International Journal of Epidemiology, 2018, 2082–2093
doi: 10.1093/ije/dyy135
Advance Access Publication Date: 5 July 2018
Education Corner

Education Corner

The use of controls in interrupted time series studies of public health interventions

James Lopez Bernal,[1]* Steven Cummins[1] and Antonio Gasparrini[1,2]

[1]Department of Public Health, Environments and Society, London School of Hygiene and Tropical Medicine, London, UK and [2]Centre for Statistical Methodology, London School of Hygiene & Tropical Medicine, London, UK

*Corresponding author. Department of Public Health, Environments and Society, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH, UK. E-mail: james.lopez-bernal@lshtm.ac.uk

Editorial decision 28 May 2018; Accepted 6 June 2018

Abstract

Interrupted time series analysis differs from most other intervention study designs in that it involves a before-after comparison within a single population, rather than a comparison with a control group. This has the advantage that selection bias and confounding due to between-group differences are limited. However, the basic interrupted time series design cannot exclude confounding due to co-interventions or other events occurring around the time of the intervention. One approach to minimizse potential confounding from such simultaneous events is to add a control series so that there is both a before-after comparison and an intervention-control group comparison. A range of different types of controls can be used with interrupted time series designs, each of which has associated strengths and limitations. Researchers undertaking controlled interrupted time series studies should carefully consider a priori what confounding events may exist and whether different controls can exclude these or if they could introduce new sources of bias to the study. A prudent approach to the design, analysis and interpretation of controlled interrupted time series studies is required to ensure that valid information on the effectiveness of health interventions can be ascertained.

Key words: Interrupted time series, quasi-experimental design, evaluation, controls, time series, natural experiments

Introduction

Evaluation of public health interventions normally relies on comparing the outcome of interest in a population exposed to an intervention with that in an external control group not subject to the same intervention.[1] Interrupted time series (ITS) is an increasingly popular design that adopts a different approach whereby comparisons are instead made across time within a single population.[2] This design is generally applied to natural experiments with an intervention introduced at a known point in time. By collecting data at regular intervals over time, a pre-post comparison can be made while accounting for underlying trends in the outcome.[2] Because the evaluation is based on observing a single population over time, the ITS design is free from problems due to between-group differences, such as selection bias or unmeasured confounders. Furthermore, by modelling the underlying trend, ITS also controls for within-group characteristics that tend to change only slowly over time, secular changes, random

Key Messages

- History bias due to other interventions or events occurring around the time of the intervention is the primary threat to the validity of interrupted time series studies.
- A wide range of different controls can be used in order to limit history bias and improve the validity of an ITS study.
- Controls should be selected by considering, a priori, the possible sources of history bias and examining for differential changes in covariates between the study series and the control series throughout the study period.
- Researchers should take care in interpreting the results of controlled interrupted time series studies, in particular when the results differ from those of simple (uncontrolled) analysis.

fluctuations from one time point to the next and regression to the mean.[3,4] Nevertheless, ITS studies cannot exclude time-varying confounders which do not form part of the underlying trend, for example other interventions or events occurring around the time of the intervention that may also affect the outcome.[5]

One approach that limits the threat of these other confounding events is to include a control series, that is a design known as a controlled (or comparative) interrupted time series (CITS) analysis. A lack of effect in a well-chosen control can provide stronger evidence to support a causal relationship between the intervention and outcome. Conversely, the presence of an effect in the control series indicates that the change may be attributable to different factors. Indeed, a number of recent within-study comparisons have provided empirical evidence of the validity of the CITS design, by demonstrating comparable results to randomized controlled trial (RCT) benchmarks.[6–9] Nevertheless, although the basic ITS design has been described in detail elsewhere and reference is made to the inclusion of a control as a method of improving the validity of the design,[2,10] there is little guidance available on the what a control series can and cannot solve and how to select an appropriate control in CITS studies. The purpose of this paper is to evaluate the use of controls in ITS studies and provide a framework for their selection, analytical approaches and the interpretation of results. We then provide an illustration of the application of this framework, using an example from a recent study where alternative types of controls can be selected and compared.

Evaluative study designs

In order to know whether an intervention has caused an effect, a comparison needs to be made between the observed change in the outcome and the counterfactual, that is, what would have happened if the intervention had not taken place. Of course, it is not possible to observe the intervention both being implemented and not being implemented in the same individuals in the same population at the same time, therefore the true counterfactual is never known. Evaluation design is therefore centred on creating the best approximation of the true counterfactual and then comparing what happened in the intervention group with the approximated counterfactual.[3] There are two main approaches to approximating the counterfactual: controlled designs and before-and-after designs.[3]

Controlled designs

Controlled designs normally compare the same outcome in the intervention group and in an external control. Randomized controlled trials and cross-sectional studies, as well as other designs less commonly used for intervention evaluations (such as cohort and case control studies), all make comparisons between a intervention group and a control. The advantage of this approach is that both intervention and control groups are compared at the same point in time, so other time-sensitive factors that would affect both populations (such as other interventions or events that might impact on the outcome of interest) can be excluded. Nevertheless, selection bias and differences between the intervention and control population may mean that observed effects could be due to other confounding factors (which may be unknown or difficult to measure) rather than the intervention.[1] Randomization addresses this limitation in experimental studies; however, this is often not desirable, feasible or practical in studies evaluating public health interventions.[1,11,12] Other approaches, such as adjusting for multiple variables in regression models or propensity score matching, can account for known characteristics that differ between the two groups, but cannot control for unmeasured confounders.[1,12,13]

