

Effect estimates can be accurately calculated with data digitally extracted from interrupted time series graphs

Simon Lee Turner  | Elizabeth Korevaar  | Miranda S. Cumpston  |
Raju Kanukula  | Andrew B. Forbes  | Joanne E. McKenzie 

School of Public Health and Preventive
Medicine, Monash University, Melbourne,
Australia

Correspondence

Simon Lee Turner, School of Public
Health and Preventive Medicine, Monash
University, Melbourne, 3004 VIC,
Australia.

Email: simon.turner@monash.edu

Funding information

Australian Government Research Training
Program; National Health and Medical
Research Council, Grant/Award Number:
GNT2009612

Abstract

Interrupted time series (ITS) studies are frequently used to examine the impact of population-level interventions or exposures. Systematic reviews with meta-analyses including ITS designs may inform public health and policy decision-making. Re-analysis of ITS may be required for inclusion in meta-analysis. While publications of ITS rarely provide raw data for re-analysis, graphs are often included, from which time series data can be digitally extracted. However, the accuracy of effect estimates calculated from data digitally extracted from ITS graphs is currently unknown. Forty-three ITS with available datasets and time series graphs were included. Time series data from each graph was extracted by four researchers using digital data extraction software. Data extraction errors were analysed. Segmented linear regression models were fitted to the extracted and provided datasets, from which estimates of immediate level and slope change (and associated statistics) were calculated and compared across the datasets. Although there were some data extraction errors of time points, primarily due to complications in the original graphs, they did not translate into important differences in estimates of interruption effects (and associated statistics). Using digital data extraction to obtain data from ITS graphs should be considered in reviews including ITS. Including these studies in meta-analyses, even with slight inaccuracy, is likely to outweigh the loss of information from non-inclusion.

KEYWORDS

digital data extraction, interrupted time series, meta-analysis, public health

Highlights

What is already known

Interrupted time series (ITS) studies are a non-randomised study design often used to examine the effects of population-level interventions and may be included in systematic reviews. Inclusion of ITS studies in meta-analyses may require re-analysis of the data, however, time series data are rarely reported in

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Research Synthesis Methods* published by John Wiley & Sons Ltd.

publications. Fortunately, most publications include graphs, from which time series data can be digitally extracted. Digital data extraction from scatter plots has been shown to be accurate, but ITS graphs are often line plots, and the accuracy of digital extraction from these is unknown.

What is new

We found that data extracted digitally from 43 ITS graphs, each by four researchers, resulted in some data extraction errors. However, when the data were analysed using segmented linear regression, these errors did not lead to any important differences in the ensuing effect estimates, their confidence intervals and *p*-values. Therefore, use of digital data extraction should be considered for ITS graphs to maximise the inclusion of such studies in meta-analysis.

Potential impact for research synthesis methods readers outside the authors' field

A methods review examining the characteristics of reviews that include ITS (Korevaar et al, *J Clin Epidemiol*, 2022;145:55–69), found that such reviews are undertaken in a range of disciplines and topics, including public health, crime, economics, war and psychology. Therefore, the findings from this research are likely to have impact across disciplines.

1 | INTRODUCTION

Interrupted time series (ITS) studies are commonly used to assess the effects of population-level public health and policy interventions or exposures such as natural disasters or pandemics (henceforth jointly referred to as 'interruptions').^{1–8} Systematic reviews examining the effects of interruptions targeted at a population level may need to include study designs beyond randomised trials, such as when randomisation is difficult or impossible (e.g., examining the effects of a policy change to an entire country), to provide evidence for the question of interest.^{9,10} In these reviews, ITS designs are often included because of their potential to minimise bias compared with other non-randomised experimental designs.^{1,3,5,8,11} Meta-analysis of results from the included ITS designs can usefully inform decision making by yielding summary effect estimates, along with an understanding of the extent of inconsistency of the effects across studies, and what the likely effect of the interruption would be in an individual setting.^{12,13}

Undertaking a meta-analysis of results from ITS studies requires the use of a consistent effect measure (e.g., immediate level change) across the studies; the use of statistical methods that appropriately account for correlation arising from time series data (known as autocorrelation); and, complete reporting of the study effect estimate along with a measure of precision (e.g., confidence interval, standard error).^{14,15} When

these criteria are not met for a particular ITS study, it will likely be excluded from the meta-analysis, resulting in a loss of information. Re-analysing the time series data could overcome the aforementioned issues by ensuring consistent effect measures, the use of appropriate statistical methods and the necessary statistics for meta-analysis; thus, facilitating inclusion of most (if not all) the available ITS studies in a meta-analysis. However, this relies on the time series data (i.e., measurements of the outcome of interest at each time point in the series) being publicly available, which is rare. Fortunately, publications of ITS studies often include a graph,^{16–18} making extraction of time series data possible (see Figure 1 e.g., References 19,20).

Digital data extraction has been shown to be accurate in several studies,^{21–24} and is recommended by the Cochrane Handbook for Systematic Reviews of Interventions when the data are not available.²⁵ However, the focus of studies examining digital data extraction to date has been to investigate the accuracy and precision of data digitally extracted from scatter plots, while graphs of ITS are frequently line plots (see Figure 1b for an example).¹⁵ Additionally, the quality of ITS graphs is not always good (e.g., line plots may not include individual data points, or axis tick marks may not align with data points), potentially hampering accurate extraction of data and the accuracy of the ensuing effect estimates.¹⁵

To our knowledge, no study has examined the accuracy of interruption effect estimates calculated from

digitally extracted ITS data. Our aim was therefore to compare effect estimates (immediate level change and slope change, and their standard errors, confidence

intervals and *p*-values) calculated from digitally extracted ITS data with those calculated from provided datasets.

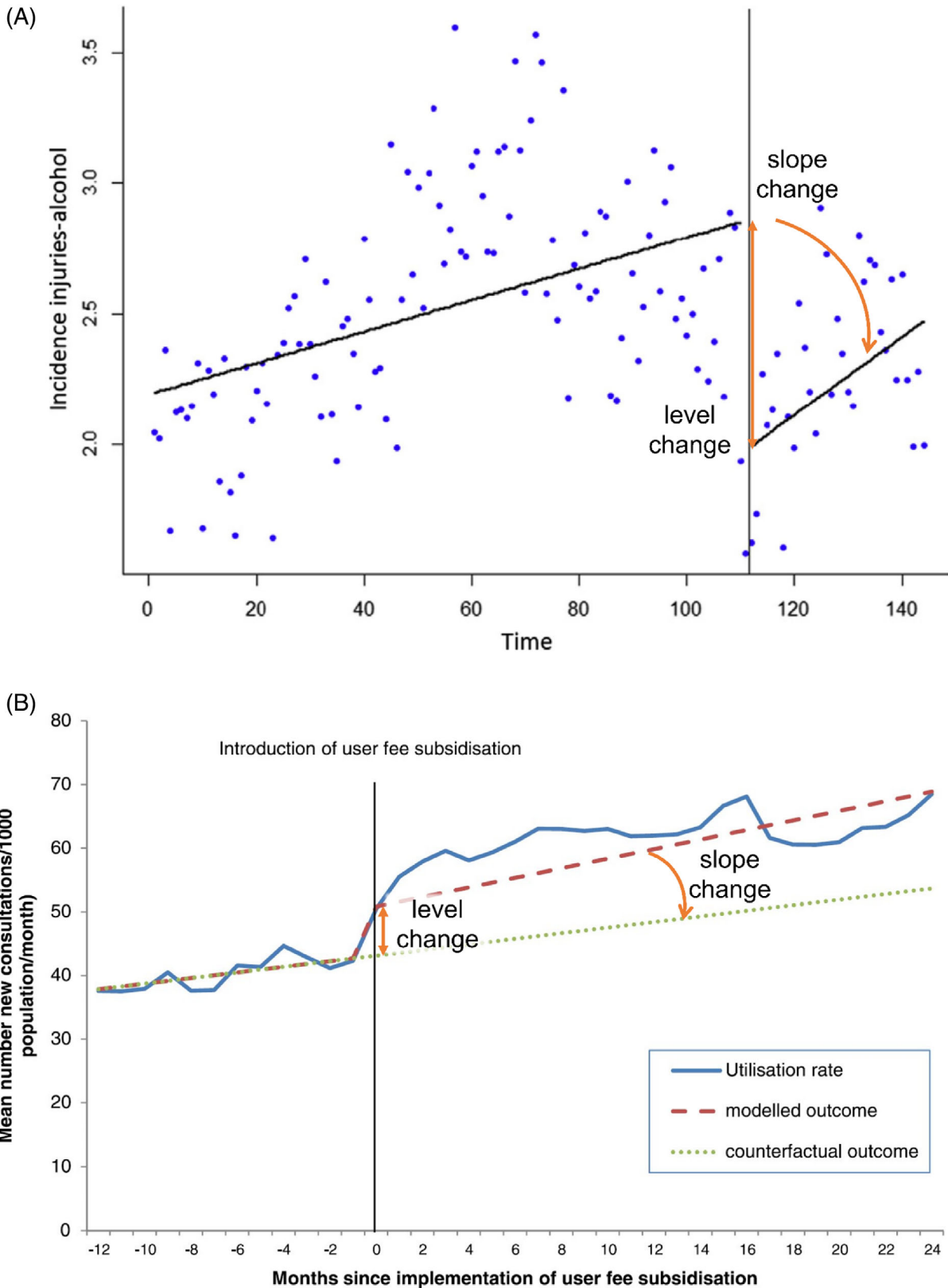


FIGURE 1 Legend on next page.

