                     local SC_`segment' = "`SC_`segment''" + " + slope_change_`segment_sub'"
                  }
               }

               lincom `SC_`segment'' `small'
               if "`model_type'" == "arima" | "`model_type'" == "mixed" {
                  local lincom_slope_dof = e(N) - e(df_m)
               }
               else {
                  local lincom_slope_dof = r(df)
               }

               *return list

               display "slope lincom dof = `lincom_slope_dof'"

               if `lincom_slope_dof' < 2 & "`model_type'" == "mixed_satt" {
                  lincom `SC_`segment'', df(2)
               }

               local slope_change_counter_`segment' = r(estimate)
               local slope_change_counter_`segment'_ll = r(lb)
               local slope_change_counter_`segment'_ul = r(ub)
               local slope_change_counter_`segment'_se = r(se)
               local slope_change_counter_`segment'_p = r(p)
            } // end if loop
         } // end segment loop

      } // end model error check.

      ////////////////////////////////////////////////////////////////////////////////
      // postvalues
      // finally put the values into the post file

      forvalues segment = 0/`num_segments' {
         local analysis_autocorr = analysis_autocorr[`time_`segment'_start']
         // in the excel file the segments begin with 1, in this file we set to zero
         // therefore we need to similarly reduce the analysis effects by 1 so they all match up
         local analysis_effects = analysis_effects[`time_`segment'_start'] - 1
         summ time if segment == `segment'
         local segment_num_points = r(N)

         // scaled adjustment by rmse
         local level_change_counter_`segment' = `level_change_counter_`segment''/`rmse'
         local level_change_counter_`segment'_ll = `level_change_counter_`segment'_ll'/`rmse'
         local level_change_counter_`segment'_ul = `level_change_counter_`segment'_ul'/`rmse'
         local level_change_counter_`segment'_se = `level_change_counter_`segment'_se'/`rmse'
         local slope_change_counter_`segment' = `slope_change_counter_`segment''/`rmse'
         local slope_change_counter_`segment'_ll = `slope_change_counter_`segment'_ll'/`rmse'
         local slope_change_counter_`segment'_ul = `slope_change_counter_`segment'_ul'/`rmse'
         local slope_change_counter_`segment'_se = `slope_change_counter_`segment'_se'/`rmse'

         post `post_values_`model_type'' (`series_id') ///
                  ("`model_type'") ///
                  (`segment') (`analysis_autocorr') (`analysis_effects') (`segment_num_points')
(`total_num_points') (`rmse') ///
                  (`level_change_counter_`segment'') (`level_change_counter_`segment'_ll')
(`level_change_counter_`segment'_ul') (`level_change_counter_`segment'_se')
(`level_change_counter_`segment'_p') /// ///
                  (`slope_change_counter_`segment'') (`slope_change_counter_`segment'_ll')
(`slope_change_counter_`segment'_ul') (`slope_change_counter_`segment'_se')
(`slope_change_counter_`segment'_p') ///
                  (`rho_est') (`rho_est_cil') (`rho_est_ciu') ///
                  (`num_iterations') (`error_code') (`converged') (`lincom_level_dof')
(`lincom_slope_dof')
      } // end segment loop

      restore
   } // end study id loop

   postclose `post_values_`model_type''
} // end of model type loop
```

H.3 Data file 3: combine_methods_published.do

```
////////////////////////////////////////////////////////////////////////////////
// This is the analysis file for the empirical study:
// "Comparison of six statistical methods for interrupted time series studies: empirical evaluation
of 190 published series"
// this do file combines the results for all the different statistical methods
// the combined file is called estimates_published_all.dta

local model_types "regress newey prais_raw mixed mixed_satt arima"

local first = 0

foreach model of local model_types {

   if `first' == 0 {
      use "estimates_published_`model'.dta", clear
      local first = 1
   }
   else {
      append using "estimates_published_`model'.dta"
   }

}

save estimates_published_all.dta, replace
```

H.4 Data file 4: get_wanted_empirical_estimates_published.do

```
////////////////////////////////////////////////////////////////////////////////
// This is the final do file for the empirical study:
// "Comparison of six statistical methods for interrupted time series studies: empirical evaluation
of 190 published series"
// It selects the effect estimate wanted for final comparisons
// excluding segments that are not required (e.g. transition periods)
// It requires the data from STurner_Empirical_study_information which contains the desired segment
number
// The final file "empirical_estimates_published.dta" is the output

// keep only the segments we need
keep if segment == analysis_effects

// tidy the file
sort series_id model_type segment
order series_id model_type segment segment_num_points total_num_points