Before-after designs

Before-after designs involve making a comparison between a period of time after the intervention has occurred and a

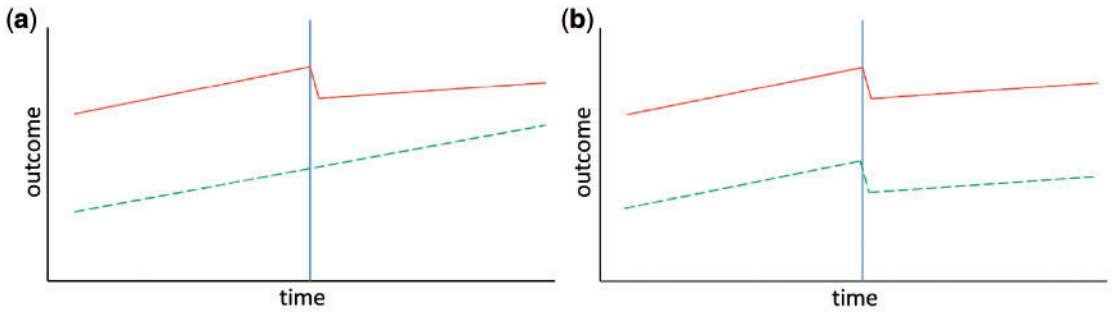

Figure 1. Controlled interrupted time series. Solid line = intervention series, dashed line = control series. (a) Here there is an effect in the intervention series (step and slope decrease) but no effect in the control series, which increases confidence that the effect is due to the intervention. (b) Here there is a step and slope decrease in both the intervention and control series, suggesting the change is due to some other event or co-intervention that affected both groups.

period of time before the intervention, within a single population. Here, the pre-intervention period effectively acts as the control. Simple pre-post designs make before-after comparisons by estimating the change from a single pre-intervention time point to a single post-intervention time point. However, these have poor internal validity as they cannot exclude underlying trends as a cause for any change. Conversely, interrupted time series use multiple pre-intervention and post-intervention observations, thereby allowing the underlying trend to be accounted for. These have the advantage that confounding is rarely a problem, as population characteristics tend to change only gradually over time.[3,14] Nevertheless, such before-after comparisons cannot exclude other events or co-interventions occurring around the same time as the intervention under investigation, as the cause of any detected change in the outcome. This phenomenon is known as history bias in Campell and Stanley's classical list of threats to internal validity.[5]

Controlled interrupted time series

A controlled (or comparative) interrupted time series (CITS) involves adding a control series, which was not exposed to the intervention, to the basic ITS design (Figure 1).[9] This results in the definition of a more complex counterfactual based on both a before-after comparison and an intervention-control comparison. The primary benefit of this approach is that it can help to control for history bias due to time-varying confounders, in particular co-interventions and other events concurrent with the intervention.[3] In a CITS, if an effect is detected in the intervention group but not in a well-chosen control (Figure 1a), this suggests that the effect is more likely to be due to the intervention; conversely if an effect is detected in both the intervention and the control series (Figure 1b), this suggests that it is due to some confounding event.

CITS is related to other study designs applied in evaluation analyses. For instance, the controlled before-after design (CBA) also involves a before-after and intervention-control comparison. Nevertheless, the CBA design involves a comparison of a single pre- and a single post-intervention, or a comparison of pre- and post-intervention means. Although both CITS and CBA designs involve a difference in difference calculation, CBA designs do not take into account baseline trends and therefore use the control group alone in order to approximate the counterfactual.[3,15]

An extension of the CITS design is the multiple baseline design. This is similar to a stepped wedge cluster randomized trial, but typically does not involve randomization. Here, following a baseline period, the intervention is first introduced in one group while one or more other groups act as a control.[16,17] The intervention is subsequently introduced in other groups at different times, with a different subset acting either as intervention or control groups at each time. In this design, the observation of an effect of similar strength and magnitude, following the intervention in multiple different groups at multiple sequential time points, can provide strong evidence that the observed effect is due to the intervention rather than other potential confounding events.[16,17]

Selecting a control

With studies that rely on the control as the sole means of approximating the counterfactual (including RCTs, cross-sectional studies and CBA studies) the central prerequisite when selecting a control is that it is as similar as possible to the intervention group. The ideal control is the same in terms of all variables other than exposure to the intervention.[1,3] RCTs accomplish this through randomization. Where randomisation is not possible, a range of methods have been developed to achieve covariate balance in cross-sectional and CBA designs including multivariable

regression, propensity score matching and synthetic controls.[18–20] Nevertheless, none of these methods can account for systematic differences in unknown variables.[18,21]

As described above, ITS studies use the pre-intervention trend to predict the counterfactual. The purpose of the control in this case is to exclude time-varying confounders, in particular co-interventions or other events occurring around the time of the intervention, as these are generally unpredictable based on modelling pre-intervention trends.[2,3] It follows that the key attribute of a control series for a CITS study should be its ability to control for known co-interventions or external events that may affect the outcome. Therefore, the control series should be exposed to any such co-interventions or events that might also affect the intervention series; however, it should not be exposed to other interventions or events that could impact on the control series alone (and not the intervention series). The latter could result in artefactual effects being detected in the CITS, which are in fact due to independent changes in the control series. Several different types of control series have been used for CITS analyses. We have broadly classified some of the most commonly used controls as follows: location-based control groups, characteristic-based control groups, behaviour-based control groups, historical cohort controls, control outcomes and control time periods. Table 1 describes these six types of controls, each of which may plausibly control for different sources of confounding events.