2 | METHODS

2.1 | Overview of the methods

We identified a cohort of ITS studies with available datasets and time series graphs. Four authors extracted data from the graphs using digital data extraction software. The quality of the graphs was assessed against ITS graphing recommendations. We analysed the errors made in the data extraction. We fitted segmented linear regression models to the extracted and provided datasets, calculated commonly used interruption effect estimates (along with their standard errors, confidence intervals and *p*-values), and compared these across the datasets.

2.2 | Retrieval of data series and graphs

In a previous methods study examining the design characteristics and statistical methods used in ITS studies evaluating public health interventions,²⁶ we identified 200 ITS studies via random sampling (details available in Turner et al¹⁸ and Appendix S2). Each ITS study could contribute multiple series if they reported outcomes that fell into more than one outcome category (i.e., binary, continuous, count). Details of the process for selecting one outcome per outcome category are available in Turner et al¹⁸ and Appendix S2. Although the original sample of studies may have included multiple series per study, for the present study, we included only one series per study by selecting the first reported outcome.

We attempted to retrieve the time series data through extraction of data from tables, Supporting Information or via email requests to authors (details available in Turner et al²⁶). Information on the time series segment lengths and interruption time (or times) were obtained from the manuscript and included graphs. For the present study we included ITS for which the time series data were available as well as an accompanying graph. The ITS graphs were extracted from portable document format (PDF) versions of the manuscripts using the snapshot tool and saved in joint photographic experts group (JPEG) format.

2.3 | Digital data extraction

2.3.1 | Data extractors, software and training

Four authors (E.K., M.S.C., R.K. and S.L.T.), with varying experience, undertook the digital data extraction. We selected the digital data extraction tool WebPlotDigitizer²⁷ as it has been shown in previous studies to be accurate in estimating data point positions on graphs.^{21–24} To provide all extractors with similar knowledge, given their differing prior experience, S.L.T. developed written and video documentation to provide guidance on how to extract data using WebPlotDigitizer (Appendix S1). This training covered how to use WebPlotDigitizer to import the images of the ITS graphs, accurately extract the data points and save the resulting data file for analysis (further details below). In addition, resources from the software developer were provided.²⁷ Extractors were asked to read through the written documentation and watch the video prior to practicing the digital data extraction from two graphs. The two practice graphs were chosen based on attributes that allowed for one easy data extraction (i.e., had a small number of clearly defined data points with clearly marked axes) or one more difficult data extraction (i.e., a line graph with no clearly defined data points and an x-axis involving dates). The training process (reading the documentation, watching the video, and extracting the data from the two graphs) took approximately 1 h. Following the training process, S.L.T. met with the extractors, provided feedback, and answered any further questions, prior to them commencing digital data extraction from the remaining graphs.

2.3.2 | Data extraction process

Each graph was given to each extractor. The order in which data extraction of the graphs was to be undertaken was randomly assigned for each extractor to account for any order effects that may occur (e.g., learning effects or fatigue). This was implemented by providing the extractors with a spreadsheet that included the list of graphs in

FIGURE 1 Examples of interrupted time series graphs. (a) shows incidence rate of injuries related to alcohol per 100,000 inhabitants over time (months) before (left) and after (right) the implementation of a law decreasing the legal blood alcohol limit for driving in Chile, South America. (b) shows the effects of user fee subsidisation on mean health-care utilisation rate 12 months prior and 24 months following their introduction for 16 health zones of the Democratic Republic of Congo (2008–2012). Level change and slope change labels and indications (orange arrows) have been added for clarity. (a) Reprinted from Public Health, 150, Nistal-Nuño B, “Segmented regression analysis of interrupted time series data to assess outcomes of a South American road traffic alcohol policy change”, 51–59, Copyright (2017), with permission from Elsevier, licence number 5376221132966. (b) Reprinted from BMC Health Services Research, 14:504, Maini 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,” Copyright (2014), under the terms of the Creative Commons Attribution Licence 2.0 (<http://creativecommons.org/licenses/by/2.0>). [Colour figure can be viewed at wileyonlinelibrary.com]

random order. Extractors were asked to follow the process outlined Table 1 (provided in detail in the guidance documentation; Appendix S1) and record any notes regarding issues they had in extracting the data (e.g., if a data point was missing). The time taken to extract the data points for each graph was recorded by one extractor (S.L.T.); this information was not recorded for all extractors because it was not a focus of the study.

Several unanticipated issues arose during the data extraction. Approaches for dealing with these were discussed and agreed at team meetings (S.L.T., E.K., A.B.F., J.E.M.). Our driver for choosing a particular approach was that it would reflect the likely approach chosen in practice. The issues (italicised) and agreed approaches follow.

- *Inadvertent extraction of data from the wrong series in graphs that included multiple series.* In these instances, we asked the extractor to extract the data from the correct series.
- *Data points missed during the extraction.* In these instances, we did not ask the extractors to re-extract the data, but instead assumed these values were missing.
- *Multiple outcome values assigned to the same time point (sometimes arising due to rounding to the nearest time point in the software).* In these instances, the first outcome value extracted was used for the time point, and the second data point was dropped from the analysis.

2.3.3 | Assessment of the quality of graphs

We assessed the quality of the included graphs by examining whether they met a subset of the core graphing recommendations for ITS proposed by Turner et al.¹⁵ The

TABLE 1 Digital data extraction process summary.

Using WebPlotDigitizer	
1	Select the required graph image file
2	Choose the 2D (X-Y) plot type
3	Align the X and Y axes by clicking on the leftmost x-axis tickmark, rightmost x-axis tickmark, lowest y-axis tickmark and highest y-axis tickmark
4	Extract the data points by moving the cursor target to the centre of each data point and left clicking. When a line graph is plotted without data points, use the cursor coordinates in the WebPlotDigitizer zoom window and left click on the line at the correct x-axis position
5	Save the data in comma separated values (CSV) format (two columns of x- and y-coordinate pairs)

recommendations chosen are those required for accurate data extraction. Specifically, these include: distinct individual points plotted, tick marks on the x-axis, data points that aligned with the tick marks on the x-axis (alignment of data points with tick marks on the x axis allows identification of the time period corresponding to the data point), tick marks on the y-axis, and y-axis labels that aligned with tick marks.

