Use of interrupted time series methods in the evaluation of health system quality improvement interventions: a methodological systematic review

Celestin Hategeka, Hinda Ruton, Mohammad Karamouzian, Larry D Lynd, Michael R Law

ABSTRACT

Background When randomisation is not possible, interrupted time series (ITS) design has increasingly been advocated as a more robust design to evaluating health system quality improvement (QI) interventions given its ability to control for common biases in healthcare QI. However, there is a potential risk of producing misleading results when this rather robust design is not used appropriately. We performed a methodological systematic review of the literature to investigate the extent to which the use of ITS has followed best practice standards and recommendations in the evaluation of QI interventions.

Methods We searched multiple databases from inception to June 2018 to identify QI intervention studies that were evaluated using ITS. There was no restriction on date, language and participants. Data were synthesised narratively using appropriate descriptive statistics. The risk of bias for ITS studies was assessed using the Cochrane Effective Practice and Organisation of Care standard criteria. The systematic review protocol was registered in PROSPERO (registration number: CRD42018094427).

Results Of 4061 potential studies and 2028 unique records screened for inclusion, 120 eligible studies assessed eight QI strategies and were from 25 countries. Most studies were published since 2010 (86.7%), reported data using monthly interval (71.4%), used ITS without a control (81%) and modelled data using segmented regression (62.5%). Autocorrelation was considered in 55% of studies, seasonality in 20.8% and non-stationarity in 8.3%. Only 49.2% of studies specified the ITS impact model. The risk of bias was high or very high in 72.5% of included studies and did not change significantly over time.

Conclusions The use of ITS in the evaluation of health system QI interventions has increased considerably over the past decade. However, variations in methodological considerations and reporting of ITS in QI remain a concern, warranting a need to develop and reinforce formal reporting guidelines to improve its application in the evaluation of health system QI interventions.

INTRODUCTION

Improving health system quality is vital to ensure that utilisation of healthcare maximises population health outcomes, economic benefit and people’s confidence in health systems. A key strategy to improving the quality of care is the use of context-appropriate health system quality improvement (QI) interventions. QI interventions employ different approaches targeting patients or users (eg, education of patients and promotion of self-management), healthcare providers (eg, clinical education and reminders) and health systems (eg, organisational change and electronic patient registry) to bolster quality of care. While evaluating these interventions is key to understanding their impact, many are designed and
implemented without predetermined evaluation plans.\textsuperscript{6–9} In such cases, rigorous quasi-experimental designs can be employed.\textsuperscript{10–13}

Concerns persist, however, about the potential biases in some weaker observational study designs.\textsuperscript{10,11} As a result, interrupted time series (ITS) analysis has increasingly been advocated as one of the more robust observational designs as it can control for the secular trends present in many health system outcomes.\textsuperscript{7,14} Additionally, ITS can easily be interpreted by lay audiences, especially when graphical displays of the results are employed. Similarly, the Cochrane Effective Practice and Organisation of Care (EPOC) recommends ITS along with experiments and controlled before and after designs for inclusion in EPOC reviews.\textsuperscript{15} There remains a risk, however, of producing misleading or incorrect results when this rather robust design is not used appropriately.\textsuperscript{16–18}

While use of ITS in health research has increased substantially over the last few decades (online supplementary appendix figure 1), previous systematic reviews have shown heterogeneity in the reporting and methodological concerns in the application of ITS.\textsuperscript{17,19–21} However, none of these reviews have focused on use of ITS in the evaluation of health system QI interventions.\textsuperscript{17,19–20} Another review examined characteristics and reporting of ITS in healthcare more broadly.\textsuperscript{22} However, this review was restricted to just 1\textsuperscript{st} year (ie, 2015) and one database and excluded controlled ITS studies.\textsuperscript{21} Similarly, they did not assess the risk of bias of included studies. Therefore, the extent to which the use of ITS has followed best practice standards and recommendations in the evaluation of health system QI interventions remains unclear.