Researchers should also consider whether the intervention under study could have an indirect effect on the control series. For example, there may be a contamination effect in location-based or characteristic-based control groups, or a substitution effect with control outcomes.[22,23] A contamination effect occurs when the effects of the intervention spread beyond the target population, for example with behaviour change interventions, whereby members of the control population learn about the new behaviour and adopt it themselves.[22] An example of a substitution effect would be an evaluation of the effect of an intervention aimed at reducing the prescription of a certain drug. In this scenario, prescriptions of a similar drug not targeted by the intervention may be considered as a control intervention, but doctors may substitute the targeted drug with the similar drug, so that it is indirectly affected by the intervention.[23] Control series that could be indirectly affected by the intervention should be excluded.

Finally, although covariate balance between the intervention and control series in ITS is not required to predict the counterfactual, and is therefore not the fundamental prerequisite that it is in other controlled designs, it remains important for two reasons: first, certain subgroups may be more susceptible to either an intervention or a confounding event than others. If such a subgroup is more concentrated in the intervention group than the control, one would expect a greater effect in the intervention group simply due to the population distribution. Second, if certain characteristics are associated with the outcome and these characteristics change differentially over time in the intervention and control groups, the trend in the outcome may change in one group but not the other simply due to differential changes in the populations under investigation. For example, there is evidence that rates of cycle head injuries are lower in females than in males.[24] In the cycle helmet legislation study by Dennis *et al.*,[25] described in Table 1, if the intervention population had a higher proportion of females at baseline than the control population this would not necessarily be a problem. Nevertheless, if the proportion of females increased more rapidly in the intervention group than in the control population following the intervention, this would be a source of confounding, as there may be a decrease in head injuries in the intervention group simply due to the population change, rather than any effect of the intervention. Matching techniques, including propensity score matching, can be used to ensure balance of known covariates at baseline, which can help to limit the effects of differential susceptibility to the intervention by population subgroup.[19,26] Furthermore, synthetic control approaches can be applied to ITS studies where multiple potential controls exist. This approach re-weights a range of different control groups, so that the weighted average of their baseline characteristics is as similar as possible to that of the study group (maximizing covariate balance).[27] Linden[27] 2018 demonstrates an example of the use of synthetic controls in interrupted time series, which produces strong covariate balance and no significant difference from the intervention group in terms of pre-intervention level and trend in the outcome. Nevertheless, whether matching or using synthetic controls, it is still important to check for covariate balance between the control and intervention group throughout the study period. If there are changes over time, variables associated with the outcome can be included in the interrupted time series regression model to adjust for confounding. However, none of these methods can control for unknown confounding, and this should be recognized as a limitation of CITS studies in common with other non-randomized controlled designs.

Analysis and interpretation of CITS studies

There are a range of analyses that can be employed when undertaking CITS studies. These can broadly be divided into two: separate analysis of the intervention series and the control series; or a single model incorporating both series. Separate analysis is the simpler approach and may be

Table 1. Types of controls

Type of control	Description	Examples	Strengths	Limitations
Location-based control	The control series is selected from another location similar to the study location but which did not receive the intervention. The type of location depends on the scale of the intervention; for large-scale interventions, this may be a different geographical area (such as a country, district or city), whereas for smaller-scale interventions, this could be a different institution or a different ward within a hospital	Dennis *et al.*[25] evaluated the impact of the introduction of helmet legislation in a number of Canadian provinces on cycling-related head injuries, by comparing outcomes in Canadian provinces that did not implement helmet legislation Lopez Bernal *et al.*[28] compared the change in hospital activity in England, following major health reforms, with those in Scotland where the reforms did not apply	Helps to control for confounding events that would affect both locations	Cannot exclude events that are unique to the intervention location. For example, in the study of helmet legislation, reductions in head injuries could be due to a protective effect of helmets (presumably the desired effect) or due to a reduction in the number of cyclists if the need to wear a helmet acts as a deterrent (which may not be a desired effect). Comparing with provinces that did not implement the legislation would not help to distinguish these[25]
Characteristic-based control	Interventions are sometimes targeted according to certain characteristics, for example only males or only females, a certain age group, a specific ethnic minority group or patients with a certain diagnosis. Controls may be chosen from those groups that were not targeted	Feigl *et al.*[29] investigated the impact of a national ban on smoking in high schools, and selected a control based on age by comparing trends in smoking prevalence among those aged 12–18 years compared with those aged 19–24 years Kontopantelis *et al.*[30] examined the impact of a national primary care financial incentive scheme on trends in consultation rates among patients with severe mental illness, compared with matched patient controls with no severe mental illness	In cross-sectional or similar designs, this type of control is not ideal as the characteristic that differentiates the two groups is a known confounder that cannot be controlled for. Nevertheless in ITS studies, where the pre-intervention trend is the primary control, characteristic control groups can help to exclude concurrent events with the intervention that both groups would have been exposed to	Interventions may have been targeted at the intervention group because of a detected deviation in the trend, for example in the smoking ban study, high schools may have been targeted because of recent increases in smoking among adolescents. Therefore trends could differ substantially from the control group[29]
Behaviour based-control	Sometimes the intervention does not affect all of those within the population at whom it is targeted. This tends to occur when the intervention targets a behaviour that some individuals never performed (either prior to the intervention starting or since). Those individuals who never performed the behaviour can therefore be used as a control group	Ross-Degnan *et al.*[23] evaluated the impact of the national withdrawal of a non-steroidal anti-inflammatory drug (Zomepirac) on the prescription of other analgesics. They used physicians who never prescribed Zomepirac (and were thus unaffected by its withdrawal) as the control group	Controls can be very similar to the intervention group other than in the specific behaviour targeted by the intervention	It may be difficult to directly identify those who did not perform the behaviour. Therefore, a proxy may have to be used, such as, age, in the alcopops study. This proxy may, however, introduce selection bias. For example, selection based on age could bias the alcopops study because age could be independently associated with both the intervention (younger people may