// save the final data file
save empirical_estimates_published.dta, replace
```

Appendix I. Additional file 2 accompanying Chapter 6 – Citation details of the studies from which datasets were obtained

"Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series"
Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, McKenzie JE.

The following are citation details of the studies that contributed data via publication, email or digital data extraction for the empirical evaluation study.

We wish to thank all of the authors who generously contributed datasets for this study.

Study	Data via publication	Data via email	Data extracted
Abegaz T, Berhane Y, Worku A, et al. Effectiveness of an improved road safety policy in Ethiopia: an interrupted time series study. BMC Public Health 2014;14(1) doi: 10.1186/1471-2458-14-539	No	No	No
Adams AS, Soumerai SB, Zhang F, et al. Effects of Eliminating Drug Caps on Racial Differences in Antidepressant Use Among Dual Enrollees With Diabetes and Depression. Clinical Therapeutics 2015;37(3):597-609. doi: 10.1016/j.clinthera.2014.12.011	No	No	Yes
Aiken AM, Wanyoro AK, Mwangi J, et al. Changing Use of Surgical Antibiotic Prophylaxis in Thika Hospital, Kenya: A Quality Improvement Intervention with an Interrupted Time Series Design. PLoS ONE 2013;8(11):e78942. doi: 10.1371/journal.pone.0078942	No	No	Yes
Akhtar S, Ziyab AH. Impact of the Penalty Points System on Severe Road Traffic Injuries in Kuwait. Traffic Injury Prevention 2013;14(7):743-48. doi: 10.1080/15389588.2012.749466	No	No	Yes
Alexandridis AA, McCort A, Ringwalt CL, et al. A statewide evaluation of seven strategies to reduce opioid overdose in North Carolina. Injury Prevention 2017;24(1):48-54. doi: 10.1136/injuryprev-2017-042396	No	No	Yes
Alpert HR, Carpenter D, Connolly GN. Tobacco industry response to a ban on lights descriptors on cigarette packaging and population outcomes. Tobacco Control 2017;27(4):390-98. doi: 10.1136/tobaccocontrol-2017-053683	Yes	N/A	Yes

Andersen SE, Knudsen JD. A managed multidisciplinary programme on multi-resistant Klebsiella pneumoniaein a Danish university hospital. BMJ Quality & Safety 2013;22(11):907-15. doi: 10.1136/bmjqs-2012-001791	No	No	Yes
Andrade AL, Minamisava R, Policena G, et al. Evaluating the impact of PCV-10 on invasive pneumococcal disease in Brazil: A time-series analysis. Human Vaccines & Immunotherapeutics 2016;12(2):285-92. doi: 10.1080/21645515.2015.1117713	No	Yes	No
Armah G, Pringle K, Enweronu-Laryea CC, et al. Impact and Effectiveness of Monovalent Rotavirus Vaccine Against Severe Rotavirus Diarrhea in Ghana. Clinical Infectious Diseases 2016;62(suppl 2):S200-S07. doi: 10.1093/cid/ciw014	No	No	Yes
Barber C, Gagnon D, Fonda J, et al. Assessing the impact of prescribing directives on opioid prescribing practices among Veterans Health Administration providers. Pharmacoepidemiology and Drug Safety 2016;26(1):40-46. doi: 10.1002/pds.4066	No	No	Yes
Barocas DA, Mallin K, Graves AJ, et al. Effect of the USPSTF Grade D Recommendation against Screening for Prostate Cancer on Incident Prostate Cancer Diagnoses in the United States. Journal of Urology 2015;194(6):1587-93. doi: 10.1016/j.juro.2015.06.075	No	No	Yes
Baskerville NB, Brown KS, Nguyen NC, et al. Impact of Canadian tobacco packaging policy on use of a toll-free quit-smoking line: an interrupted time-series analysis. CMAJ Open 2016;4(1):E59-E65. doi: 10.9778/cmajo.20150104	No	No	Yes
Been JV, Mackay DF, Millett C, et al. Impact of smoke-free legislation on perinatal and infant mortality: a national quasi-experimental study. Scientific Reports 2015;5(1) doi: 10.1038/srep13020	No	No	No
Been JV, Szatkowski L, van Staa T-P, et al. Smoke-free legislation and the incidence of paediatric respiratory infections and wheezing/asthma: interrupted time series analyses in the four UK nations. Scientific Reports 2015;5(1) doi: 10.1038/srep15246	No	No	No
Bell S, Davey P, Nathwani D, et al. Risk of AKI with Gentamicin as Surgical Prophylaxis. Journal of the American Society of Nephrology 2014;25(11):2625-32. doi: 10.1681/asn.2014010035	No	Yes	Yes

Bendzsak AM, Baxter NN, Darling GE, et al. Regionalization and Outcomes of Lung Cancer Surgery in Ontario, Canada. Journal of Clinical Oncology 2017;35(24):2772-80. doi: 10.1200/jco.2016.69.8076	No	No	Yes
Berkowitz SA, Percac-Lima S, Ashburner JM, et al. Building Equity Improvement into Quality Improvement: Reducing Socioeconomic Disparities in Colorectal Cancer Screening as Part of Population Health Management. Journal of General Internal Medicine 2015;30(7):942-49. doi: 10.1007/s11606-015-3227-4	No	No	Yes
Bernat DH, Maldonado-Molina M, Hyland A, et al. Effects of Smoke-Free Laws on Alcohol-Related Car Crashes in California and New York: Time Series Analyses From 1982 to 2008. American Journal of Public Health 2013;103(2):214-20. doi: 10.2105/ajph.2012.300906	No	No	No
Blais E, Carnis L. Improving the safety effect of speed camera programs through innovations: Evidence from the French experience. Journal of Safety Research 2015;55:135-45. doi: 10.1016/j.jsr.2015.08.007	No	No	No
Bobo WV, Epstein RA, Hayes RM, et al. The effect of regulatory advisories on maternal antidepressant prescribing, 1995–2007: an interrupted time series study of 228,876 pregnancies. Archives of Women's Mental Health 2013;17(1):17-26. doi: 10.1007/s00737-013-0383-6	No	No	Yes
Boden DG, Agarwal A, Hussain T, et al. Lowering levels of bed occupancy is associated with decreased inhospital mortality and improved performance on the 4-hour target in a UK District General Hospital. Emergency Medicine Journal 2015;33(2):85-90. doi: 10.1136/emermed-2014-204479	No	No	Yes
Boel J, Andreasen V, Jarløv JO, et al. Impact of antibiotic restriction on resistance levels ofEscherichia coli: a controlled interrupted time series study of a hospital-wide antibiotic stewardship programme. Journal of Antimicrobial Chemotherapy 2016;71(7):2047-51. doi: 10.1093/jac/dkw055	No	Yes	Yes
Bonander C, Nilson F, Andersson R. The effect of the Swedish bicycle helmet law for children: An interrupted time series study. Journal of Safety Research 2014;51:15-22. doi: 10.1016/j.jsr.2014.07.001	No	Yes	Yes

Borde JP, Kern WV, Hug M, et al. Implementation of an intensified antibiotic stewardship programme targeting third-generation cephalosporin and fluoroquinolone use in an emergency medicine department. Emergency Medicine Journal 2014;32(7):509-15. doi: 10.1136/emermed-2014-204067	No	No	Yes
Bowden JA, Dono J, John DL, et al. What happens when the price of a tobacco retailer licence increases? Tobacco Control 2013;23(2):178-80. doi: 10.1136/tobaccocontrol-2012-050615	No	No	Yes
Bozorgmehr K, Razum O. Effect of Restricting Access to Health Care on Health Expenditures among Asylum-Seekers and Refugees: A Quasi-Experimental Study in Germany, 1994–2013. PLOS ONE 2015;10(7):e0131483. doi: 10.1371/journal.pone.0131483	No	No	No
Branas CC, Kastanaki AE, Michalodimitrakis M, et al. The impact of economic austerity and prosperity events on suicide in Greece: a 30-year interrupted time-series analysis. BMJ Open 2015;5(1):e005619-e19. doi: 10.1136/bmjopen-2014-005619	No	No	No