**Methodological considerations with interrupted time series**

The ITS design relies on data collected at multiple intervals over time (ie, time series data) before and after an intervention to establish a causal relationship between an intervention (eg, QI) and an outcome of interest (eg, health outcomes).\textsuperscript{22} The Cochrane EPOC recommends ITS studies with at least three data points before and after an intervention for inclusion in its reviews.\textsuperscript{15} ITS can be either retrospective or prospective, and can further be classified into two types based on whether one group was used (single ITS) or a control group was added (controlled ITS).\textsuperscript{10,22} Single ITS (SITS) compares longitudinal changes before and after the intervention in the exposed group, whereas controlled ITS (CITS) compares longitudinal changes before and after the intervention between an exposed group and a control group (figures 1 and 2).\textsuperscript{7,18,22–25}

Single ITS assumes that level and trend in a given outcome measure in the group exposed to the intervention would have remained the same absent the intervention (figure 1). In contrast, the counterfactual assumption for controlled ITS is that the level and trend in the group exposed to the intervention would have changed in the same way as was observed in the control group (figure 2).\textsuperscript{7,18,22–25} It is recommended to specify anticipated changes in the level or trend in outcome measures a priori, for example, based on expert knowledge.\textsuperscript{17,26,27} Graphical figures of results of ITS studies are vital as they allow readers to see pre-existing levels and trends, the implementation of the intervention and the impact of the intervention.\textsuperscript{18,20,28}

CITS is arguably a stronger quasi-experimental design as it has two controls (baseline trend in the intervention group and control group), allowing both within-group and between-group comparisons.\textsuperscript{10,11} This enhances its capacity to control for potential threats to internal validity such as history (co-intervention), maturation, instrumentation, regression to the mean and attrition.\textsuperscript{10,18,22,24,29} However, SITS may be more vulnerable to these threats to internal validity. For example, Baicker and Svoronos recently reanalysed data from the Oregon Health Insurance experiment using SITS and found results that were not consistent with the randomised trial findings.\textsuperscript{30} Therefore, it is generally recommended to add a control group to ITS wherever possible.\textsuperscript{10,22} Bernal and colleagues’ study reviewed various types of controls that can be added to strengthen ITS design, including location and characteristic-based controls and control outcome.\textsuperscript{24} When QI interventions are multisite and implemented at different times, this needs to be taken into account when assigning index time to a control group using techniques such as matching and random assignment of index time to controls.
Statistical analysis of interrupted time series

While there are many statistical approaches used for estimating effects of interventions in ITS studies, segmented regression analysis of ITS is the most commonly used, followed by autoregressive integrated moving average (ARIMA) models. Details on segmented regression analysis can be found in the seminal paper of Wagner and colleagues. While most ITS studies use segmented regression analysis on time series data aggregated within periods of time, analysis on individual-level data is feasible using mixed effects models. Saeed and colleagues discuss how to combine generalised mixed effects models with segmented regression to evaluate the effectiveness of healthcare interventions. Similarly, Ewusie suggests incorporating patient variability and sample size as weights in a weighted segmented regression model of ITS to account for potential heterogeneity introduced by population variability within and/or between sites. Power in ITS depends on several factors, including the number and distribution of data points before and after the intervention, the stability of data points, strength of effect, autocorrelation and seasonality. Missing observations are common in time series data of health service utilisation and quality of care and should be dealt with using appropriate techniques such as multiple imputation and maximum likelihood estimation.