(Continued)

Table 1. Continued

Type of control	Description	Examples	Strengths	Limitations
		Kiseley *et al.*[31] used a CITS to evaluate the impact of an increase in taxation of 'alcopops' on alcohol-related harm, by comparing the effect in young people aged 15–29 with the effect in those aged 30–49. Alcopops tend to be favoured by young people, so it was expected that older groups would be largely unaffected		be lower earners and thus more affected by a tax increase) and the outcome (if rates of alcohol-related harm vary with age)[31]
Historical cohort control	Historical cohorts are commonly used in the evaluation of education interventions but have also been used for healthcare evaluations.[8] This is possible where a cohort periodically progresses to another level (for example moving from one school year to the next) and is replaced by another cohort. The intervention cohort can then be compared with a previous or subsequent cohort[3]	Schneeweiss *et al.*[32] evaluated the impact of a restriction of state funding of nebulized respiratory medication. The intervention time series used monthly observations of nebulized drug expenditure, primary care visits and admissions to emergency department for a year (6 months prior to the policy and 6 months after the policy). Control series were taken from the same population 1 year and 2 years before	Historical cohorts help to rule out seasonal effects (such as stockpiling of drugs in the Scheeweiss *et al.*[32] study) and events that occur on an annual basis	They would not control for events that are unique to the year in which the intervention was implemented
Control outcome	Where no control group is possible, another option is to compare the effect on the primary outcome with that in a related 'control outcome' (or 'non-equivalent dependent variable') within the same group. Such an outcome should not be affected by the intervention, but would be affected by confounding events	Walter *et al.*[33] (2011) conducted a study on the impact of helmet legislation on head injuries in Australia (similar to that by Dennis *et al.*,[25] described above). Rather than other locations, they used limb injuries as a control outcome to exclude other effects on cycling Lopez Bernal *et al.*[34] (2013) used accidental deaths as a control outcome in their ITS study of the impact of the financial crisis on suicides in Spain, as both suicides and accidental deaths undergo similar judicial review and recording methods. This enabled them to control for other events that could have impacted on these processes.	Uses the same group as an intervention population, therefore it is not sensitive to many of the between-group differences that can affect other controls Can often be used to control for potential confounders that would only affect the intervention group. For example by using limb injuries as a control outcome, Walter *et al.* were able to control for any changes in the number of cyclists where comparing with different states could not[5,33]	Can only control for factors that would affect both the primary outcome and the control outcome

(Continued)

Table 1. Continued

Type of control	Description	Examples	Strengths	Limitations
Control time period	It may be possible to use the primary outcome as its own control for interventions that are only active at certain times (certain times of day or days of the week). In this case the outcome during times in which the intervention is inactive act as the control	Ross *et al.*[35] studied the impact of the 1967 British Road Safety Act, which increased the use of breathalysers to reduce drink-driving, on traffic casualties. They compared the effect on the weekend evenings, when pubs are busiest and accidents are more likely to be due to drink-driving, with that in commuting hours when pubs are closed and accidents are less likely to be due to drink-driving	Uses the same group as the intervention group; therefore it is not sensitive to many of the between-group differences that can affect other controls	Can only be used for short-term outcomes with rapid onset The outcome must be recorded at a sufficiently high time resolution to allow identification of when the intervention is active and inactive. For example, this may be to the nearest hour if the intervention is only active at night time[36]

suitable, particularly if there is no change in the control series. A single model can be developed by including indicator variables for the intervention or control series as interaction terms (Supplementary Appendix 1, available as Supplementary data at *IJE* online) or by generating a new series of the ratio or difference between the intervention and control series at each time point.[6,37] This approach provides a test of the differential effects of the intervention (level or slope change) across the groups. The benefit of this approach is that if there are trend changes in the control series which could be due to some confounding event, any additional effect of the intervention can still be calculated.

Even if a single model combining the intervention and control series is selected, we would recommend starting with a simple (uncontrolled) ITS of the intervention group. Both the uncontrolled ITS and the CITS should always be planned a priori and the results reported with equal prominence. If the result of the simple ITS mirrors that of the CITS, this provides a greater degree of confidence that any association between intervention and effect is likely to be causal. Results should be interpreted more cautiously if: either the simple ITS shows an effect but the CITS shows no effect (or a smaller effect); or if the CITS shows an effect but the simple ITS does not. If the simple ITS shows an effect but the CITS does not, then there may have been a change in both the intervention and the control series—this suggests possible history bias due to some simultaneous event or co-intervention.[3] If the CITS shows an effect but the simple ITS does not, the change may be due (at least in part) to a change in the control series, as a result of some other event that affected the control population but not the intervention group. This framework for analysing and interpreting CITS studies is summarized in Figure 2.