Bugden S, Friesen KJ, Falk J. Voluntary warnings and the limits of good prescribing behavior: the case for de-adoption of meperidine. Journal of Pain Research 2015:879. doi: 10.2147/jpr.s96625	No	No	Yes
Burke LK, Brown CP, Johnson TM. Historical Data Analysis of Hospital Discharges Related to the Amerithrax Attack in Florida. Perspect Health Inf Manag 2016;13(Fall):1c-1c.	No	No	No
Busch SH, McGinty EE, Stuart EA, et al. Was federal parity associated with changes in Out-of-network mental health care use and spending? BMC Health Services Research 2017;17(1) doi: 10.1186/s12913-017-2261-9	No	No	Yes
Cairns KA, Jenney AWJ, Abbott IJ, et al. Prescribing trends before and after implementation of an antimicrobial stewardship program. The Medical Journal of Australia 2013;198(5):262-66. doi: 10.5694/mja12.11683	No	No	Yes
Carracedo-Martínez E, Pia-Morandeira A, Figueiras A. Trends in celecoxib and etoricoxib prescribing following removal of prior authorization requirement in Spain. Journal of Clinical Pharmacy and Therapeutics 2016;42(2):185-88. doi: 10.1111/jcpt.12490	No	Yes	Yes

Cecil E, Bottle A, Sharland M, et al. Impact of UK Primary Care Policy Reforms on Short-Stay Unplanned Hospital Admissions for Children With Primary Care-Sensitive Conditions. The Annals of Family Medicine 2015;13(3):214-20. doi: 10.1370/afm.1786	No	No	No
Chandran A, Pérez-Núñez R, Bachani AM, et al. Early Impact of a National Multi-Faceted Road Safety Intervention Program in Mexico: Results of a Time-Series Analysis. PLoS ONE 2014;9(1):e87482. doi: 10.1371/journal.pone.0087482	No	Yes	Yes
Chang C-H, Lin J-W, Wu L-C, et al. National Antiviral Treatment Program and the Incidence of Hepatocellular Carcinoma and Associated Mortality in Taiwan. Medical Care 2013;51(10):908-13. doi: 10.1097/mlr.0b013e3182a502ba	No	No	No
Chen IL, Lee C-H, Su L-H, et al. Effects of implementation of an online comprehensive antimicrobial-stewardship program in ICUs: A longitudinal study. Journal of Microbiology, Immunology and Infection 2018;51(1):55-63. doi: 10.1016/j.jmii.2016.06.007	No	No	Yes
Cheng C-L, Chao P-H, Hsu JC-S, et al. Utilization patterns of Antihyperuricemic Agents Following Safety Announcement on Allopurinol and Benzbromarone by Taiwan Food and Drug Administration. Pharmacoepidemiology and Drug Safety 2013;23(3):309-13. doi: 10.1002/pds.3550	No	No	Yes
Cheng J, Benassi P, de Oliveira C, et al. Impact of a mass media mental health campaign on psychiatric emergency department visits. Canadian Journal of Public Health 2016;107(3):e303-e11. doi: 10.17269/cjph.107.5265	No	No	Yes
Chua K-P, Shrime MG, Conti RM. Effect of FDA Investigation on Opioid Prescribing to Children After Tonsillectomy/Adenoidectomy. Pediatrics 2017;140(6):e20171765. doi: 10.1542/peds.2017-1765	No	No	Yes
Chung YK, Kim J-S, Lee SS, et al. Effect of daily chlorhexidine bathing on acquisition of carbapenem-resistant Acinetobacter baumannii (CRAB) in the medical intensive care unit with CRAB endemicity. American Journal of Infection Control 2015;43(11):1171-77. doi: 10.1016/j.ajic.2015.07.001	No	No	Yes
Čižman M, Plankar Srovin T, Blagus R, et al. The long-term effects of restrictive interventions on consumption and costs of antibiotics. Journal of Global Antimicrobial Resistance 2015;3(1):31-35. doi: 10.1016/j.jgar.2014.11.004	No	No	Yes

Corcoran P, Griffin E, Arensman E, et al. Impact of the economic recession and subsequent austerity on suicide and self-harm in Ireland: An interrupted time series analysis. International Journal of Epidemiology 2015;44(3):969-77. doi: 10.1093/ije/dyv058	No	No	Yes
Cunningham JK, Liu L-M, Callaghan RC. Essential (“Precursor”) chemical control for heroin: Impact of acetic anhydride regulation on US heroin availability. Drug and Alcohol Dependence 2013;133(2):520-28. doi: 10.1016/j.drugalcdep.2013.07.014	No	No	Yes
Damiani G, Federico B, Anselmi A, et al. The impact of Regional co-payment and National reimbursement criteria on statins use in Italy: an interrupted time-series analysis. BMC Health Services Research 2014;14(1) doi: 10.1186/1472-6963-14-6	No	No	Yes
Denkel LA, Schwab F, Garten L, et al. Protective Effect of Dual-Strain Probiotics in Preterm Infants: A Multi-Center Time Series Analysis. PLOS ONE 2016;11(6):e0158136. doi: 10.1371/journal.pone.0158136	No	Yes	Yes
Desai SP, Lu B, Szent-Gyorgyi LE, et al. Increasing pneumococcal vaccination for immunosuppressed patients: A cluster quality improvement trial. Arthritis & Rheumatism 2012;65(1):39-47. doi: 10.1002/art.37716	No	No	Yes
Deslandes PN, Jenkins KSL, Haines KE, et al. A change in the trend in dosulepin usage following the introduction of a prescribing indicator but not after two national safety warnings. Journal of Clinical Pharmacy and Therapeutics 2016;41(2):224-28. doi: 10.1111/jcpt.12376	No	Yes	Yes
Dicks KV, Lofgren E, Lewis SS, et al. A Multicenter Pragmatic Interrupted Time Series Analysis of Chlorhexidine Gluconate Bathing in Community Hospital Intensive Care Units. Infection Control & Hospital Epidemiology 2016;37(7):791-97. doi: 10.1017/ice.2016.23	No	Yes	No
Dik J-WH, Hendrix R, Lo-Ten-Foe JR, et al. Automatic day-2 intervention by a multidisciplinary antimicrobial stewardship-team leads to multiple positive effects. Frontiers in Microbiology 2015;06 doi: 10.3389/fmicb.2015.00546	No	No	Yes

DiMaggio C, Chen Q, Muennig PA, et al. Timing and effect of a safe routes to school program on child pedestrian injury risk during school travel hours: Bayesian changepoint and difference-in-differences analysis. Injury Epidemiology 2014;1(1) doi: 10.1186/s40621-014-0017-0	No	No	Yes
Doernberg SB, Dudas V, Trivedi KK. Implementation of an antimicrobial stewardship program targeting residents with urinary tract infections in three community long-term care facilities: a quasi-experimental study using time-series analysis. Antimicrobial Resistance and Infection Control 2015;4(1) doi: 10.1186/s13756-015-0095-y	No	No	Yes
Dresden SM, Powell ES, Kang R, et al. Increased Emergency Department Use in Illinois After Implementation of the Patient Protection and Affordable Care Act. Annals of Emergency Medicine 2017;69(2):172-80. doi: 10.1016/j.annemergmed.2016.06.026	No	No	Yes
Druetz T, Fregonese F, Bado A, et al. Abolishing Fees at Health Centers in the Context of Community Case Management of Malaria: What Effects on Treatment-Seeking Practices for Febrile Children in Rural Burkina Faso? PLOS ONE 2015;10(10):e0141306. doi: 10.1371/journal.pone.0141306	No	No	No
Emmerick ICM, Campos MR, Luiza VL, et al. Retrospective interrupted time series examining hypertension and diabetes medicines usage following changes in patient cost sharing in the 'Farmácia Popular' programme in Brazil. BMJ Open 2017;7(11):e017308. doi: 10.1136/bmjopen-2017-017308	No	No	Yes
Faryar KA, Freeman CL, Persaud AK, et al. The Effects of Kentucky's Comprehensive Opioid Legislation on Patients Presenting with Prescription Opioid or Heroin Abuse to One Urban Emergency Department. The Journal of Emergency Medicine 2017;53(6):805-14. doi: 10.1016/j.jemermed.2017.08.066	No	No	Yes