As successive time series observations may be correlated, it is recommended to assess for autocorrelation using appropriate modelling techniques. There are several tests for assessing autocorrelation including Durbin-Watson tests and visual plots of the autocorrelation function and partial autocorrelation function. Moreover, seasonality is common in time series data and should be adjusted for as appropriate. There are several techniques for adjusting for seasonality in time series analyses, including deseasonalising time series data prior to fitting statistical models, adding dummy calendar variable (eg, calendar month), using Fourier terms (pairs of sine and cosine functions) and splines. When using monthly time series data, at least 12 data points before and after the intervention are recommended to be able to adjust for seasonality meaningfully. Additionally, non-stationarity—a time series exhibits an underlying trend that is not related to the intervention—should be assessed and, where significant, stationarised as appropriate. Non-stationarity can be assessed using the Augmented Dickey-Fuller test. When autocorrelation, seasonality and/or non-stationarity are present in time series data and are not adjusted for, this can result in biased results.

Despite the rich literature on how to approach ITS studies from a methodological and statistical standpoint, we do not have good information on the degree to which evaluations of healthcare QI interventions meet these criteria. Thus, we systematically reviewed the literature to investigate the extent to which these evaluations have followed best practice standards.

METHODS

This methodological systematic review was conducted according to a prespecified study protocol registered in PROSPERO (registration number: CRD42018094427) on 11 June 2018.

Search strategy

Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist (online supplemental appendix table 1), we searched for healthcare QI intervention studies that were evaluated using ITS and were published in peer-reviewed journals from inception to June 2018. We also searched the grey literature to capture any relevant studies. The search strategy is summarised in table 1 and full details are provided in online supplemental appendix table 2.

Inclusion and exclusion criteria

The population, interventions, comparisons, outcomes and study designs considered for review are listed in table 2. Guided by the Shojania et al’s taxonomy for QI...
Table 2  Population, interventions, comparisons, outcomes and study designs for study inclusion

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>All types of patients/consumers and healthcare professionals/providers are eligible. All levels of healthcare delivery are eligible.</td>
</tr>
</tbody>
</table>
| Intervention      | Interventions to improve quality of healthcare\(^4\)\(^5\):  
|                   | - Provider reminder systems;  
|                   | - Facilitated relay of clinical data to providers;  
|                   | - Audit and feedback;  
|                   | - Provider education;  
|                   | - Patient education;  
|                   | - Patient reminder systems;  
|                   | - Promotion of self-management;  
|                   | - Organisational change; and  
|                   | - Financial incentives, regulation and policy |
| Comparison        | Not relevant given this review is not focused on any particular comparator. |
| Outcomes          | Not relevant given this is a methodological review. |
| Study designs     | Interrupted time series |

strategies, we identified interventions targeting health systems, health providers and/or patients for inclusion.\(^4\)\(^5\) Studies that did not have at least three preintervention and postintervention data points were excluded, consistent with the Cochrane EPOC recommendation.\(^15\)

Similarly, given that this review focused on empirical applications of ITS, systematic reviews/meta-analyses, letters, commentaries, methodological contributions, study protocols and conference abstracts were excluded.

Data extraction and analysis

The titles and abstracts of unique results from all databases and grey literature were reviewed independently by two researchers (CH and HR) for potential inclusion. The full texts of studies retained at the title and abstract screening stage were retrieved and independently assessed for inclusion by two researchers (CH and HR). Any discrepancies were resolved through discussion and consensus, with a third researcher (MRL) consulted wherever necessary. Data extraction on each included study was conducted using a data extraction tool, developed and piloted a priori drawing on a checklist on the methodological and reporting recommendations for ITS studies adapted from the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement (online supplemental appendix table 3).\(^20\)

Data elements included study characteristics (eg, publication year, country), intervention characteristics (eg, QI strategies, single/multiple interventions), methodological considerations (eg, ITS model, autocorrelation, use of control). We synthesised extracted data narratively using descriptive statistics and following the review protocol registered in PROSPERO. Data extraction was done by one researcher (CH); however, a validation was done by a second researcher (HR) who verified a random subsample consisting of 10% of eligible studies. The agreement between the two researchers (CH and HR) was very high (>95%). Additionally, where necessary, CH consulted with the researcher team members during data extraction. Lastly, assessment of the risk of bias of included studies was performed by two researchers (CH and HR).