Analysis of CITS studies requires careful consideration of a number of statistical issues particular to time series data, including overdispersion, autocorrelation and seasonality. Furthermore, where multiple controls or intervention groups are used, clustering effects need to be taken into consideration. These analytical issues are beyond the scope of this paper but have been described in more detail elsewhere.[2,38–40]

It should be noted that the CITS model, outlined above and in Supplementary Appendix 1 (available as Supplementary data at *IJE* online), works best where the underlying trend is linear. Where non-linear trends exist, non-linear terms can be included within the time series model. Nevertheless, the more complex the trend, the more difficult it becomes to differentiate intervention effects from underlying fluctuations in the trend.[41] Where complex pre-intervention trends exist, it may be preferable to use a generalized difference in difference approach.

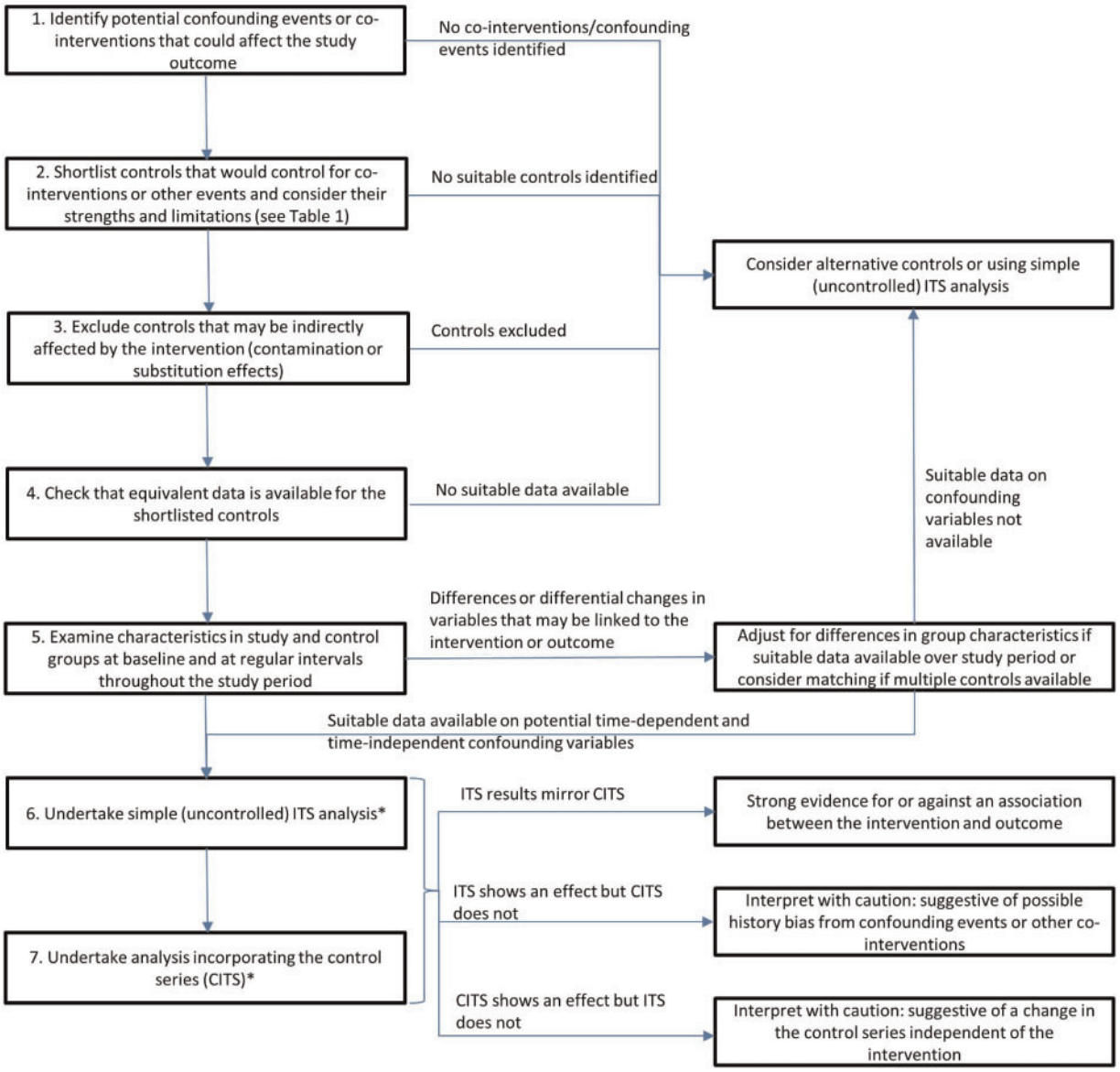

Figure 2. Suggested steps for undertaking a controlled interrupted time series study. *Both analyses should be undertaken and reported.

This has fewer restrictions on the shape of the time trend, but does assume that the treatment and control groups follow parallel trends. In either case, it is important that the assumption of linearity or parallel trends is checked.