Filippidis FT, Gerovasili V, Millett C, et al. Medium-term impact of the economic crisis on mortality, health-related behaviours and access to healthcare in Greece. Scientific Reports 2017;7(1) doi: 10.1038/srep46423	Yes	N/A	No

Finnell KJ, John R, Thompson DM. 1% low-fat milk has perks!: An evaluation of a social marketing intervention. Preventive Medicine Reports 2017;5:144-49. doi: 10.1016/j.pmedr.2016.11.017	No	No	Yes
Fisher D, Tambyah PA, Lin RTP, et al. Sustained meticillin-resistant Staphylococcus aureus control in a hyper-endemic tertiary acute care hospital with infrastructure challenges in Singapore. Journal of Hospital Infection 2013;85(2):141-48. doi: 10.1016/j.jhin.2013.07.005	No	Yes	Yes
Flett KB, Ozonoff A, Graham DA, et al. Impact of Mandatory Public Reporting of Central Line–Associated Bloodstream Infections on Blood Culture and Antibiotic Utilization in Pediatric and Neonatal Intensive Care Units. Infection Control & Hospital Epidemiology 2015;36(8):878-85. doi: 10.1017/ice.2015.100	No	No	Yes
Flynn D, Ford GA, Rodgers H, et al. A Time Series Evaluation of the FAST National Stroke Awareness Campaign in England. PLoS ONE 2014;9(8):e104289. doi: 10.1371/journal.pone.0104289	No	No	Yes
Fournier P, Dumont A, Tourigny C, et al. The Free Caesareans Policy in Low-Income Settings: An Interrupted Time Series Analysis in Mali (2003–2012). PLoS ONE 2014;9(8):e105130. doi: 10.1371/journal.pone.0105130	No	No	Yes
Gadzhanova SV, Roughead EE, Bartlett MJ. Improving cardiovascular disease management in Australia: NPS MedicineWise. The Medical Journal of Australia 2013;199(3):192-95. doi: 10.5694/mja12.11779	No	No	Yes
Gale M, Muscatello DJ, Dinh M, et al. Alcopops, taxation and harm: a segmented time series analysis of emergency department presentations. BMC Public Health 2015;15(1) doi: 10.1186/s12889-015-1769-3	No	No	Yes
Gallini A, Andrieu S, Donohue JM, et al. Trends in use of antipsychotics in elderly patients with dementia: Impact of national safety warnings. European Neuropsychopharmacology 2014;24(1):95-104. doi: 10.1016/j.euroneuro.2013.09.003	No	No	Yes
Gamble J-M, Johnson JA, Majumdar SR, et al. Evaluating the introduction of a computerized prior-authorization system on the completeness of drug exposure data. Pharmacoepidemiology and Drug Safety 2013;22(5):551-55. doi: 10.1002/pds.3427	No	No	Yes

Garnett M, Charyk Stewart T, Miller MR, et al. Did Amendments to the Ontario Highway Traffic Act in 2009-2010 Affect the Proportion of Alcohol-Related Motor Vehicle Collisions Seen at a Level I Trauma Centre over a 10-year Period? CJEM 2016;19(2):106-11. doi: 10.1017/cem.2016.343	No	Yes	Yes
Gaudreau K, Sanford CJ, Cheverie C, et al. The Effect of a Smoking Ban on Hospitalization Rates for Cardiovascular and Respiratory Conditions in Prince Edward Island, Canada. PLoS ONE 2013;8(3):e56102. doi: 10.1371/journal.pone.0056102	No	No	Yes
Gebrehiwot TG, San Sebastian M, Edin K, et al. The Health Extension Program and Its Association with Change in Utilization of Selected Maternal Health Services in Tigray Region, Ethiopia: A Segmented Linear Regression Analysis. PLOS ONE 2015;10(7):e0131195. doi: 10.1371/journal.pone.0131195	Yes	N/A	Yes
Gefenaite G, Bijlsma M, Bos H, et al. Did introduction of pneumococcal vaccines in the Netherlands decrease the need for respiratory antibiotics in children? Analysis of 2002 to 2013 data. Eurosurveillance 2014;19(44):20948. doi: 10.2807/1560-7917.es2014.19.44.20948	No	No	No
Gilbert C, Darlow B, Zin A, et al. Educating Neonatal Nurses in Brazil: A Before-and-After Study with Interrupted Time Series Analysis. Neonatology 2014;106(3):201-08. doi: 10.1159/000362532	No	No	Yes
Glantz SA, Gibbs E. Changes in Ambulance Calls After Implementation of a Smoke-Free Law and Its Extension to Casinos. Circulation 2013;128(8):811-13. doi: 10.1161/circulationaha.113.003455	No	No	Yes
Gobin M, Verlander N, Maurici C, et al. Do sexual health campaigns work? An outcome evaluation of a media campaign to increase chlamydia testing among young people aged 15–24 in England. BMC Public Health 2013;13(1) doi: 10.1186/1471-2458-13-484	No	Yes	Yes
Godman B, Persson M, Miranda J, et al. Changes in the Utilization of Venlafaxine after the Introduction of Generics in Sweden. Applied Health Economics and Health Policy 2013;11(4):383-93. doi: 10.1007/s40258-013-0037-x	No	No	Yes

Godman B, Wettermark B, Miranda J, et al. Influence of multiple initiatives in Sweden to enhance ARB prescribing efficiency following generic losartan; findings and implications for other countries. International Journal of Clinical Practice 2013;67(9):853-62. doi: 10.1111/ijcp.12130	No	No	Yes
Gold R, Nelson C, Cowburn S, et al. Feasibility and impact of implementing a private care system's diabetes quality improvement intervention in the safety net: a cluster-randomized trial. Implementation Science 2015;10(1) doi: 10.1186/s13012-015-0259-4	No	No	Yes
Graves AJ, Kozhimannil KB, Kleinman KP, et al. The Association between High-Deductible Health Plan Transition and Contraception and Birth Rates. Health Services Research 2015;51(1):187-204. doi: 10.1111/1475-6773.12326	No	No	No
Guthrie B, Clark SA, Reynish EL, et al. Differential Impact of Two Risk Communications on Antipsychotic Prescribing to People with Dementia in Scotland: Segmented Regression Time Series Analysis 2001–2011. PLoS ONE 2013;8(7):e68976. doi: 10.1371/journal.pone.0068976	No	Yes	Yes
Haas JP, Menz J, Dusza S, et al. Implementation and impact of ultraviolet environmental disinfection in an acute care setting. American Journal of Infection Control 2014;42(6):586-90. doi: 10.1016/j.ajic.2013.12.013	No	No	Yes
Haggins A, Patrick S, Demonner S, et al. When Coverage Expands: Children's Health Insurance Program as a Natural Experiment in Use of Health Care Services. Academic Emergency Medicine 2013;20(10):1026-32. doi: 10.1111/acem.12236	No	No	Yes
Halim S, Jiang H. The effect of Operation 24 Hours on reducing collision in the City of Edmonton. Accident Analysis & Prevention 2013;58:106-14. doi: 10.1016/j.aap.2013.04.031	No	No	Yes
Hamilton I, Lloyd C, Bland JM, et al. The impact of assertive outreach teams on hospital admissions for psychosis: a time series analysis. Journal of Psychiatric and Mental Health Nursing 2015;22(7):484-90. doi: 10.1111/jpm.12239	No	No	Yes

Hanatani T, Sai K, Tohkin M, et al. Evaluation of two Japanese regulatory actions using medical information databases: a ‘Dear Doctor’ letter to restrict oseltamivir use in teenagers, and label change caution against co-administration of omeprazole with clopidogrel. Journal of Clinical Pharmacy and Therapeutics 2014;39(4):361-67. doi: 10.1111/jcpt.12153	No	No	Yes
Hansen BT, Østergaard SD, Sønderskov KM, et al. Increased Incidence Rate of Trauma- and Stressor-Related Disorders in Denmark After the September 11, 2001, Terrorist Attacks in the United States. American Journal of Epidemiology 2016;184(7):494-500. doi: 10.1093/aje/kww089	No	Yes	No
Hansen BT, Sønderskov KM, Hageman I, et al. Daylight Savings Time Transitions and the Incidence Rate of Unipolar Depressive Episodes. Epidemiology 2017;28(3):346-53. doi: 10.1097/ede.0000000000000580	Yes	N/A	Yes