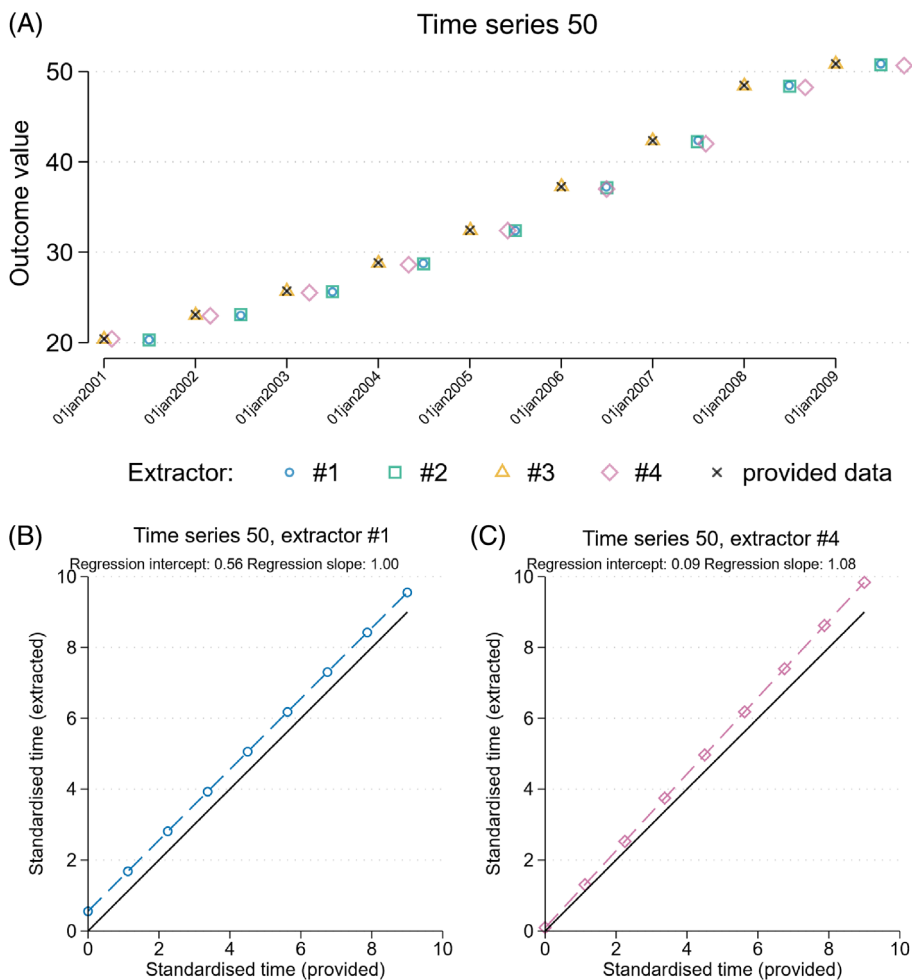
2.3.4 | Analysis of the data extraction errors

We undertook analyses to quantify the extent of error in extraction of the x- and y- coordinates of the points, and the extent of missing data points. To quantify errors in the extracted time points (x-coordinates) of the data points, an ordinary least squares (OLS) regression of the extracted time points (x-coordinates) versus the provided time points was undertaken (Figure 2). These analyses were undertaken separately by extractor and for each data series. The intercept of the regression line provides an estimate of the difference between the extracted and provided time points (which can potentially arise when the data points are not lined up with the axis tick marks). When there is agreement between the extracted and provided time points, the estimated intercept will be (close to) zero. The slope of the regression line provides an indication of whether the extracted time points became increasingly (estimated slope >1), or decreasingly (estimated slope <1), discordant compared with the provided data over the series (which can occur when the x-axis is not defined correctly by the extractor). When there is no change in the error across the series, the estimated slope will be (close to) 1.

We standardised the estimates of intercept and slope within each time series by dividing by the length of the provided series (e.g., 2 years) and multiplying by the number of time periods (e.g., 24, when data are collected monthly). This yields differences between extracted and provided times in fractions of time periods. If the extracted data time points are more than half a time period different (a standardised difference of ≥ 0.5) to those in the provided data, the time of the interruption may be incorrect, and this may impact the interruption effect estimates. For this reason, we have denoted differences of greater than or equal to 0.5 as “important errors.” Within extractor, we then calculated the means (with standard deviations) and medians (with interquartile ranges) of the standardised intercept and slope estimates across the ITS studies.

We undertook the same analyses for the y-coordinates. However, we standardised estimates of intercept and slope within each time series by dividing by the range of the

FIGURE 2 Example demonstrating how different types of errors in the extraction of time points (x-coordinates) from time series plots can be detected via linear regression. (a) Scatterplot of provided (black × symbol) and extracted data from four extractors (represented by different coloured symbols). The blue circle (and green square) represents extracted data points that are consistently half of a time unit greater than the provided data. The purple diamond represents extracted data that is close to the correct time at the start of the series, but with increasingly large error over the series (see c) for how linear regression can be used to quantify this error). (b) and (c) demonstrate how linear regression can be used to quantify these errors. (b) Scatterplot of x-coordinates extracted by one extractor (#1 indicated by blue circles) against the provided data. The extracted data is consistently half of a time unit greater than the provided data. This is reflected in the regression intercept estimate of 0.56. (c) Scatterplot of x-coordinates extracted by one extractor (#4 indicated by purple diamonds) against the provided data. The extracted data, which is close to the provided data at the start of the series, has increasingly larger error over the series. This is reflected in the regression slope estimate of 1.08. [Colour figure can be viewed at wileyonlinelibrary.com]



provided time series so that the standardised effects were between 0 and 1. This allowed us to make comparisons across the ITS studies.

We quantified the extent of missing data points by coding, for each time series and extractor, whether the extracted time series had fewer data points than the provided time series. We then summarised, by extractor, the number of series for which there was at least one missing data point.

2.4 | Analysis of interrupted time series

The CSV data from the extractors was imported into Stata version 17²⁸ for analysis by S.L.T. Segmented linear regression models were fitted, as they are frequently used in practice to analyse ITS studies.^{14,16–18,29,30} The models

were fitted (using the parameterisation of Huitema and McKean^{31,32}) as:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 D_t + \beta_3 [t - T_1] D_t + \varepsilon_t, \quad (1)$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + w_t.$$

Here, Y_t represents the outcome measured at time point t ; β_0 represents the intercept with the y-axis; β_1 represents the slope of the pre-interruption line segment; and, β_2 represents the immediate change in level between the pre- and post-interruption line segments, which are defined by D_t , an indicator variable that is 0 before the interruption time (T_1) and 1 after. β_3 represents the change in slope between the pre- and post-interruption line segments (Figure 1). Finally, ε_t represents the error term, allowing for deviations from the fitted model. It is

TABLE 2 Prevalence of adherence to graphing recommendations for graphs included in the present study and the original study.

Graph characteristic	Present study (N = 43)*		Original study** (N = 217)*	
	n	%	n	%
Distinct individual points plotted	26	60	130	60
Tick marks on the x-axis	38	88	195	90
Data points align with tick marks on the x-axis (n = 38,195)	23	61	109	56
Tick marks on the y-axis	43	100	217	100
Labels aligned with tick marks on the y-axis	43	100	217	100

*N = number of ITS graphs.

**Turner 2020.¹⁵

common in time series for data points to be correlated over time,³³ and *model 1* represents a process whereby the error is assumed to only be influenced by the previous time point (lag-1 autocorrelation), where ρ is the magnitude of the autocorrelation (ranging from -1 to 1) and w_t represents “white noise,” which we assume to be normally distributed $w_t \sim N(0, \sigma^2)$. While modelling of longer lags is possible, in this paper we restrict our attention to lag-1 autocorrelation.

We analysed each time series using restricted maximum likelihood (REML), allowing for lag-1 autocorrelation. In instances where REML failed to converge for at least one of the provided or extracted datasets associated with a particular time series, we analysed all the associated series using OLS regression.

Selection of the interruption time was obtained from the papers. If the number of time points pre- and post-interruption were reported, we used this information in setting up the ITS model; however, if only the date of the interruption was reported, we used this information. These options have different consequences when there are misalignments between the extracted time points and provided time points. The first option leads to the same model as if there was no misalignment, while the second option will lead to a model where the time of the interruption differs. In series with multiple interruptions, only the first interruption was analysed.

2.4.1 | Comparison of results calculated from digitally extracted and provided time series data

The effect measures of interest in this study were selected on the basis of their common use in practice^{16,18,29}; namely, the immediate level change, β_2 , and slope change, β_3 , along with their associated standard errors, confidence intervals and *p*-values. We calculated these

estimates for each of the provided and extracted datasets. Across the ITS studies, different outcomes were measured, which necessitated the need to standardise the estimates of level and slope change for comparison across the time series. We achieved this by dividing these estimates by the range of the outcome of the provided time series data (i.e., the maximum observed value of the outcome minus the minimum observed value). We chose this method of standardisation ahead of others (e.g., standardising by the root mean square estimated (RMSE) from OLS regression²⁶) to overcome complications where series with very small RMSE estimates yield exaggerated interruption effect estimates.