Sensitivity analysis

Different ITS model assumptions can be checked using sensitivity analyses. Specific to CITS designs, different types of controls may control for different sources of bias or confounding events. Therefore where possible, researchers should undertake sensitivity analyses using different types of controls to control for those potential sources of bias that have been identified a priori. Similar to the primary model, sensitivity analyses should be clearly pre-specified to avoid the possibility of 'data dredging'.

Illustrative example

Steinbach *et al.*[36] recently used a CITS design to evaluate the impact of a range of changes to streetlights in various regions of the UK on road traffic crashes and crime at night.[36,42] The purpose of the intervention was to save energy and costs. The intervention consisted of reductions in the brightness of streetlights, replacement of bulbs with lower-energy consumption bulbs, reducing the hours during which streetlights were turned on at night (i.e. turning on later and turning off earlier) and reducing the ambient light threshold at which sensors would activate streetlights.

The authors hypothesized that although the intervention might save costs, reduced street lighting might unintentionally increase road traffic crashes and crime at night. To illustrate the design and interpretation of CITS studies, we used an extract of these data on minor roads in the Birmingham and Black Country region, to analyse the impact of the intervention (introduced from 2010) on the number of casualties from road traffic crashes. Outcome data were taken from the STATS19 Road Accident dataset; a STATS19 report form is completed by police officers for all accidents involving human injury or death. This includes information on the location, date and time of the accident and the severity of the injury. Note that, for simplicity of this illustration, we make the assumption that the intervention was introduced simultaneously in 2010 throughout the region and that it would have a step change effect. A number of different controls can be considered for the analysis, and we work through the process of selecting controls and analysing the CITS.

Data on road traffic crash casualties included variables on the region, the road type and the time of the road traffic crash. Therefore, three potential controls could be considered: (1) another region as a location-based control; (ii) comparison of casualties from road traffic crashes on minor roads with those on major roads as a characteristic based control; or (iii) comparison of road traffic crash casualties at night with road traffic crash casualties during the day when street lights are not in use, as a control time period.

Our first step in selecting a control is to identify potential confounding events or co-interventions that would affect the study outcome. In this study other changes to roads, such as changes to road layout or new road safety measures, were identified as a potential confounding event that could impact on road traffic crashes independently of the street lighting interventions. Another potential concern was instrumentation effects due to unidentified changes to data collection. Data quality reports suggest that 'local circumstances (for example organizational changes, reviews of coding practice and local initiatives) may affect the data and trends over time'. Considering each of the controls in turn: the location-based control would not be able to control for the identified confounding factors, as road changes may have differed from one region to the next and data collection was separate in each region. The characteristic controls (different road types) would control for changes to data collection processes within a region but would not be able to control for road changes, as these are likely to differ between minor and major roads. In this example, the control time period is the most appropriate, as this uses the same roads and same data source and should therefore adequately control for all known potential confounders. No other interventions or events that would only affect daytime road traffic crashes were identified, and it was considered unlikely that the intervention would have any indirect effect on this control. Daytime road traffic crashes were therefore selected as the control series.

The next step was to check characteristics of the control and intervention series at baseline and throughout the study period for covariate balance. We know that the data come from the same roads, and therefore this will not be different between night and day. However, no data on the characteristics of the population of night-time drivers compared with daytime drivers were available. One might assume that there are fewer elderly drivers with visual impairments at night, but this is unlikely to change differentially between the intervention and control group over the study period independently of the intervention.

Figure 3 shows the results of the analysis. First, an uncontrolled ITS analysis (Figure 3a) was undertaken. This shows a significant decrease in road traffic crash casualties following the intervention, contrary to the hypothesized increase. Nevertheless, when a CITS analysis using daytime road traffic crash casualties is run (Figure 3b), the decrease is also present in the control series and there is no evidence of any additional effect in the intervention series. This suggests that the effect is due to a change occurring at the same time as the intervention and biasing the previously estimated association.

To demonstrate the possible consequences of poor control selection, in Figure 3c a location-based control is used instead. We select the most closely matched region according to baseline characteristics (including number of roads in the region, population size, age distribution, sex distribution and level of unemployment). There is also no significant difference in baseline trends between the control and intervention group. In this case the results are very similar to the uncontrolled analysis, showing strong evidence of a decrease in road traffic crash casualties following the intervention. Nevertheless, this control group is clearly unable to account for changes to road layout or changes to data collection that are unique to the region, and could result in erroneous conclusions about the effect of the intervention. This highlights the potential pitfalls of selecting controls without first carefully considering potential confounding events or co-interventions specific to the study context, even when there is good covariate balance between the intervention and control group.