Harper S, Bruckner TA. Did the Great Recession increase suicides in the USA? Evidence from an interrupted time-series analysis. Annals of Epidemiology 2017;27(7):409-14.e6. doi: 10.1016/j.annepidem.2017.05.017	No	Yes	No
Hartung DM, Middleton L, Markwardt S, et al. Changes in Long-acting β-agonist Utilization After the FDA’s 2010 Drug Safety Communication. Clinical Therapeutics 2015;37(1):114-23.e1. doi: 10.1016/j.clinthera.2014.10.025	No	No	Yes
Hassanian-Moghaddam H, Ghorbani F, Rahimi A, et al. Federation Internationale de Football Association (FIFA) 2014 World Cup Impact on Hospital-Treated Suicide Attempt (Overdose) in Tehran. Suicide and Life-Threatening Behavior 2017;48(3):367-75. doi: 10.1111/sltb.12359	No	Yes	Yes
Hawton K, Bergen H, Geulayov G, et al. Impact of the recent recession on self-harm: Longitudinal ecological and patient-level investigation from the Multicentre Study of Self-harm in England. Journal of Affective Disorders 2016;191:132-38. doi: 10.1016/j.jad.2015.11.001	No	No	Yes
Hawton K, Bergen H, Simkin S, et al. Long term effect of reduced pack sizes of paracetamol on poisoning deaths and liver transplant activity in England and Wales: interrupted time series analyses. BMJ 2013;346(feb07 1):f403-f03. doi: 10.1136/bmj.f403	No	No	Yes

Hingwala J, Bhangoo S, Hiebert B, et al. Evaluating the Implementation Strategy for Estimated Glomerular Filtration Rate Reporting in Manitoba: The Effect on Referral Numbers, Wait Times, and Appropriateness of Consults. Canadian Journal of Kidney Health and Disease 2014;1:9. doi: 10.1186/2054-3581-1-9	No	Yes	Yes
Høgli JU, Garcia BH, Skjold F, et al. An audit and feedback intervention study increased adherence to antibiotic prescribing guidelines at a Norwegian hospital. BMC Infectious Diseases 2016;16(1) doi: 10.1186/s12879-016-1426-1	No	No	Yes
Honein-AbouHaidar GN, Rabeneck L, Paszat LF, et al. Evaluating the impact of public health initiatives on trends in fecal occult blood test participation in Ontario. BMC Cancer 2014;14(1) doi: 10.1186/1471-2407-14-537	No	No	Yes
Horton DB, Gerhard T, Davidow A, et al. Impact of the black triangle label on prescribing of new drugs in the United Kingdom: lessons for the United States at a time of deregulation. Pharmacoepidemiology and Drug Safety 2017;26(11):1307-13. doi: 10.1002/pds.4304	No	No	No
Hostenkamp G, Fischer KE, Borch-Johnsen K. Drug safety and the impact of drug warnings: An interrupted time series analysis of diabetes drug prescriptions in Germany and Denmark. Health Policy 2016;120(12):1404-11. doi: 10.1016/j.healthpol.2016.09.020	No	No	No
Hsu JC, Cheng C-L, Ross-Degnan D, et al. Effects of safety warnings and risk management plan for Thiazolidinediones in Taiwan. Pharmacoepidemiology and Drug Safety 2015;24(10):1026-35. doi: 10.1002/pds.3834	No	Yes	Yes
Hsu JC, Lu CY, Wagner AK, et al. Impacts of drug reimbursement reductions on utilization and expenditures of oral antidiabetic medications in Taiwan: An interrupted time series study. Health Policy 2014;116(2-3):196-205. doi: 10.1016/j.healthpol.2013.11.005	No	Yes	Yes
Huitema BE, Van Houten R, Manal H. Time-series intervention analysis of pedestrian countdown timer effects. Accident Analysis & Prevention 2014;72:23-31. doi: 10.1016/j.aap.2014.05.025	No	No	No

Humphreys DK, Gasparrini A, Wiebe DJ. Evaluating the Impact of Florida’s “Stand Your Ground” Self-defense Law on Homicide and Suicide by Firearm. JAMA Internal Medicine 2017;177(1):44. doi: 10.1001/jamainternmed.2016.6811	No	Yes	Yes
Iams W, Heck J, Kapp M, et al. A Multidisciplinary Housestaff-Led Initiative to Safely Reduce Daily Laboratory Testing. Academic Medicine 2016;91(6):813-20. doi: 10.1097/acm.0000000000001149	No	No	Yes
Jenkins TC, Knepper BC, Shihadeh K, et al. Long-Term Outcomes of an Antimicrobial Stewardship Program Implemented in a Hospital with Low Baseline Antibiotic Use. Infection Control & Hospital Epidemiology 2015;36(6):664-72. doi: 10.1017/ice.2015.41	No	No	Yes
Jiao B, Kim S, Hagen J, et al. Cost-effectiveness of neighbourhood slow zones in New York City. Injury Prevention 2017;25(2):98-103. doi: 10.1136/injuryprev-2017-042499	Yes	N/A	No
Johri M, Ridde V, Heinmüller R, et al. Estimation of maternal and child mortality one year after user-fee elimination: an impact evaluation and modelling study in Burkina Faso. Bulletin of the World Health Organization 2014;92(10):706-15. doi: 10.2471/blt.13.130609	No	No	No
Katikireddi SV, Der G, Roberts C, et al. Has Childhood Smoking Reduced Following Smoke-Free Public Places Legislation? A Segmented Regression Analysis of Cross-Sectional UK School-Based Surveys. Nicotine & Tobacco Research 2016;18(7):1670-74. doi: 10.1093/ntr/ntw018	Yes	N/A	Yes
Kesselheim AS, Donneyong M, Dal Pan GJ, et al. Changes in prescribing and healthcare resource utilization after FDA Drug Safety Communications involving zolpidem-containing medications. Pharmacoepidemiology and Drug Safety 2017;26(6):712-21. doi: 10.1002/pds.4215	No	No	Yes
Kim B, Kim K, Lee J, et al. Impact of bacteremia prediction rule in CAP: Before and after study. The American Journal of Emergency Medicine 2018;36(5):758-62. doi: 10.1016/j.ajem.2017.10.005	No	No	Yes

Kim J-S, Chung YK, Lee SS, et al. Effect of daily chlorhexidine bathing on the acquisition of methicillin-resistant Staphylococcus aureus in a medical intensive care unit with methicillin-resistant S aureus endemicity. American Journal of Infection Control 2016;44(12):1520-25. doi: 10.1016/j.ajic.2016.04.252	No	No	Yes
Kim SH, Cho BL, Shin DW, et al. The Effect of Asthma Clinical Guideline for Adults on Inhaled Corticosteroids PrescriptionTrend: A Quasi-Experimental Study. Journal of Korean Medical Science 2015;30(8):1048. doi: 10.3346/jkms.2015.30.8.1048	No	No	Yes
Kiran T, Wilton AS, Moineddin R, et al. Effect of Payment Incentives on Cancer Screening in Ontario Primary Care. The Annals of Family Medicine 2014;12(4):317-23. doi: 10.1370/afm.1664	No	Yes	Yes
Kisely S, Crowe E, Lawrence D, et al. A time series analysis of presentations to Queensland health facilities for alcohol-related conditions, following the increase in ‘alcopops’ tax. Australasian Psychiatry 2013;21(4):383-88. doi: 10.1177/1039856213486307	No	No	Yes
Klein EG, Forster JL, Toomey TL, et al. Did a local clean indoor air policy increase alcohol-related crime around bars and restaurants? Tobacco Control 2011;22(2):113-17. doi: 10.1136/tobaccocontrol-2011-050010	No	No	No
Kolhatkar A, Cheng L, Chan FKI, et al. The impact of medication reviews by community pharmacists. Journal of the American Pharmacists Association 2016;56(5):513-20.e1. doi: 10.1016/j.japh.2016.05.002	No	Yes	Yes
Kontopantelis E, Olier I, Planner C, et al. Primary care consultation rates among people with and without severe mental illness: a UK cohort study using the Clinical Practice Research Datalink. BMJ Open 2015;5(12):e008650. doi: 10.1136/bmjopen-2015-008650	Yes	N/A	Yes
Kontopantelis E, Reeves D, Valderas JM, et al. Recorded quality of primary care for patients with diabetes in England before and after the introduction of a financial incentive scheme: a longitudinal observational study. BMJ Quality & Safety 2012;22(1):53-64. doi: 10.1136/bmjqs-2012-001033	No	Yes	Yes