Assessment of risk of bias

The risk of bias assessment was guided by the Cochrane Effective Practice and Organisation of Care (EPOC) standard criteria for assessing risk of bias for ITS studies.\(^17\)\(^27\)\(^41\) Specifically, we assessed the following EPOC standard criteria:

1. Was the intervention independent of other changes?  
2. Was the shape of the intervention effect prespecified?  
3. Was the intervention unlikely to affect data collection?  
4. Was the primary outcome measured objectively?  
5. Were incomplete outcome data adequately addressed?  
6. Was the study free of selective outcome reporting?  
7. Was the study analysed appropriately using interrupted time series techniques?

Criteria 1, 3 and 4 assess the threat of history, instrumentation and testing, respectively. Criterion 2 assesses whether the ITS impact model was specified a priori, while criterion 5 assesses whether missing data were dealt with appropriately. Criterion 6 assesses whether all relevant outcomes that were part of the study objectives were reported. Lastly, criterion 7 assesses whether data were analysed appropriately such as using appropriate models (eg, segmented regression and ARIMA) and considering key methodological recommendations (eg, autocorrelation and seasonality).

Online supplemental appendix table 4 describes how we performed risk of bias scoring among included studies. Each criterion scored ‘0’ if low risk and ‘1’ otherwise. For each study, we created an aggregate score by combining scores across the seven criteria. Subsequently, the study-level risk of bias aggregate score was categorised as low (score=0), moderate (score=1 or 2), high (score=3 or 4) and very high (score >4). Trends in risk of bias over time were also assessed. The time for each included study was defined as the year of the publication. Time trend for risk of bias was assessed by fitting a logistic regression model comparing the proportion of studies with a low or moderate risk of bias to the studies with high or very high risk of bias as has previously been done.\(^42\) As the literature search end date was June 2018, studies published in 2018 that were captured by our search strategy may not have been representative of all studies published in 2018. As such, for time trend analyses, studies published in 2018 were excluded. All statistical analyses were conducted using R version 4.0.2
Patient and public involvement
This research was done without patient involvement. Patients were not invited to comment on the study design and were not consulted to develop patient relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy.

RESULTS
The search strategy identified 4061 results, of which 2028 unique eligible records were screened for inclusion. Abstract and full-text screening identified 120 studies that met the inclusion criteria. A summary of this process is presented in the PRISMA flow diagram in figure 3.

Characteristics of included studies
The 120 studies represented a diversity of QI strategies from different geographical settings. Included studies were from 25 countries, with 106 (88.3%) from high-income countries and 14 (11.7%) from low- and middle-income countries (LMICs) as defined by the 2019 World Bank Classification (online supplemental appendix figure 2). The number of studies published over time increased considerably, with 104 (86.7%) studies published in 2010 or later (online supplemental appendix figure 3). The included studies assessed eight QI strategies, with provider education (39.2%), audit and feedback (30.0%), organisational change (24.2%) and provider reminder systems (19.2%) being the most reported QI strategies (figure 4). Nearly half of studies were multisite (44.2%).

Reporting and methodological considerations
Interrupted time series methodological consideration of included studies
All included studies reported the use of an ITS design in their titles and/or abstracts and provided clear timing related to the QI implementation (table 3). Nearly all studies provided a background/rationale for using ITS (98.3%) and provided a description of the intervention (96.7%). A fifth of studies reported on evaluation of multiple phases of QI interventions, of which 76.0% adjusted for multiple phase of QI interventions in their analyses. All included studies reported on the study population, inclusion criteria and outcomes. Nearly all studies (99.2%) used data collected at regular intervals and reported data sources, of which 57% were administrative...
data including Health Management Information System data, 35% were patient records and 8% were survey data. Most studies (71.4%) used monthly data points, with a median preintervention and postintervention period of 18 months (range, 3–120) and 20 months (range 4–90), respectively (table 3). Less than half (42.5%) of studies provided a rationale for the number and spacing of data points described. Just over half reported the unit of analysis clearly (42.5% used data aggregated at the study site level and 9.2% used disaggregated data). Most studies used ITS without a control (81%) and modelled data using segmented regression (62.5%), while 1 in 10 (11.6%) did not report what model was used. For studies that included a control group, 59.1% of studies used a location-based control and about half (54.5%) modelled intervention and control group in a separate ITS model.