2.4.2 | Estimates of immediate level and slope changes

We used Bland–Altman plots to assess pairwise agreement in the results (level change, slope change and their standard errors) calculated from the data extracted by each extractor and the provided data.³⁴ For each pairwise comparison (e.g., extractor 1’s data versus provided data) and each time series, the difference in the standardised effect estimates were plotted on the y-axis versus their average on the x-axis. For 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.³⁴ For each pairwise comparison, we calculated the mean difference in the standardised effect estimates and 95% limits of agreement (calculated as the mean of the differences ± 1.96 * standard deviation). The Bland–Altman plots were displayed in a matrix, depicting the agreement between each pairwise comparison. We used dot plots (by extractor) to display the distribution of differences in the standardised effects (immediate level change, slope change) calculated from extracted and provided data.

Summary statistics (e.g., mean differences, limits of agreement, medians) were tabulated.

2.4.3 | Confidence intervals for the interruption effect estimates

We compared the widths of the confidence intervals (for the interruption effect estimates) calculated using each of the four extracted datasets and the provided dataset. Specifically, for each pairwise comparison (e.g., extractor 1 vs. extractor 2) and each time series, we calculated the ratio of confidence interval widths, and scaled these so that the comparator (e.g., extractor 2) confidence interval spanned -0.5 to 0.5 (Appendix S2). For each pairwise comparison, a plot of the ratios of confidence interval widths (depicted by vertical lines) for all datasets was constructed. These plots were combined in a matrix of plots representing all pairwise comparisons.

2.4.4 | *p*-Values

We compared the *p*-values of the interruption effect estimates calculated using each of the four extracted time series and provided time series by categorising the *p*-values based on commonly used levels of statistical significance. We categorised *p*-values by dichotomising them at a 5% level of statistical significance (i.e., *p*-value <0.05 and ≥ 0.05) and also at a finer gradation (i.e., *p*-value ≤ 0.01 , $0.01 < p\text{-value} \leq 0.05$, $0.05 < p\text{-value} \leq 0.1$, *p*-value > 0.1). For each pairwise comparison between the extracted and provided time series, the percentage of time series for which there was agreement in the category of statistical significance was calculated.

3 | RESULTS

3.1 | Time series data acquisition

Our previous empirical study of 200 ITS studies included 230 ITS. Of these 230 ITS, data for 10 time series were available in the publications (e.g., as published Supporting Information) and data from a further 50 time series were obtained from contact with the authors.²⁶ These 60 time series, each of which was from a unique study, were considered potentially eligible for inclusion in the present study. Seventeen were excluded for the following reasons: an appropriate segmented linear regression model could not be used ($n = 4$); errors were identified in the time series ($n = 3$);

mismatch between the provided time series and the manuscript graph ($n = 3$); only summaries of the data were plotted in the manuscript graph ($n = 2$); data points were unable to be individually distinguished in the manuscript graph ($n = 5$). The remaining 43 time series form the cohort for the present study. The median series length was 40 time points (IQR 19–58, range 7–188).

3.2 | Quality of the included graphs

Fewer than two thirds (60%, 26/43) of the graphs had distinct individual points plotted (Table 2). Although most of the graphs had tick marks on the x-axis (88%, 38/43), the tick marks aligned with the data points in fewer than two thirds of these (61%, 23/38). In comparison, the y-axis always had tick marks with aligned labels. These findings reflected those of the original methods study which included 217 graphs¹⁵ (Table 2).

3.3 | Data extraction

Errors in the extracted time points (x-coordinates) were identified from the regression analyses (Table 3). Important errors (indicated by a standardised difference ≥ 0.5) in the extraction of time points varied across extractors from 9% (4/43) to 35% (15/43) (illustrative examples shown in Figure 2). A likely common cause of these errors was misalignment of the tick marks and data points in the original graph. Extraction of the outcome values (y-coordinates) was very accurate, as indicated by the summary statistics for the standardised intercepts and slopes being close to 0 and 1, respectively (Table 3). Most time series were extracted without any missing data points. However, when data points were missed, they frequently occurred in graphs which had no data points plotted (i.e., where only lines were plotted). The average time one extractor took to extract the data points was 3 min 41 s (median 3 m; IQR 2 m 20 s to 4 m 18 s; range 1 m 40 s to 16 m 20 s).

3.4 | Comparison of results calculated from digitally extracted and provided time series

Of the 43 time series, 33 were analysed with REML and 10 were analysed with OLS (due to REML failing to converge in at least one extracted or provided version of the

TABLE 3 Data extractor errors ($N = 43$).

Extractor ID	1	2	3	4
Regression of extracted time points versus provided time points				
Median (IQR) of standardised intercepts	0.023 (−0.006 to 0.480)	0.005 (−0.014 to 0.037)	0.003 (−0.009 to 0.012)	−0.000 (−0.036 to 0.037)
Mean (SD) of standardised intercepts	0.165 (0.308)	0.082 (0.269)	−0.011 (0.290)	−0.387 (1.898)
Median (IQR) of standardised slopes	1.000 (0.999 to 1.001)	1.000 (0.999 to 1.000)	1.000 (0.999 to 1.000)	1.000 (1.000 to 1.005)
Mean (SD) of standardised slopes	1.001 (0.015)	0.999 (0.012)	1.001 (0.009)	1.008 (0.040)
Important time point errors ¹	14/43	8/43	4/43	15/43
X-axis unaligned with data points where important time point errors occurred ²	11/14	4/8	0/4	9/15
Regression of extracted outcome values versus provided outcome values				
Median (IQR) of standardised intercepts	−0.000 (−0.001 to 0.003)	0.001 (−0.001 to 0.003)	0.000 (−0.002 to 0.004)	0.003 (−0.001 to 0.012)
Mean (SD) of standardised intercepts	0.009 (0.053)	0.009 (0.053)	0.001 (0.006)	0.012 (0.025)
Median (IQR) of standardised slopes	0.999 (0.997 to 1.002)	0.998 (0.994 to 1.001)	0.999 (0.996 to 1.002)	0.996 (0.985 to 1.000)
Mean (SD) of standardised slopes	0.980 (0.126)	0.979 (0.126)	0.998 (0.006)	0.976 (0.063)
Graphs from which at least one data point was missing	6/43	1/43	2/43	7/43
No data points plotted on	4/6	0/1	0/2	6/7

TABLE 3 (Continued)

Extractor ID	1	2	3	4
graph where at least one data point was missing ³				

Abbreviations: IQR, inter-quartile range given as the 25th and 75th centiles; SD, standard deviation.

¹A difference of greater than or equal to 0.5 between the extracted and provided data was deemed important.

²For example, if the data points were plotted between the tick marks.

³For example, if a line graph was used without any data points plotted.

TABLE 4 Level and slope change estimate differences between extracted and provided data.

Extractor ID	1	2	3	4
Level change difference between extracted and provided data				
Mean (LoA ¹)	-0.002 (-0.027 to 0.023)	-0.001 (-0.029 to 0.027)	-0.002 (-0.026 to 0.023)	0.002 (-0.037 to 0.040)
Median (IQR ²)	0.000 (-0.003 to 0.002)	-0.001 (-0.003 to 0.002)	0.000 (-0.002 to 0.002)	0.001 (-0.007 to 0.013)
Geometric mean ratio of standard errors for level change between extracted and provided data				
Mean (LoA)	1.001 (0.925 to 1.083)	1.002 (0.929 to 1.079)	1.000 (0.920 to 1.086)	1.015 (0.875 to 1.179)
Median (IQR)	1.000 (0.992 to 1.009)	0.999 (0.995 to 1.005)	1.000 (0.992 to 1.005)	1.001 (0.988 to 1.032)
Slope change difference between extracted and provided data				
Mean (LoA)	0.000 (-0.005 to 0.006)	0.000 (-0.007 to 0.006)	0.000 (-0.005 to 0.006)	0.000 (-0.006 to 0.007)
Median (IQR)	0.000 (0.000 to 0.000)	0.000 (0.000 to 0.000)	0.000 (0.000 to 0.000)	0.000 (-0.001 to 0.001)
Geometric mean ratio of standard errors for slope change between extracted and provided data				
Mean (LoA)	1.003 (0.913 to 1.103)	1.004 (0.917 to 1.099)	1.000 (0.908 to 1.103)	1.022 (0.866 to 1.207)
Median (IQR)	1.000 (0.991 to 1.009)	1.000 (0.994 to 1.007)	1.001 (0.992 to 1.005)	1.001 (0.990 to 1.044)

¹LoA: Limits of agreement calculated as the average ± 1.96 * standard deviation of the differences.