Kracalik I, Abdullayev R, Asadov K, et al. Changing Patterns of Human Anthrax in Azerbaijan during the Post-Soviet and Preemptive Livestock Vaccination Eras. PLoS Neglected Tropical Diseases 2014;8(7):e2985. doi: 10.1371/journal.pntd.0002985	No	No	Yes
Kracalik IT, Abdullayev R, Asadov K, et al. Human Brucellosis Trends: Re-emergence and Prospects for Control Using a One Health Approach in Azerbaijan (1983-2009). Zoonoses and Public Health 2015;63(4):294-302. doi: 10.1111/zph.12229	No	No	Yes
Kruik-Kollöffel WJ, van der Palen J, Kruik HJ, et al. Prescription behavior for gastroprotective drugs in new users as a result of communications regarding clopidogrel - proton pump inhibitor interaction. Pharmacology Research & Perspectives 2016;4(4):e00242. doi: 10.1002/prp2.242	No	Yes	Yes
Larney S, Lai W, Dolan K, et al. Monitoring a Prison Opioid Treatment Program Over a Period of Change to Clinical Governance Arrangements, 2007–2013. Journal of Substance Abuse Treatment 2016;70:58-63. doi: 10.1016/j.jsat.2016.08.001	No	No	Yes
Lavergne MR, Law MR, Peterson S, et al. Effect of incentive payments on chronic disease management and health services use in British Columbia, Canada: Interrupted time series analysis. Health Policy 2018;122(2):157-64. doi: 10.1016/j.healthpol.2017.11.001	No	Yes	Yes
Lee KR, Bagga B, Arnold SR. Reduction of Broad-Spectrum Antimicrobial Use in a Tertiary Children's Hospital Post Antimicrobial Stewardship Program Guideline Implementation*. Pediatric Critical Care Medicine 2016;17(3):187-93. doi: 10.1097/pcc.0000000000000615	No	No	Yes
Lee TC, Frenette C, Jayaraman D, et al. Antibiotic Self-stewardship: Trainee-Led Structured Antibiotic Time-outs to Improve Antimicrobial Use. Annals of Internal Medicine 2014;161(10_Supplement):S53. doi: 10.7326/m13-3016	No	No	Yes
Lee Y-J, Chen J-Z, Lin H-C, et al. Impact of active screening for methicillin-resistant Staphylococcus aureus (MRSA) and decolonization on MRSA infections, mortality and medical cost: a quasi-experimental study in surgical intensive care unit. Critical Care 2015;19(1) doi: 10.1186/s13054-015-0876-y	No	Yes	Yes

Li Z, Li M, Fink G, et al. User–fee–removal improves equity of children's health care utilization and reduces families' financial burden: evidence from Jamaica. Journal of Global Health 2017;7(1) doi: 10.7189/jogh.07.010502	No	No	Yes
Lieberman DA, Polinski JM, Choudhry NK, et al. Unintended Consequences of a Medicaid Prescription Copayment Policy. Medical Care 2014;52(5):422-27. doi: 10.1097/mlr.0000000000000119	No	No	Yes
Lin CM, Liao CM. Inpatient expenditures on alcohol-attributed diseases and alcohol tax policy: a nationwide analysis in Taiwan from 1996 to 2010. Public Health 2014;128(11):977-84. doi: 10.1016/j.puhe.2014.09.004	No	Yes	Yes
Lopez Bernal J, Gasparrini A, Artundo C, et al. RE: The effect of the late 2000s financial crisis on suicides in Spain: an interrupted time-series analysis. The European Journal of Public Health 2014;24(2):183-84. doi: 10.1093/eurpub/ckt215	No	No	Yes
López-Ruiz M, Martínez JM, Pérez K, et al. Impact of road safety interventions on traffic-related occupational injuries in Spain, 2004–2010. Accident Analysis & Prevention 2014;66:114-19. doi: 10.1016/j.aap.2014.01.012	No	No	Yes
Lu CY, Zhang F, Lakoma MD, et al. Asthma Treatments and Mental Health Visits After a Food and Drug Administration Label Change for Leukotriene Inhibitors. Clinical Therapeutics 2015;37(6):1280-91. doi: 10.1016/j.clinthera.2015.03.027	No	No	Yes
Ma T, Byrne PA, Haya M, et al. Working in tandem: The contribution of remedial programs and roadside licence suspensions to drinking and driving deterrence in Ontario. Accident Analysis & Prevention 2015;85:248-56. doi: 10.1016/j.aap.2015.09.017	No	No	Yes
Maini R, Van den Bergh R, van Griensven J, et al. Picking up the bill - improving health-care utilisation in the Democratic Republic of Congo through user fee subsidisation: a before and after study. BMC Health Services Research 2014;14(1) doi: 10.1186/s12913-014-0504-6	No	Yes	Yes
Martin CL, Aldridge PJ, Harris AM, et al. Opening a New Level II Trauma Center Near an Established Level I Trauma Center. Journal of Orthopaedic Trauma 2016;30(10):517-23. doi: 10.1097/bot.0000000000000640	No	No	Yes

Marufu O, Desai N, Aldred D, et al. Analysis of interventions to reduce the incidence of Clostridium difficile infection at a London teaching hospital trust, 2003–2011. Journal of Hospital Infection 2015;89(1):38-45. doi: 10.1016/j.jhin.2014.10.003	No	No	Yes
Marwick CA, Guthrie B, Pringle JEC, et al. A multifaceted intervention to improve sepsis management in general hospital wards with evaluation using segmented regression of interrupted time series. BMJ Quality & Safety 2013;23(12):e2-e2. doi: 10.1136/bmjqs-2013-002176	No	Yes	Yes
McAlister FA, Bakal JA, Kaul P, et al. Changes in Heart Failure Outcomes After a Province-Wide Change in Health Service Provision A Natural Experiment in Alberta, Canada. Circulation: Heart Failure 2013;6(1):76-82. doi: 10.1161/circheartfailure.112.971119	No	No	No
McFarlane WR, Susser E, McCleary R, et al. Reduction in Incidence of Hospitalizations for Psychotic Episodes Through Early Identification and Intervention. Psychiatric Services 2014;65(10):1194-200. doi: 10.1176/appi.ps.201300336	No	Yes	No
McKirdy A, Imbuldeniya AM. The clinical and cost effectiveness of a virtual fracture clinic service. Bone & Joint Research 2017;6(5):259-69. doi: 10.1302/2046-3758.65.bjr-2017-0330.r1	No	No	Yes
McLeod A, Weir A, Aitken C, et al. Rise in testing and diagnosis associated with Scotland's Action Plan on Hepatitis C and introduction of dried blood spot testing. Journal of Epidemiology and Community Health 2014;68(12):1182-88. doi: 10.1136/jech-2014-204451	No	Yes	Yes
McLintock K, Russell AM, Alderson SL, et al. The effects of financial incentives for case finding for depression in patients with diabetes and coronary heart disease: interrupted time series analysis. BMJ Open 2014;4(8):e005178-e78. doi: 10.1136/bmjopen-2014-005178	No	No	No
McPhedran S, Mauser G. Lethal Firearm-Related Violence Against Canadian Women: Did Tightening Gun Laws Have an Impact on Women's Health and Safety? Violence and Victims 2013;28(5):875-83. doi: 10.1891/0886-6708.vv-d-12-00145	No	No	Yes

Mead EL, Cruz-Cano R, Bernat D, et al. Association between Florida's smoke-free policy and acute myocardial infarction by race: A time series analysis, 2000–2013. Preventive Medicine 2016;92:169-75. doi: 10.1016/j.ypmed.2016.05.032	No	No	Yes
Meirambayeva A, Vingilis E, Zou G, et al. Evaluation of Deterrent Impact of Ontario's Street Racing and Stunt Driving Law on Extreme Speeding Convictions. Traffic Injury Prevention 2014;15(8):786-93. doi: 10.1080/15389588.2014.890721	No	No	No
Mellon L, Hickey A, Doyle F, et al. Can a media campaign change health service use in a population with stroke symptoms? Examination of the first Irish stroke awareness campaign. Emergency Medicine Journal 2013;31(7):536-40. doi: 10.1136/emermed-2012-202280	No	No	Yes