About half (55%) of studies reported checking/adjusting for autocorrelation in their ITS models, with the Durbin-Watson test statistic being the test frequently used to check for autocorrelation (36.4%). A fifth (20.8%) of studies reported checking and adjusting for seasonality. Similarly, only 10 (8.3%) studies reported checking/adjusting for non-stationarity in their ITS models, with the augmented Dickey-Fuller test being the most frequently used (70%). Stratified analysis showed that these items were more likely to be reported with ARIMA models compared with segmented regression (online supplemental appendix table 5). Only about half (49.2%) of studies specified the ITS impact model. Over a tenth (12.5%) of studies considered sensitivity analyses.

**Figure 4** Type of quality improvement (QI) strategy reported in the included studies.

Most studies (89.2%), reported CI or SE (81.7%), and reported graphical figures displaying results (92.5%). Of the studies that included graphical figures, 57.6% showed observed trends and 12.6% showed the counterfactual. Further, 29.2% of studies reported absolute and/or relative changes and their significance, for example, clinical and policy. Most studies reported key results and discussed potential limitations of the study. Less than half of studies explicitly discussed potential history threats such as co-interventions. Similarly, 42.1% of studies that mentioned potential limitations also discussed the direction or magnitude of these potential bias(es). About a quarter of studies discussed the stability of outcome data over time.

**Risk of biases of included studies**

Risk of bias was low for 3 (2.5%) studies, moderate for 30 (25.0%), high for 64 (53.3%) and very high for 23 (19.2%). The proportion of studies with a low or moderate risk of bias did not significantly change over time (OR 1.15, 95% CI 0.96 to 1.40, p=0.14) (figure 5). Figure 6 shows the risk of bias across the seven EPOC standard criteria for assessing bias in ITS studies. Details on risk of bias assessment for each included study are provided elsewhere.6

**DISCUSSION**

Over the past few decades, ITS has been an increasingly popular method to evaluate QI interventions in healthcare. We investigated whether the use of ITS in such evaluations has followed best practice standards and recommendations. Our findings showed variations in methodological considerations and reporting of ITS in QI. In particular, we found that most studies identified used ITS without a control and only about half of studies specified the ITS impact model and considered autocorrelation in

---

BMJ Global Health 2020;5:e003567. doi:10.1136/bmjgh-2020-003567

Consistent with previous systematic reviews, this review highlights variations in methodological considerations (or application) and reporting of ITS.19–21 For example, the ITS analyses. Of concern, the risk of bias was high or very high in about three-quarters of included studies and did not appear to have changed significantly over time.

While ITS is arguably one of the strongest quasi-experimental designs available, ITS studies may be subjected to several threats to internal validity.10 18 22 These internal validity threats can often be minimised by adding a control time series group to ITS where feasible.10 18 22 24 However, this review found that a control group was used in less than 20% of included studies, with a location-based control being the most commonly used.