²IQR: Inter-quartile range given as the 25th and 75th centiles.

time series analysis). The average differences in immediate level change calculated from the extracted and provided time series were not importantly different (Table 4, Figures 3 and 4). The largest limits of agreement across the extractors (extractor 4) were ± 0.04 (on a scale ranging from 0 to 1). The limits of agreement were generally driven by a few large differences, but the interquartile ranges indicated that for the central 50% of the time series, the differences were negligible. Similarly, the average

difference in the estimated standard errors of the level change was not importantly different. The largest limits of agreement across the extractors (extractor 4) showed that the estimated standard errors ranged from 13% smaller to 18% larger.

The average differences in slope change calculated from the extracted and provided time series were not importantly different (Table 4, Figures 5 and 6). The limits of agreement ranged from ± 0.007 for all extractors. The

interquartile ranges indicate that for the central 50% of the time series, the differences were negligible. The standard error limits of agreement were again larger for extractor 4 (ranging from 13% smaller to 21% larger) with the other three extractors ranging from 9% smaller to 10% larger.

3.5 | Confidence intervals

Pairwise comparisons of immediate level and slope change estimates calculated from extracted and provided time series yielded very similar confidence

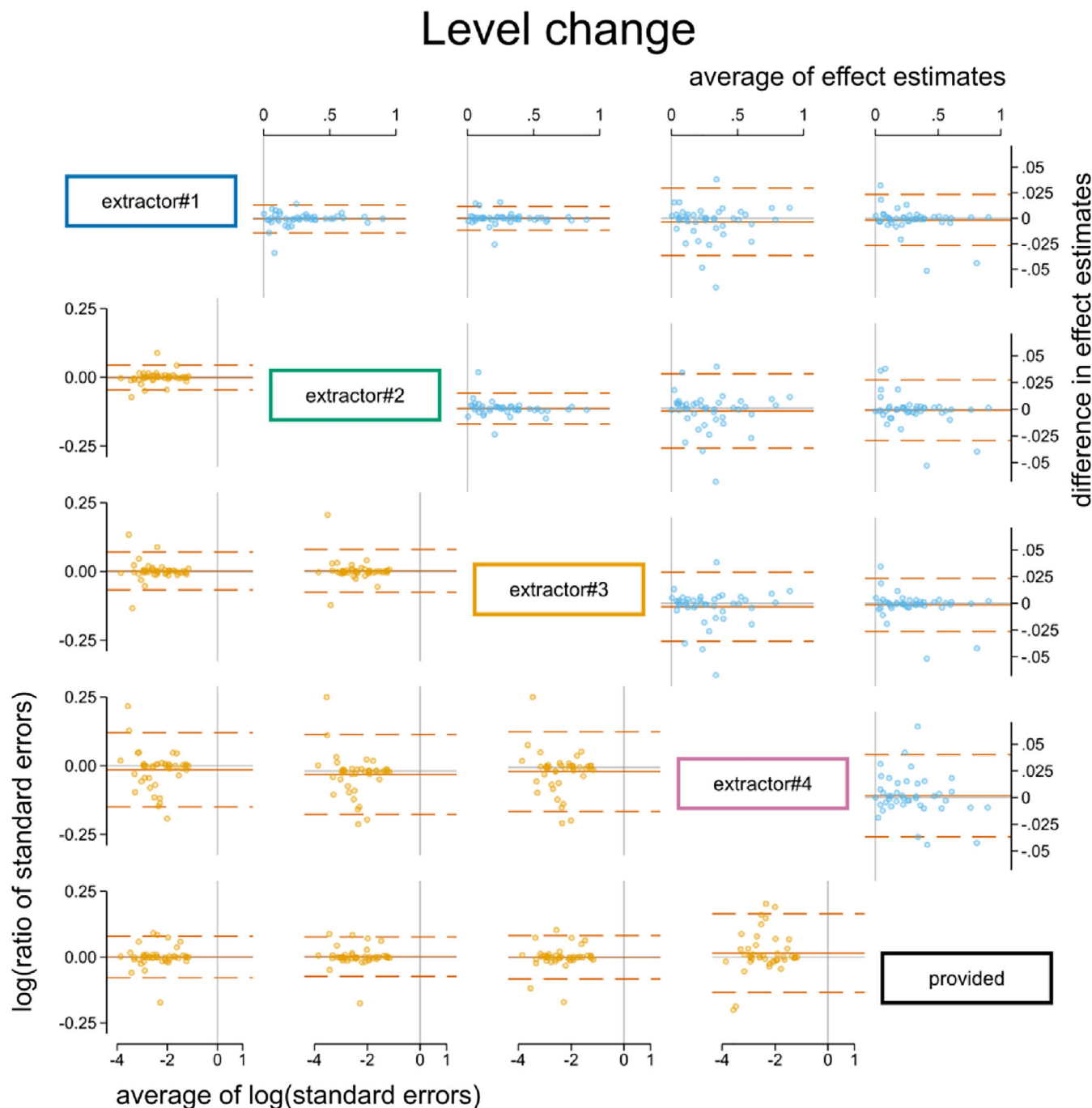


FIGURE 3 Bland Altman plot of standardised level change. Plots in the top triangle (blue points) show the difference in point estimates (row data source—column data source) on the vertical axis and average of the parameter estimates on the horizontal axis. Plots in the bottom triangle (orange points) show differences in standard errors on the vertical axis ($=\log(\text{ratio of standard errors})$) (column data source—row data source) 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 $\pm 1.96 \times$ standard deviation of the differences). Grey lines indicate zero. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

interval widths (Figure 7). In none of the 43 times, series did an estimated immediate level or slope change calculated from extracted time series fall outside of the confidence interval for the effect calculated from the provided time series.

3.6 | *p*-Values

The agreement in the statistical significance (dichotomised at the 5% significance level) for estimates of immediate level change and slope change calculated from extracted and provided time series were near identical (Appendix S2). The only exceptions to this arose for one extractor (extractor 4), in which there was discordance in one time series for the immediate level change (1/43, 2%) and two time series for the slope change (2/43, 5%). Examining agreement using the finer gradation of statistical significance categories showed that discordance between time series was rare, but when it arose, it generally occurred in the adjacent category (e.g., results from one extracted time series with a p -value ≤ 0.01 and result from the provided time series with a $0.01 < p$ -value ≤ 0.05).

4 | DISCUSSION

Four authors digitally extracted data from 43 ITS using the tool WebPlotDigitizer²⁷ and we compared the accuracy of the extracted x-axis and y-axis coordinates to the time series used to create the original graphs. We analysed the extracted and provided time series using segmented linear regression models and compared estimates of immediate level change, slope change, their associated standard errors, confidence intervals and p -values between the extracted and provided time series. We found that although there were some errors in the data extraction, primarily in the time points (x-coordinates), this did not translate into important differences in analysis results (across all metrics) between the digitally extracted and the provided time series.

Data extraction accuracy was generally poorer for the x-axis than the y-axis. This may have been because for the x-axis there was more often misalignment between the tick marks and data points as compared with the y-axis. Two of the x-axis errors occurred because the graph in the original manuscript did not include all of the time points on the axis so the data scaling made by WebPlotDigitizer was incorrect (which works by