Where multiple possible confounding events exist, it may be best to use multiple different types of controls that can exclude different factors and can provide a more detailed picture of the intervention effect. For example, it is possible that the reduction in streetlighting could result in a substitution effect whereby people with poor vision are

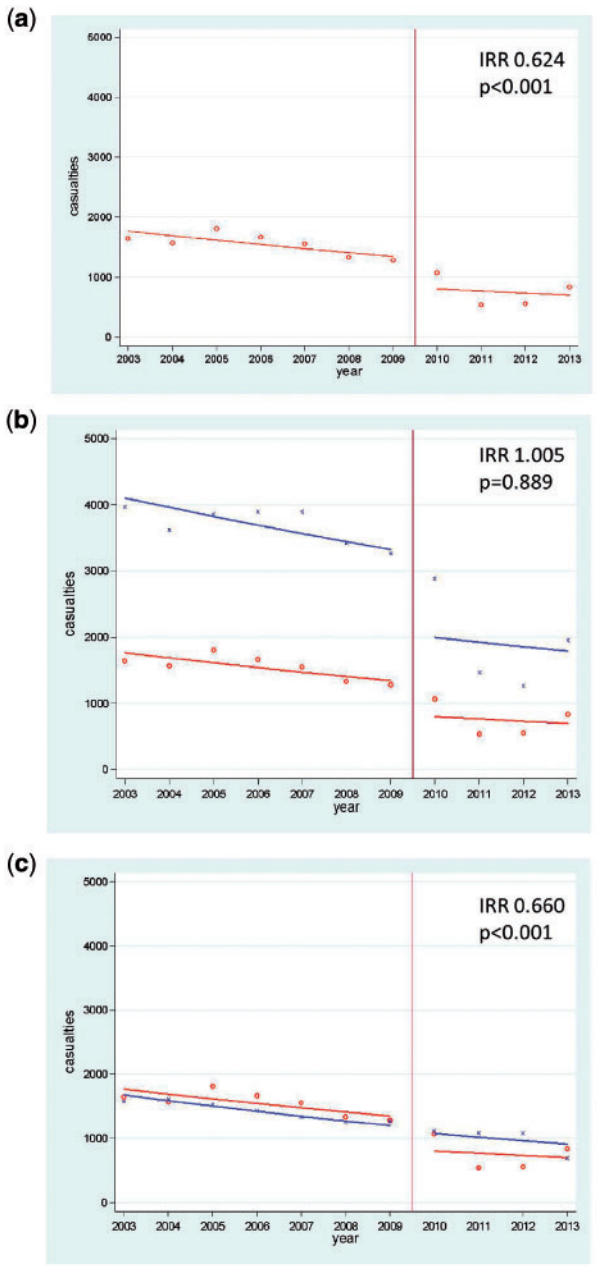

Figure 3. The effect of the Birmingham and Black Country street lighting intervention on road traffic crash casualties. Regression line with circles is the intervention series (night-time road traffic casualties on minor roads in Birmingham and the Black Country); regression line with crosses is the control series: (a) no control; (b) control time period: daytime road traffic crash casualties on minor roads in Hertfordshire; (c) location-based control: night-time road traffic crash casualties on major roads in West Yorkshire. The vertical red line is the intervention point. The incident rate ratio (IRR) is the step change in road traffic crash casualties following the intervention compared with before the intervention In (b) and (c), the IRR is the step change in the intervention series over and above any step change in the control series.

less inclined to drive at night following the intervention, due to poorer lighting, and do more of their driving during the day. This could therefore actually result in a reduction in night-time accidents. In order to examine this, one might consider comparing an analysis using the control time period and location-based control.

Conclusion

In this paper we have highlighted how ITS studies differ from other evaluation designs by making within-group rather than between-group comparisons. Although this has the advantage of limiting confounding by factors that change only slowly through time, history bias can still threaten the validity of ITS studies. A wide range of different controls can be used in order to limit history bias and improve the validity of an ITS study. Nevertheless, it is important to systematically consider a priori the degree of risk of history bias associated with any particular study, what control series are available and whether these will adequately control for history bias. Finally, researchers should take care in interpreting the results of CITS studies, in particular when the results of CITS analysis differ from those of simple (uncontrolled) ITS analysis. If the results of the CITS and the ITS analysis are aligned, CITS studies can provide strong evidence on the effectiveness of public health interventions, and when appropriate controls are selected the design ranks second only to randomized controlled designs in terms of capacity to control for bias.[14]

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

This work was supported by the UK Medical Research Council [MR/L011891/1].

Conflict of interest: None declared.

References

1. Bonell CP, Hargreaves J, Cousens S *et al*. Alternatives to randomisation in the evaluation of public health interventions: design challenges and solutions. *J Epidemiol Community Health* 2009;**65**:582–87.
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. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin, 2002.
4. 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**:613–23.