When QI interventions were multisite and implemented at different times, approaches used to assign control were not often discussed clearly. Most of studies included in this review were from high-income countries; however, with expanding routine health information systems in LMICs, there is opportunity to conduct well-designed ITS studies for evaluating healthcare interventions including QI in these settings.25 163

Table 3  ITS methodological considerations of included studies

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITS design reported in the title and/or abstract</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>Background/rationale reported</td>
<td>118</td>
<td>98.3</td>
</tr>
<tr>
<td>Study objectives reported</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>Description of QI intervention</td>
<td>116</td>
<td>96.6</td>
</tr>
<tr>
<td>Start (and end) of QI intervention reported</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>Multiple QI interventions</td>
<td>25</td>
<td>20.8</td>
</tr>
<tr>
<td>▶ 2 interventions</td>
<td>22</td>
<td>88</td>
</tr>
<tr>
<td>▶ 3 interventions</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>▶ 4 interventions</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Adjusted multiple interventions in the analyses</td>
<td>19</td>
<td>76</td>
</tr>
<tr>
<td>Study setting reported</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>Multisite</td>
<td>53</td>
<td>44.2</td>
</tr>
<tr>
<td>Study period reported</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>Study population reported</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>Cohort definition reported</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>Inclusion criteria reported</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>Data sources reported</td>
<td>119</td>
<td>99.2</td>
</tr>
<tr>
<td>Data completeness and validity reported</td>
<td>14</td>
<td>11.6</td>
</tr>
<tr>
<td>Time point intervals, monthly</td>
<td>85</td>
<td>71.4</td>
</tr>
<tr>
<td>Data collected regularly (regular interval)</td>
<td>119</td>
<td>99.2</td>
</tr>
<tr>
<td>Time points clearly reported</td>
<td>119</td>
<td>99.2</td>
</tr>
<tr>
<td>Rationale for the number and spacing of data points described</td>
<td>51</td>
<td>42.5</td>
</tr>
<tr>
<td>Outcome measure(s) reported</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>Format of outcome(s) reported</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>Unit of analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Aggregated</td>
<td>51</td>
<td>42.5</td>
</tr>
<tr>
<td>▶ Individual</td>
<td>11</td>
<td>9.2</td>
</tr>
<tr>
<td>▶ Not reported</td>
<td>58</td>
<td>48.3</td>
</tr>
<tr>
<td>ITS models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Segmented regression</td>
<td>75</td>
<td>62.5</td>
</tr>
<tr>
<td>▶ ARIMA</td>
<td>19</td>
<td>15.8</td>
</tr>
<tr>
<td>▶ Other models (eg, GEE, linear regression, mixed effect model, spline regression, poison regression, Prais-Winsten regression, logistic regression)</td>
<td>17</td>
<td>14.2</td>
</tr>
<tr>
<td>▶ Not reported</td>
<td>14</td>
<td>11.6</td>
</tr>
<tr>
<td>Autocorrelation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Checked/adjusted as appropriate</td>
<td>66</td>
<td>55</td>
</tr>
<tr>
<td>Test(s) used to check for autocorrelation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Durbin-Watson test statistic</td>
<td>24</td>
<td>36.4</td>
</tr>
<tr>
<td>Continued</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 Continued

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ Other tests (eg, ACF and PACF, Ljung-Box χ² test, residual plots and Breusch-Godfrey test, Breusch-Godfrey test, Cumby-Huizinga test)</td>
<td>10</td>
<td>15.2</td>
</tr>
<tr>
<td>▶ Not reported</td>
<td>33</td>
<td>50</td>
</tr>
<tr>
<td>Seasonality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Checked/adjusted</td>
<td>25</td>
<td>20.8</td>
</tr>
<tr>
<td>Non-stationarity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Checked/adjusted</td>
<td>10</td>
<td>8.3</td>
</tr>
<tr>
<td>Test(s) used to check</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Augmented Dickey-Fuller test</td>
<td>7</td>
<td>70</td>
</tr>
<tr>
<td>▶ Not reported</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Control group used</td>
<td>22</td>
<td>18.3</td>
</tr>
<tr>
<td>Type of control, location-based control</td>
<td>13</td>
<td>59.1</td>
</tr>
<tr>
<td>Analyses of controlled ITS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Combined</td>
<td>5</td>
<td>22.7</td>
</tr>
<tr>
<td>▶ Separated</td>
<td>12</td>
<td>54.5</td>
</tr>
<tr>
<td>▶ Difference</td>
<td>2</td>
<td>9.1</td>
</tr>
<tr>
<td>▶ Not reported</td>
<td>3</td>
<td>13.6</td>
</tr>
<tr>
<td>Specify ITS impact model (or provided basic ITS model structure)</td>
<td>59</td>
<td>49.2</td>
</tr>
<tr>
<td>Use of lag period</td>
<td>34</td>
<td>28.4</td>
</tr>
<tr>
<td>Sensitivity analyses</td>
<td>15</td>
<td>12.5</td>
</tr>
<tr>
<td>Reported statistical software used</td>
<td>97</td>
<td>80.8</td>
</tr>
</tbody>
</table>