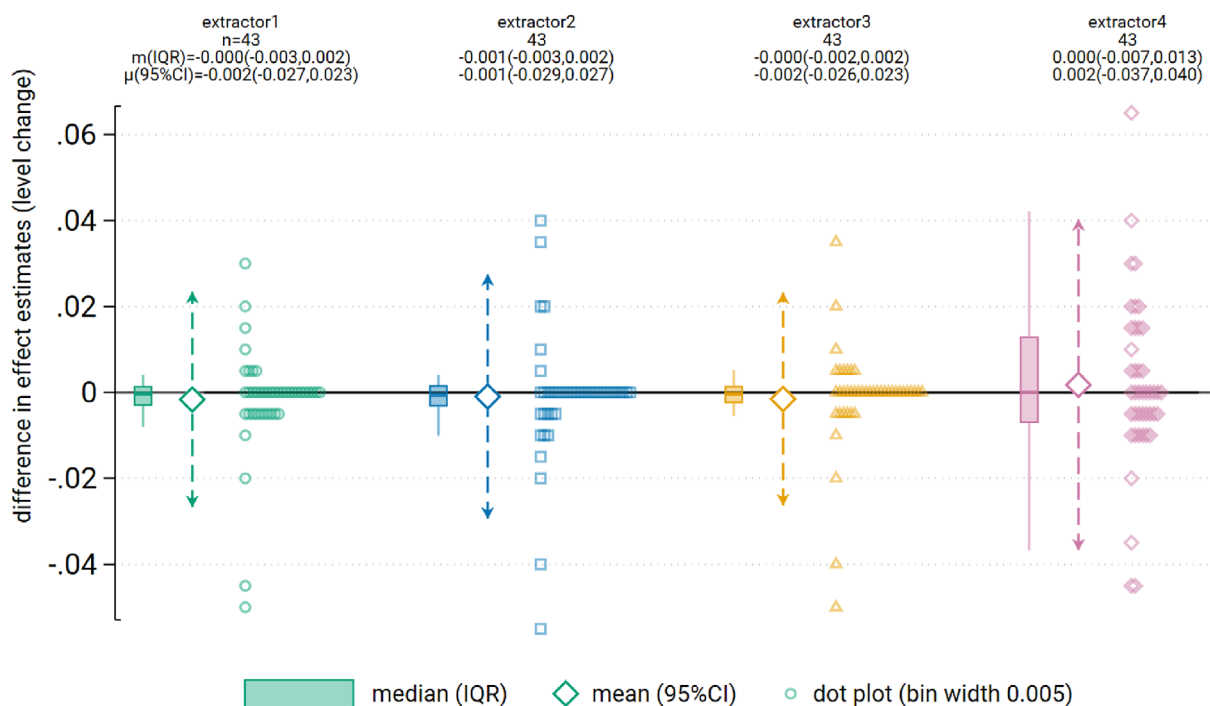


FIGURE 4 Dot plot showing difference in level change point estimates between extractor and provided data. Dot plot data has been aggregated to the nearest 0.005. Box plots show the median (m) (solid horizontal line), interquartile range (box) and lower and upper adjacent values (vertical lines). Large diamonds show the mean (μ) with 95% limits of agreement (dashed arrows). [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

Slope change

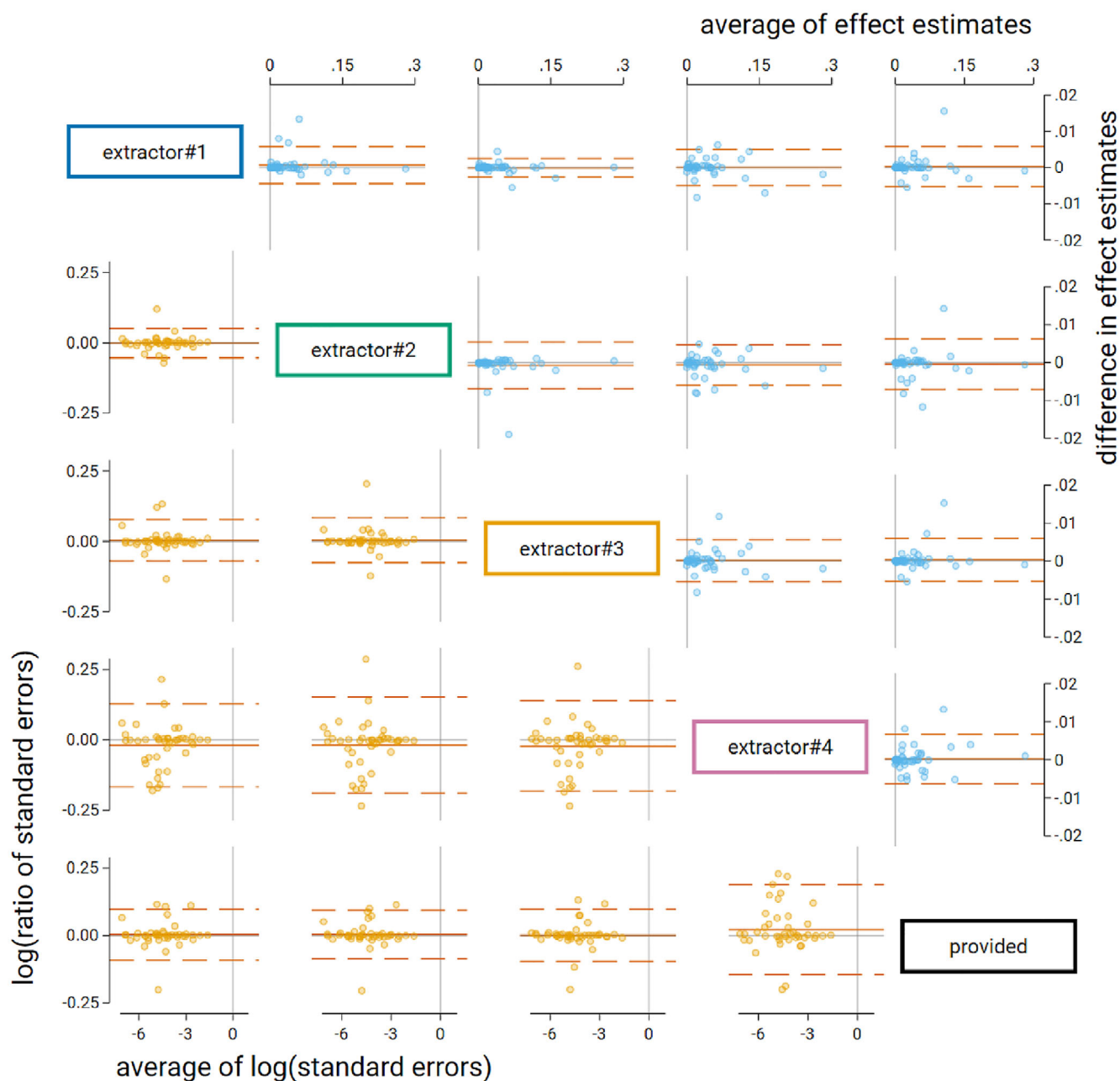


FIGURE 5 Bland Altman plot of standardised slope change. Plots in the top triangle (blue points) show the difference in point estimates (row data source—column data source) on the vertical axis and average of the parameter estimates on the horizontal axis. Plots in the bottom triangle (orange points) show differences in standard errors on the vertical axis ($-\log(\text{ratio of standard errors})$) (column data source—row data source) 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 $\pm 1.96 \times$ standard deviation of the differences). Grey lines indicate zero. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/jsm.1464)]

calculating the screen distance between the two defined end points; in one case there was no January, in another there was no zero; Appendix S2). One y-axis error occurred due to a point in the provided data not being plotted on the graph (the value was higher than the

plotted y-axis scale range). Two extractors missed data points in several graphs, with the majority of these occurring when line graphs were plotted without data points. Extractor 4 assigned two observations to the same month in several different time series, which arose due to

FIGURE 6 Dot plot showing difference in slope change point estimates between extractor and provided data. Dot plot data has been aggregated to the nearest 0.005. Box plots show the median (m) (solid horizontal line), interquartile range (box) and lower and upper adjacent values (vertical lines). Large diamonds show the mean (μ) with 95% limits of agreement (dashed arrows). [Colour figure can be viewed at wileyonlinelibrary.com]