5. Campbell DT, Stanley JC. *Experimental and Quasi-experimental Designs for Research*. Boston, MA: Houghton Mifflin, 1963.
6. Fretheim A, Soumerai SB, Zhang F, Oxman AD, Ross-Degnan D. Interrupted time-series analysis yielded an effect estimate concordant with the cluster-randomized controlled trial result. *J Clin Epidemiol* 2013;**66**:883–87.
7. Fretheim A, Zhang F, Ross-Degnan D *et al*. A re-analysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation. *J Clin Epidemiol* 2015;**68:**324–33.
8. St Clair T, Cook TD, Hallberg K. Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *Am J Eval* 2014;**35:**311–27.
9. St Clair T, Hallberg K, Cook TD. The validity and precision of the comparative interrupted time-series design: three within-study comparisons. *J Educ Behav Stat* 2015;**41:**269–99.
10. 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**:299–309.
11. Petticrew M, Cummins S, Ferrell C *et al*. Natural experiments: an underused tool for public health? *Public Health* 2005;**119**: 751–57.
12. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;**312**:1215.
13. 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**: 155–61.
14. Soumerai SB, Starr D, Majumdar SR. How do you know which health care effectiveness research you can trust? A guide to study design for the perplexed. *Prev Chronic Dis* 2015;**12:** E101.
15. Fu AZ, Dow WH, Liu GG. Propensity score and difference-in-difference methods: a study of second-generation antidepressant use in patients with bipolar disorder. *Health Serv Outcomes Res Methodol* 2005;**8**:23–38.
16. Biglan A, Ary D, Wagenaar AC. The value of interrupted time-series experiments for community intervention research. *Prev Sci* 2000;**1**:31–49.
17. Hawkins NG, Sanson-Fisher RW, Shakeshaft A, D'Este C, Green LW. The multiple baseline design for evaluating population-based research. *Am J Prev Med* 2007;**33**:162–68.
18. Cousens S, Hargreaves J, Bonell C *et al*. Alternatives to randomisation in the evaluation of public-health interventions: Statistical analysis and causal inference. *J Epidemiol Community Health* 2011;**65**:576–81.
19. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011;**46**:399–424.
20. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *J Am Stat Assoc* 2010;**105**:493–505.
21. Winkelmayer WC, Kurth T. Propensity scores: help or hype? *Nephrol Dial Transplant* 2004;**19**:1671–73.
22. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? *BMJ* 2001;**322**:355–57.
23. Ross-Degnan D, Soumerai SB, Fortess EE, Gurwitz JH. Examining product risk in context: market withdrawal of zomepirac as a case study. *JAMA* 1993;**270**:1937–42.
24. Klauber MR, Barrett-Connor E, Marshall LF, Bowers SA. The epidemiology of head injury: a prospective study of an entire community - San Diego County, California, 1978. *Am J Epidemiol* 1981;**113**:500–09.
25. Dennis J, Ramsay T, Turgeon AF, Zarychanski R. Helmet legislation and admissions to hospital for cycling related head injuries in Canadian provinces and territories: interrupted time series analysis. *BMJ* 2013;**346**:f2674.
26. Linden A, Adams JL. Applying a propensity score-based weighting model to interrupted time series data: improving causal inference in programme evaluation. *J Eval Clin Pract* 2011;**17**: 1231–38.
27. Linden A. Combining synthetic controls and interrupted time series analysis to improve causal inference in program evaluation. *J Eval Clin Pract* 2018;**24:**447–53.
28. Lopez Bernal JA, Lu CY, Gasparrini A, Cummins S, Wharham JF, Soumerai SB. Association between the 2012 Health and Social Care Act and specialist visits and hospitalisations in England: a controlled interrupted time series analysis. *PLOS Med* 2017;**14**:e1002427.
29. Feigl AB, Salomon JA, Danaei G, Ding EL, Calvo E. Teenage smoking behaviour following a high-school smoking ban in Chile: interrupted time-series analysis. *Bull World Health Organ* 2015;**93**:468–75.
30. 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**:e008650.
31. Kisely SR, Pais J, White A *et al*. Effect of the increase in "alcopops" tax on alcohol-related harms in young people: a controlled interrupted time series. *Med J Aust* 2011;**195**:690–93.
32. Schneeweiss S, Maclure M, Carleton B, Glynn RJ, Avorn J. Clinical and economic consequences of a reimbursement restriction of nebulised respiratory therapy in adults: direct comparison of randomised and observational evaluations. *BMJ* 2004;**328**: 560.
33. Walter SR, Olivier J, Churches T, Grzebieta R. The impact of compulsory cycle helmet legislation on cyclist head injuries in New South Wales, Australia. *Accid Anal Prev* 2011;**43**: 2064–71.
34. Lopez Bernal J, Gasparrini A, Artundo C, McKee M. The effect of the late 2000s financial crisis on suicides in Spain: an interrupted time-series analysis. *Eur J Public Health* 2013;**23**:732–36.
35. Ross HL, Campbell DT, Glass GV. Determining the social effects of a legal reform The British "breathalyser" crackdown of 1967. *Am Behav Sci* 1970;**13**:493–509.
36. Steinbach R, Perkins C, Tompson L *et al*. The effect of reduced street lighting on road casualties and crime in England and Wales: controlled interrupted time series analysis. *J Epidemiol Community Health* 2015;**69:**1118–24.
37. Linden A. Conducting interrupted time series analysis for single and multiple group comparisons. *Stata J* 2015;**15**:480–500.
38. Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, Armstrong B. Time series regression studies in environmental epidemiology. *Int J Epidemiol* 2013;**42**:1187–95.

39. Xiao Y, Abrahamowicz M. Bootstrap-based methods for estimating standard errors in Cox's regression analyses of clustered event times. *Stat Med* 2010;**29**:915–23.
40. Desai M, Bryson SW, Robinson T. On the use of robust estimators for standard errors in the presence of clustering when clustering membership is misspecified. *Contemp Clin Trials* 2013;**34**: 248–56.
41. Lopez Bernal JA, Soumerai S, Gasparrini A. A methodological framework for model selection in interrupted time series studies. *J Clin Epidemiol* 2018, June 6. doi: 10.1016/j.jclinepi.2018.05.026.
42. Perkins C, Steinbach R, Tompson L *et al*. *What is the Effect of Reduced Street Lighting on Crime and Road Traffic Injuries at Night? A Mixed-methods Study*. Southampton (UK): NIHR Journals Library, 2015.