ACF, autocorrelation function; ARIMA, autoregressive integrated moving average; GEE, generalised estimating equation; ITS, interrupted time series; PACF, partial autocorrelation function; QI, quality improvement.;
Table 4  Reporting of ITS study results and interpretation

<table>
<thead>
<tr>
<th>Results</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Participants</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>► Characteristics in each study group</td>
<td>84</td>
<td>70</td>
</tr>
<tr>
<td>► Flow diagram of study participant selection</td>
<td>11</td>
<td>9.2</td>
</tr>
<tr>
<td><strong>Outcomes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>► Reported all outcomes examined over the study period</td>
<td>105</td>
<td>87.5</td>
</tr>
<tr>
<td>► Report the average, minimum and maximum number of outcomes across time intervals</td>
<td>22</td>
<td>18.3</td>
</tr>
<tr>
<td>► Reported level/trend changes</td>
<td>107</td>
<td>89.2</td>
</tr>
<tr>
<td>► Report absolute and/or relative changes and their significance, eg, clinical, policy and statistical</td>
<td>35</td>
<td>29.2</td>
</tr>
<tr>
<td>► Report CI or SE</td>
<td>98</td>
<td>81.7</td>
</tr>
<tr>
<td>► Graphical figures to display results</td>
<td>111</td>
<td>92.5</td>
</tr>
<tr>
<td>► Fitted lines (trend)</td>
<td>64</td>
<td>57.6</td>
</tr>
<tr>
<td>► Counterfactual lines</td>
<td>14</td>
<td>12.6</td>
</tr>
<tr>
<td>► Used time lag and showed it on figure in results</td>
<td>23</td>
<td>67.6</td>
</tr>
<tr>
<td>► Results of sensitivity analyses if relevant</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td><strong>Interpretation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>► Key results</td>
<td>119</td>
<td>99.2</td>
</tr>
<tr>
<td>► Context (related to possible confounding)</td>
<td>113</td>
<td>97.4</td>
</tr>
<tr>
<td>► Relevant co-interventions</td>
<td>56</td>
<td>46.7</td>
</tr>
<tr>
<td>► Stability of the participant characteristics over time</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>► Stability of outcome coding over time</td>
<td>28</td>
<td>23.3</td>
</tr>
<tr>
<td><strong>Limitations of the study</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>► Discussion of limitations of the study</td>
<td>114</td>
<td>95</td>
</tr>
<tr>
<td>► Data variability/appropriateness of number data points</td>
<td>31</td>
<td>25.8</td>
</tr>
<tr>
<td>► Discussion direction/magnitude of any potential bias</td>
<td>48</td>
<td>42.1</td>
</tr>
</tbody>
</table>

Jandoc et al’s (2015) systematic review of application of ITS in drug utilisation research found that about three-quarters of studies examined drug utilisation over monthly intervals, and two-thirds used segmented regression techniques. They also found that key methodological considerations in ITS including seasonality and non-stationarity were considered in less than half of studies. However, autocorrelation was considered in two-thirds of studies included in Jandoc et al’s review unlike in our review where it was considered in only about half of studies.