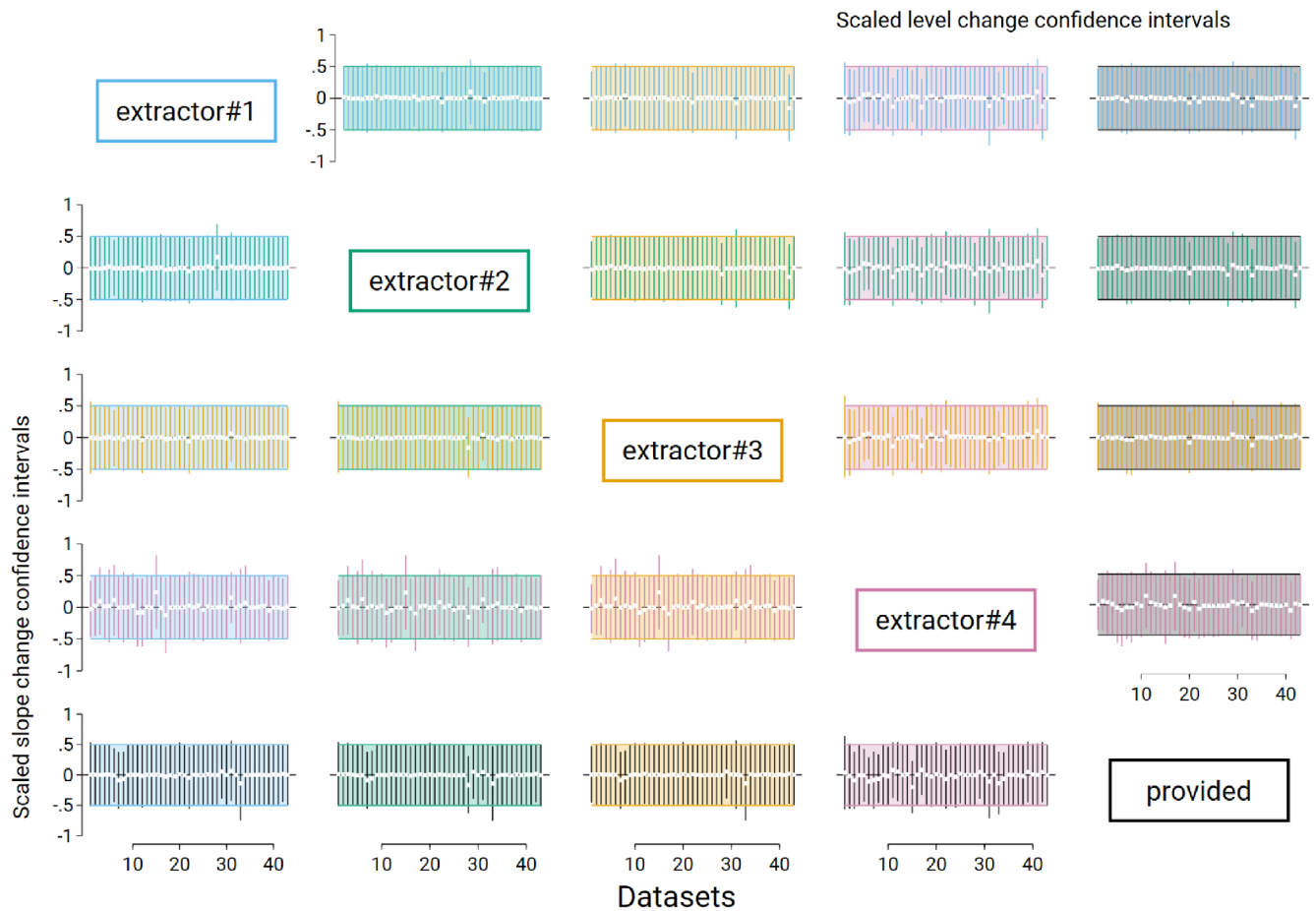
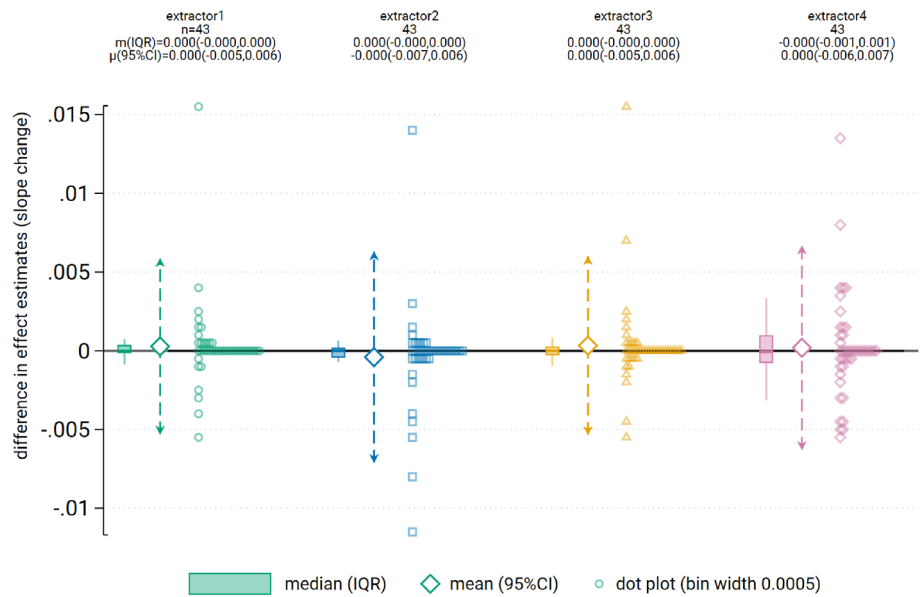


FIGURE 7 Pairwise confidence interval comparisons for immediate level change (top right triangle) and slope change (bottom left triangle). Each plot displays the 43 confidence intervals (CIs; depicted as vertical lines), with each scaled so that the confidence interval from the reference data source spans -0.5 to 0.5 (shaded area). The reference method is the column data source [e.g., the plot in the second row, fifth column shows extractor 2 level change CIs (green) compared to provided (black)]. Vertical lines falling entirely within the shaded area have smaller confidence intervals than the comparison, while lines extending beyond the shaded area have larger confidence intervals than the comparison. White dots indicate the point estimates. [Colour figure can be viewed at wileyonlinelibrary.com]

rounding time points to the nearest month. In these cases, the second data point was dropped from the analysis.

Level and slope change parameter estimates calculated from the time series obtained by the extractors and the provided data were very similar. The confidence intervals were almost identical for most time series. Furthermore, there was only one extracted time series for which the statistical significance (categorised at the 5% level) of the level change estimate was discordant to that calculated from the provided data, and two such instances for the slope change estimate. Many of the data extraction errors in the time points (x-coordinates) did not impact the interruption effect estimates (and their associated statistics) because the *actual* time points are not used in the model. Instead, time is modelled as a consecutive integer, where each value represents the number of time points from the start of the series to a particular time point. The largest discrepancies between extracted data and provided data occurred when the manuscript graphs contained errors (Appendix S2), the individual data points were not plotted (e.g., Figure 1b) or there was misalignment between x-axis tick and the data points.

Many studies examining the accuracy of digital data extraction have focused on plots with clearly defined data points (e.g., scatter plots), and found high levels of agreement between extracted and provided data. deOliveira found all intraclass correlation coefficients between extractors and original data >0.985 for three scatterplots with two data extractors.²³ Burda et al found the intraclass correlation coefficient between extractors >0.95 and percentage differences between the extractors and the original data ranged from 0.3% to 8.92% for two short scatterplots (fewer than 10 data points) and 15 data extractors.²⁴ van der Mierden et al found concordance correlation coefficients between original and extracted data were >0.99 for outcome data and >0.92 for standard error of the mean data for 26 bar charts and two scatter plots (36 data points all together) with six extractors.²¹

Other studies in the area of survival analysis have investigated whether digitally extracting data from Kaplan–Meier curves, and feeding this into an algorithm, leads to accurate recreation of the individual participant time-to-event data.^{35,36} These studies found that analysis of the recreated data yielded a high degree of accuracy for most statistics (e.g., survival probabilities, median survival times) compared with the original data.

4.1 | Implications for practice

For researchers of primary ITS studies, following recommendations for creating ITS graphs suggested by Turner

et al¹⁵ is encouraged. These recommendations were formed to achieve two goals. First, to provide an accurate visual display of the ITS data, and second, to display essential details for accurate data extraction. In addition, we encourage researchers to share their time series data (e.g., using Supporting Information).³⁷ Improved time series graphs and data sharing will facilitate inclusion of ITS studies in systematic reviews and meta-analyses.

For systematic review authors, we encourage the use of digital data extraction of time series data. Our findings demonstrate that accurate estimates of interruption effect estimates and their standard errors can be obtained from extracted data, and that this can be achieved with minimal time investment. The inclusion of these studies, even at the risk of slight inaccuracy, is expected to outweigh the loss of information from non-inclusion. Furthermore, the chosen statistical estimation method may in fact have more influence on the results than any errors in the data extraction. In most circumstances, extraction of data by one extractor will be sufficient. For graphs where data points are not plotted, extraction by multiple extractors may be beneficial.

One limitation of using digitally extracted time series data is that it will not be possible to adjust for any time-varying confounders beyond seasonality or those captured by the modelled linear trend.³⁸ For short to medium series, large changes in population characteristics (e.g., age or ethnicity distributions), that might potentially confound effect estimates, are unlikely. However, review authors need to be alert to potential time-varying confounding variables, and factor this into their risk of bias assessments.³⁹

4.2 | Strengths and limitations

A strength of this study was the representative sample of ITS graphs. The cohort of ITS included in the present paper was a subset of a randomly selected sample of 200 ITS; both samples included time series of similar lengths (median 40 in the present study compared to 48) and had graphs with very similar characteristics. Our cohort of ITS included a range of graph types that had different characteristics (e.g., with and without data points plotted), thus providing evidence of the accuracy of digital data extraction on the range of graphs that are likely to be encountered in practice. We chose a data extraction tool (WebPlot-Digitizer²⁷) that has been shown to be accurate in other contexts.^{22,24} Finally, we not only examined the accuracy of the extracted data, but also went a step further to examine whether errors in the data extraction translated to important differences in the interruption effect estimates (and their associated statistics).