Melvin KE, Hart JC, Sorvig RD. Second-Generation Antipsychotic Prescribing Patterns for Pediatric Patients Enrolled in West Virginia Medicaid. Psychiatric Services 2017;68(10):1061-67. doi: 10.1176/appi.ps.201600489	No	No	Yes
Milder EA, Rizzi MD, Morales KH, et al. Impact of a New Practice Guideline on Antibiotic Use With Pediatric Tonsillectomy. JAMA Otolaryngology–Head & Neck Surgery 2015;141(5):410. doi: 10.1001/jamaoto.2015.95	No	No	Yes
Miller P, Curtis A, Palmer D, et al. Changes in injury-related hospital emergency department presentations associated with the imposition of regulatory versus voluntary licensing conditions on licensed venues in two cities. Drug and Alcohol Review 2014;33(3):314-22. doi: 10.1111/dar.12118	No	No	Yes
Miwa S, Visintainer P, Engelman R, et al. Effects of an Ambulation Orderly Program Among Cardiac Surgery Patients. The American Journal of Medicine 2017;130(11):1306-12. doi: 10.1016/j.amjmed.2017.04.044	No	Yes	Yes
Moyo P, Simoni-Wastila L, Griffin BA, et al. Impact of prescription drug monitoring programs (PDMPs) on opioid utilization among Medicare beneficiaries in 10 US States. Addiction 2017;112(10):1784-96. doi: 10.1111/add.13860	No	Yes	Yes
Muoto I, Darney BG, Lau B, et al. Shifting Patterns in Cesarean Delivery Scheduling and Timing in Oregon before and after a Statewide Hard Stop Policy. Health Services Research 2017;53:2839-57. doi: 10.1111/1475-6773.12797	No	No	Yes

Myung W, Lee G-H, Won H-H, et al. Paraquat Prohibition and Change in the Suicide Rate and Methods in South Korea. PLOS ONE 2015;10(6):e0128980. doi: 10.1371/journal.pone.0128980	Yes	N/A	Yes
Nakahara S, Ichikawa M, Nakajima Y. Effects of Increasing Child Restraint Use in Reducing Occupant Injuries Among Children Aged 0–5 Years in Japan. Traffic Injury Prevention 2014;16(1):55-61. doi: 10.1080/15389588.2014.897698	No	No	Yes
Narayan H, Thomas SHL, Eddleston M, et al. Disproportionate effect on child admissions of the change in Medicines and Healthcare Products Regulatory Agency guidance for management of paracetamol poisoning: an analysis of hospital admissions for paracetamol overdose in England and Scotland. British Journal of Clinical Pharmacology 2015;80(6):1458-63. doi: 10.1111/bcp.12779	Yes	N/A	Yes
Nazif-Munoz JI, Quesnel-Vallée A, van den Berg A. Did Chile's traffic law reform push police enforcement? Understanding Chile's traffic fatalities and injuries reduction. Injury Prevention 2014;21(3):159-65. doi: 10.1136/injuryprev-2014-041358	No	Yes	Yes
Newitt S, Myles PR, Birkin JA, et al. Impact of infection control interventions on rates of Staphylococcus aureus bacteraemia in National Health Service acute hospitals, East Midlands, UK, using interrupted time-series analysis. Journal of Hospital Infection 2015;90(1):28-37. doi: 10.1016/j.jhin.2014.12.016	No	No	Yes
Nistal-Nuño B. Segmented regression analysis of interrupted time series data to assess outcomes of a South American road traffic alcohol policy change. Public Health 2017;150:51-59. doi: 10.1016/j.puhe.2017.04.025	No	No	Yes
Norstrom T, Stickley A. Alcohol tax, consumption and mortality in tsarist Russia: is a public health perspective applicable? The European Journal of Public Health 2012;23(2):340-44. doi: 10.1093/eurpub/cks079	No	No	Yes
O'Brien NP, Foss RD, Goodwin AH, et al. Supervised hours requirements in graduated driver licensing: Effectiveness and parental awareness. Accident Analysis & Prevention 2013;50:330-35. doi: 10.1016/j.aap.2012.05.007	No	No	Yes

Okasha O, Rinta-Kokko H, Palmu AA, et al. Population-level impact of infant 10-valent pneumococcal conjugate vaccination on adult pneumonia hospitalisations in Finland. Thorax 2017;73(3):262-69. doi: 10.1136/thoraxjnl-2017-210440	No	No	No
Osman M, Parnell AC. Effect of the First World War on suicide rates in Ireland: an investigation of the 1864–1921 suicide trends. BJPsych Open 2015;1(2):164-65. doi: 10.1192/bjpo.bp.115.000539	No	Yes	Yes
Ostrowsky B, Sharma S, DeFino M, et al. Antimicrobial Stewardship and Automated Pharmacy Technology Improve Antibiotic Appropriateness for Community-Acquired Pneumonia. Infection Control & Hospital Epidemiology 2013;34(6):566-72. doi: 10.1086/670623	No	No	Yes
Owens CL, Peterson D, Kamineni A, et al. Effects of transitioning from conventional methods to liquid-based methods on unsatisfactory Papanicolaou tests. Cancer Cytopathology 2013;121(10):568-75. doi: 10.1002/cncy.21309	No	No	Yes
Pace LE, Dusetzina SB, Keating NL. Early Impact of the Affordable Care Act on Uptake of Long-acting Reversible Contraceptive Methods. Medical Care 2016;54(9):811-17. doi: 10.1097/mlr.0000000000000551	No	No	Yes
Pan SW, Chong HH, Kao H-C. Unintentional injury mortality among indigenous communities of Taiwan: trends from 2002 to 2013 and evaluation of a community-based intervention. Injury Prevention 2017;25(1):26-30. doi: 10.1136/injuryprev-2017-042321	No	Yes	Yes
Panagiotoglou D, Law MR, McGrail K. Effect of Hospital Closures on Acute Care Outcomes in British Columbia, Canada. Medical Care 2017;55(1):50-56. doi: 10.1097/mlr.0000000000000619	No	No	Yes
Panatto D, Domnich A, Gasparini R, et al. An eHealth Project on Invasive Pneumococcal Disease: Comprehensive Evaluation of a Promotional Campaign. Journal of Medical Internet Research 2016;18(12):e316. doi: 10.2196/jmir.6205	No	Yes	No
Parikh K, Hall M, Teach SJ. Bronchiolitis Management Before and After the AAP Guidelines. Pediatrics 2013;133(1):e1-e7. doi: 10.1542/peds.2013-2005	No	No	No

Patel PR, Yi SH, Booth S, et al. Bloodstream Infection Rates in Outpatient Hemodialysis Facilities Participating in a Collaborative Prevention Effort: A Quality Improvement Report. American Journal of Kidney Diseases 2013;62(2):322-30. doi: 10.1053/j.ajkd.2013.03.011	No	No	Yes
Pegues DA, Han J, Gilmar C, et al. Impact of Ultraviolet Germicidal Irradiation for No-Touch Terminal Room Disinfection on Clostridium difficile Infection Incidence Among Hematology-Oncology Patients. Infection Control & Hospital Epidemiology 2016;38(1):39-44. doi: 10.1017/ice.2016.222	No	Yes	Yes
Pellegrin KL, Krenk L, Oakes SJ, et al. Reductions in Medication-Related Hospitalizations in Older Adults with Medication Management by Hospital and Community Pharmacists: A Quasi-Experimental Study. Journal of the American Geriatrics Society 2016;65(1):212-19. doi: 10.1111/jgs.14518	No	No	Yes
Petereit D, Omidpanah A, Boylan A, et al. A Multi-faceted Approach to Improving Breast Cancer Outcomes in a Rural Population, and the Potential Impact of Patient Navigation. South Dakota medicine : the journal of the South Dakota State Medical Association 2016;69(6):268-73. [published Online First: 2016/07/23]	No	No	Yes
Petrou P. The Ariadne's thread in co-payment, primary health care usage and financial crisis: findings from Cyprus public health care sector. Public Health 2015;129(11):1503-09. doi: 10.1016/j.puhe.2015.07.032	No	No	Yes
Pinheiro SP, Kang EM, Kim CY, et al. Concomitant use of isotretinoin and contraceptives before and after iPledge in the United States. Pharmacoepidemiology and Drug Safety 2013;22(12):1251-57. doi: 10.1002/pds.3481	No	No	Yes