One limitation of our study is that the chosen segmented linear regression model and statistical estimation

method may not have resulted in the best fitting model for the datasets. However, the purpose of our re-analysis was to compare interruption effect estimates calculated across extracted and provided datasets, and not to focus on the results of the analyses themselves.

A further limitation of our study is that we only examined one model structure (i.e., one that included both a level and slope change). While the model structure we chose is commonly used in practice,¹⁸ it is possible that our results do not generalise to more complex model structures (e.g., those which include splines or other non-linear functions).

5 | CONCLUSION

Publications of ITS studies rarely provide time series data, but often include a time series graph, thus providing the opportunity for digital data extraction, re-analysis and inclusion of the study in meta-analyses. In a cohort of 43 ITS studies, with four data extractors extracting data from each, we found that although there were some errors in extraction of time points, this did not translate into important differences in interruption effects (and associated statistics) estimated from segmented linear regression models. We therefore encourage systematic review authors to digitally extract time series data from ITS graphs to minimise the unnecessary loss of data in meta-analyses.

AUTHOR CONTRIBUTIONS

Simon Lee Turner: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; software; validation; visualization; writing – original draft; writing – review and editing. **Elizabeth Korevaar:** Investigation; writing – review and editing. **Miranda Cumpston:** Investigation; writing – review and editing. **Raju Kanukula:** Investigation; writing – review and editing. **Andrew Forbes:** Supervision; writing – review and editing. **Joanne E McKenzie:** Funding acquisition; investigation; methodology; resources; supervision; writing – original draft; writing – review and editing.

ACKNOWLEDGMENTS

The authors wish to thank all of the authors who generously contributed datasets for this study (Appendix S3). Open access publishing facilitated by Monash University, as part of the Wiley - Monash University agreement via the Council of Australian University Librarians.

FUNDING INFORMATION

Joanne E. McKenzie is supported by an NHMRC Investigator Grant (GNT2009612), Simon Lee Turner is funded by the Research Support Package of this grant. Miranda S. Cumpston and Elizabeth Korevaar are

supported by the Australian Government Research Training Program. The funders had no role in study design, decision to publish or preparation of the manuscript.


CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data and the Stata 17 code used to analyse the data and produce the tables and figures in this manuscript are available from the Figshare repository <https://figshare.com/s/10633a410a14fabf73d5>.

ORCID

Simon Lee Turner  <https://orcid.org/0000-0001-9163-4524>

Elizabeth Korevaar  <https://orcid.org/0000-0001-5808-7813>

Miranda S. Cumpston  <https://orcid.org/0000-0001-6564-8615>

Raju Kanukula  <https://orcid.org/0000-0003-0793-786X>

Andrew B. Forbes  <https://orcid.org/0000-0003-4269-914X>

Joanne E. McKenzie  <https://orcid.org/0000-0003-3534-1641>

REFERENCES

- 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*. 2015;350:h2750.
- 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.
- Lopez Bernal J, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int J Epidemiol*. 2016;46:dyw098.
- Sanson-Fisher RW, Bonevski B, Green LW, D'este C. Limitations of the randomized controlled trial in evaluating population-based health interventions. *Am J Prev Med*. 2007;33(2):155-161.
- Biglan A, Ary D, Wagenaar A. The value of interrupted time-series experiments for community intervention research. *Prev Sci*. 2000;1(1):31-49.
- Craig P, Cooper C, Gunnell D, et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. *J Epidemiol Community Health*. 1979;66(12):1182-1186.
- Victoria CG, Habicht J-P, Bryce J. Evidence-based public health: moving beyond randomized trials. *Am J Public Health*. 2004;94(3):400-405.
- Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. *Acad Pediatr*. 2013;13(6):S38-S44.
- McKenzie JE, Brennan SE, Ryan RE, Thomson HJ, Johnston RV, Thomas J. *Cochrane Handbook for Systematic*

- Reviews of Interventions version 6.3. 2022 [cited May 2022]. In: Chapter 3: Defining the criteria for including studies and how they will be grouped for the synthesis [Internet].
10. Reeves BC, Deeks JJ, Higgins JPT, Shea B, Tugwell P, Wells GA. Cochrane Handbook for Systematic Reviews of Interventions version 6.3. 2022 [cited May 2022]. In: Chapter 24: Including non-randomized studies on intervention effects [Internet].
 11. Wagner AK, Soumerai 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.
 12. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ.* 2011;342(7804):964-967.
 13. McKenzie JE, Beller EM, Forbes AB. Introduction to systematic reviews and meta-analysis. *Respiology.* 2016;21(4):626-637.
 14. Korevaar E, Karahalios A, Turner SL, et al. Methodological systematic review recommends improvements to conduct and reporting when meta-analyzing interrupted time series studies. *J Clin Epidemiol.* 2022;145:55-69.
 15. Turner SL, Karahalios A, Forbes AB, et al. Creating effective interrupted time series graphs: review and recommendations. *Res Synth Methods.* 2021;12(1):106-117.
 16. 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 Glob Health.* 2020;5(10):e003567.
 17. 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. *J Multidiscip Healthc.* 2020;13:411-423.
 18. Turner SL, Karahalios A, Forbes AB, et al. Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: a review. *J Clin Epidemiol.* 2020;122:1-11.
 19. 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.
 20. 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 Serv Res.* 2014;14(1):504.
 21. Van der Mierden S, Spineli LM, Talbot SR, et al. Extracting data from graphs: a case-study on animal research with implications for meta-analyses. *Res Synth Methods.* 2021;12(6):701-710.
 22. Drevon D, Fursa SR, Malcolm AL. Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. *Behav Modif.* 2017;41(2):323-339.
 23. de Oliveira IR, Santos-Jesus R, Po ALW, Poolsup N. Extracting numerical data from published reports of pharmacokinetics investigations: method description and validation. *Fundam Clin Pharmacol.* 2003;17(4):471-472.
 24. Burda BU, O'Connor EA, Webber EM, Redmond N, Perdue LA. Estimating data from figures with a web-based program: considerations for a systematic review. *Res Synth Methods.* 2017;8(3):258-262.
 25. Li T, Higgins JPT, Deeks JJ. Cochrane Handbook for Systematic Reviews of Interventions version 6.3. 2022 [cited May 2022]. In: Chapter 5: Collecting data [Internet]. <https://training.cochrane.org/handbook/current/chapter-05>
 26. 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. *BMC Med Res Methodol.* 2021;21:134.
 27. Rohatgi A. *WebPlotDigitizer.* 2019. <https://automeris.io/WebPlotDigitizer/>.
 28. StataCorp. *Stata Statistical Software: Release 17.* StataCorp LLC; 2021 <https://www.stata.com/>
 29. 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.
 30. 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.
 31. Huitema BE. *Analysis of Covariance and Alternatives Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies.* 2nd ed. Wiley; 2011.
 32. Huitema BE, Mckean JW. Design specification issues in time-series intervention models. *Educ Psychol Meas.* 2000;60(1):38-58.
 33. Gebiski 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):2131-2141.
 34. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8(2):135-160.
 35. Rogula B, Lozano-Ortega G, Johnston KM. A method for reconstructing individual patient data from Kaplan-Meier survival curves that incorporate marked censoring times. *MDM Policy Pract.* 2022;7(1):23814683221077643.
 36. Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol.* 2012;12(1):9.
 37. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data.* 2016;3(1):160018.
 38. Lopez Bernal J, Soumerai S, Gasparrini A. A methodological framework for model selection in interrupted time series studies. *J Clin Epidemiol.* 2018;103:82-91.
 39. Sterne JAC, Hernán MA, McAleenan A, Reeves BC, Higgins JPT. *Assessing Risk of Bias in a Non-randomized Study.* John Wiley & Sons; 2019:621-641.

SUPPORTING INFORMATION

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

How to cite this article: Turner SL, Korevaar E, Cumpston MS, Kanukula R, Forbes AB, McKenzie JE. Effect estimates can be accurately calculated with data digitally extracted from interrupted time series graphs. *Res Syn Meth.* 2023; 14(4):622-638. doi:10.1002/jrsm.1646