Poluzzi E, Veronese G, Piccinni C, et al. Switching among Equivalents in Chronic Cardiovascular Therapies: 'Real World' Data from Italy. Basic & Clinical Pharmacology & Toxicology 2015;118(1):63-69. doi: 10.1111/bcpt.12442	No	Yes	Yes
Pradhan A, Anasuya A, Pradhan MM, et al. Trends in Malaria in Odisha, India—An Analysis of the 2003–2013 Time-Series Data from the National Vector Borne Disease Control Program. PLOS ONE 2016;11(2):e0149126. doi: 10.1371/journal.pone.0149126	No	No	No

Pridemore WA, Chamlin MB, Andreev E. Reduction in Male Suicide Mortality Following the 2006 Russian Alcohol Policy: An Interrupted Time Series Analysis. American Journal of Public Health 2013;103(11):2021-26. doi: 10.2105/ajph.2013.301405	No	No	Yes
Prinja S, Kaur G, Gupta R, et al. Out-of-pocket expenditure for health care: District level estimates for Haryana state in India. The International Journal of Health Planning and Management 2018;34(1):277-93. doi: 10.1002/hpm.2628	No	No	Yes
Puig-Junoy J, Rodríguez-Feijoó S, Lopez-Valcarcel BG. Paying for Formerly Free Medicines in Spain After 1 Year of Co-Payment: Changes in the Number of Dispensed Prescriptions. Applied Health Economics and Health Policy 2014;12(3):279-87. doi: 10.1007/s40258-014-0097-6	No	Yes	No
Pun VC, Lin H, Kim JH, et al. Impacts of alcohol duty reductions on cardiovascular mortality among elderly Chinese: a 10-year time series analysis. Journal of Epidemiology and Community Health 2013;67(6):514-18. doi: 10.1136/jech-2012-201859	No	No	Yes
Rhodes D, Cheng AC, McLellan S, et al. Reducing Staphylococcus aureus bloodstream infections associated with peripheral intravenous cannulae: successful implementation of a care bundle at a large Australian health service. Journal of Hospital Infection 2016;94(1):86-91. doi: 10.1016/j.jhin.2016.05.020	No	No	No
Rooholamini SN, Clifton H, Haaland W, et al. Outcomes of a Clinical Pathway to Standardize Use of Maintenance Intravenous Fluids. Hospital Pediatrics 2017;7(12):703-09. doi: 10.1542/hpeds.2017-0099	No	Yes	Yes
Rosenthal MB, Friedberg MW, Singer SJ, et al. Effect of a Multipayer Patient-Centered Medical Home on Health Care Utilization and Quality. JAMA Internal Medicine 2013;173(20):1907. doi: 10.1001/jamainternmed.2013.10063	No	Yes	Yes
Rutman L, Wright DR, O’Callaghan J, et al. A Comprehensive Approach to Pediatric Pneumonia. Journal for Healthcare Quality 2017;39(4):e59-e69. doi: 10.1097/jhq.0000000000000048	No	No	Yes

Ryu S, Lau CL, Chun BC. The impact of Livestock Manure Control Policy on human leptospirosis in Republic of Korea using interrupted time series analysis. Epidemiology and Infection 2017;145(7):1320-25. doi: 10.1017/s0950268817000218	No	No	Yes
Sakai R, Wang W, Yamaguchi N, et al. The Impact of Japan's 2004 Postgraduate Training Program on Intra-Prefectural Distribution of Pediatricians in Japan. PLoS ONE 2013;8(10):e77045. doi: 10.1371/journal.pone.0077045	No	No	Yes
Santa-Ana-Tellez Y, Mantel-Teeuwisse AK, Dreser A, et al. Impact of Over-the-Counter Restrictions on Antibiotic Consumption in Brazil and Mexico. PLoS ONE 2013;8(10):e75550. doi: 10.1371/journal.pone.0075550	No	No	Yes
Santa-Ana-Tellez Y, Mantel-Teeuwisse AK, Leufkens HGM, et al. Seasonal Variation in Penicillin Use in Mexico and Brazil: Analysis of the Impact of Over-the-Counter Restrictions. Antimicrobial Agents and Chemotherapy 2014;59(1):105-10. doi: 10.1128/aac.03629-14	No	No	Yes
Scherb HH, Mori K, Hayashi K. Increases in perinatal mortality in prefectures contaminated by the Fukushima nuclear power plant accident in Japan. Medicine 2016;95(38):e4958. doi: 10.1097/md.0000000000004958	No	No	Yes
Sicsic J, Saint-Lary O, Rouveix E, et al. Impact of a primary care national policy on HIV screening in France: a longitudinal analysis between 2006 and 2013. British Journal of General Practice 2016;66(653):e920-e29. doi: 10.3399/bjgp16x687529	No	Yes	Yes
Singh K, Speizer I, Handa S, et al. Impact evaluation of a quality improvement intervention on maternal and child health outcomes in Northern Ghana: early assessment of a national scale-up project. International Journal for Quality in Health Care 2013;25(5):477-87. doi: 10.1093/intqhc/mzt054	No	No	Yes
Sinnott S-J, Normand C, Byrne S, et al. Copayments for prescription medicines on a public health insurance scheme in Ireland. Pharmacoepidemiology and Drug Safety 2015;25(6):695-704. doi: 10.1002/pds.3917	No	No	No

Slattery C, Freund M, Gillham K, et al. Increasing smoking cessation care across a network of hospitals: an implementation study. Implementation Science 2015;11(1) doi: 10.1186/s13012-016-0390-x	No	No	Yes
Smith RL, Hayashi VN, Lee YI, et al. The Medical Emergency Team Call. Critical Care Medicine 2014;42(2):322-27. doi: 10.1097/ccm.0b013e3182a27413	No	No	Yes
Snowden LR, Wallace N, Cordell K, Graaf G. Increased mental health treatment financing, community-based organization’s treatment programs, and Latino-White children’s financing disparities. Journal of Mental Health Policy and Economics 2017;20:137-145.	No	No	No
Staras SAS, Livingston MD, Christou AM, et al. Heterogeneous population effects of an alcohol excise tax increase on sexually transmitted infections morbidity. Addiction 2014;109(6):904-12. doi: 10.1111/add.12493	No	Yes	Yes
Stelfox HT, Bastos J, Niven DJ, et al. Critical care transition programs and the risk of readmission or death after discharge from ICU. Intensive Care Medicine 2016;42(3):401-10. doi: 10.1007/s00134-015-4173-7	No	No	Yes
Taber DJ, DuBay D, McGillicuddy JW, et al. Impact of the New Kidney Allocation System on Perioperative Outcomes and Costs in Kidney Transplantation. Journal of the American College of Surgeons 2017;224(4):585-92. doi: 10.1016/j.jamcollsurg.2016.12.009	No	Yes	Yes
Thijssen WAMH, Wijnen-van Houts M, Koetsenruijter J, et al. The Impact on Emergency Department Utilization and Patient Flows after Integrating with a General Practitioner Cooperative: An Observational Study. Emergency Medicine International 2013;2013:1-8. doi: 10.1155/2013/364659	No	Yes	Yes
Tiwari A, Osbert N, Matimelo SM, et al. Assessing the Impact of Leveraging Traditional Leadership on Access to Sanitation in Rural Zambia. The American Journal of Tropical Medicine and Hygiene 2017;97(5):1355-61. doi: 10.4269/ajtmh.16-0612	No	Yes	No
Troelstra S, Bosdriesz J, de Boer M, et al. Effect of tobacco control policies on information seeking for smoking cessation in the Netherlands: A Google Trends study. European Journal of Public Health 2014;24(suppl_2) doi: 10.1093/eurpub/cku164.043	No	Yes	No

Ullman M, Parlier G, Warren J, et al. The Economic Impact of Starting, Stopping, and Restarting an Antibiotic Stewardship Program: A 14-Year Experience. Antibiotics 2013;2(2):256-64. doi: 10.3390/antibiotics2020256	No	No	No
Yarnell CJ, Shadowitz S, Redelmeier DA. Hospital Readmissions Following Physician Call System Change: A Comparison of Concentrated and Distributed Schedules. The American Journal of Medicine 2016;129(7):706-14.e2. doi: 10.1016/j.amjmed.2016.02.022	No	Yes	Yes
Zhao A, Chen R, Qi Y, et al. Evaluating the Impact of Criminalizing Drunk Driving on Road-Traffic Injuries in Guangzhou, China: A Time-Series Study. Journal of Epidemiology 2016;26(8):433-39. doi: 10.2188/jea.je20140103	No	No	No