Our findings showed that considerations of autocorrelation, seasonality and non-stationarity were more likely to be reported when ARIMA was used compared with segmented regression, which mirrors prior reviews in other subject areas. Similar to this review, a more recent review by Hudson et al (2019) examined characteristics and reporting of single ITS in healthcare in studies published in 2015 and found that only about half of studies considered autocorrelation, 28% considered seasonality and 10% considered non-stationarity. Consistent with previous reviews, this review showed that studies used various statistical approaches that may have different effect on study results, highlighting a need for further research on how various statistical analyses used in ITS can affect intervention estimates.

While the number of ITS studies has increased over time, the conduct and reporting of ITS studies still leaves room for improvement. Consistent with previous reviews, variations in methodological considerations and reporting of ITS remain a concern, warranting a need to develop and reinforce formal reporting guidelines to improve the application of ITS in the evaluation of...
effectiveness of healthcare interventions. Jandoc et al (2015) proposed methodological and reporting recommendations for ITS studies that built on STROBE, and Bernal (2018) proposed a framework for enhanced reporting of ITS framework that can be adapted to help improve design and reporting of ITS in the evaluation of QI.20 23 Additionally, the Cochrane EPOC standard criteria for assessing risk of bias for ITS should be used to improve the design of ITS in the evaluation of QI to reduce risk of bias.14 27 Although these EPOC standard criteria do not provide guidance on how to account for key methodological considerations such as autocorrelation, non-stationarity, seasonality and multiple interventions, there are other resources that are accessible to QI evaluators to help conduct ITS analyses appropriately, such as Wagner et al’s seminal paper on segmented regression analysis of ITS, Bernal et al’s tutorial on use of ITS in the evaluation of public health interventions and Saeed et al’s work on use of segmented generalised mixed effects models to evaluate health outcomes.18 20 31

This methodological systematic review has a number of limitations that need to be highlighted. First, our results are based on what were reported (or not) as performed in eligible studies; we did not reanalyse reported time series data to check whether the statistical analyses undertaken in these studies were performed appropriately. As such, it is possible that we underestimated the risk of bias, especially as related to whether the study data were analysed appropriately using time series techniques—one of the seven EPOC standard criteria for assessing risk of bias for ITS studies. Ramsay and colleagues reanalysed ITS studies data from two systematic reviews and found that approximately 50% of studies that had reported a significant difference were not actually significant when ITS methods were applied appropriately.17 Second, although we did not apply any restrictions to our search strategy, we could still have missed some relevant studies. Lastly, while the recommended conventional double screening was employed in this study, data extraction was done by a single researcher.164 However, a validation on a random subsample consisting of 10% of eligible studies found a very high agreement, making the data extraction less of a concern. In conclusion, while use of ITS in the evaluation of health system QI interventions has increased considerably over the past decade, the quality of ITS studies still shows room for improvement. Moreover, variations in methodological considerations and reporting of ITS in QI research remain a concern, warranting a need to develop and reinforce formal reporting guidelines to improve application of ITS.

Author affiliations
1Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA
2Centre for Health Services and Policy Research, School of Population and Public Health, The University of British Columbia, Vancouver, British Columbia, Canada
3School of Public Health, University of Rwanda, Kigali, Rwanda
4School of Population and Public Health, The University of British Columbia, Vancouver, British Columbia, Canada
5HIV/STI Surveillance Research Centre, and WHO Collaborating Centre for HIV Surveillance, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran
6Collaboration for Outcomes Research and Evaluation, Faculty of Pharmaceutical Sciences, The University of British Columbia, Vancouver, British Columbia, Canada
7Center for Health Evaluation and Outcome Sciences, Providence Health Research Institute, Vancouver, British Columbia, Canada

Twitter Celestin Hategeka @DrHategeka and Michael R Law @myclaw

Acknowledgements Dr Hategeka received support through a Vanier Canada Graduate Scholarship and a Banting Postdoctoral Fellowship from the Canadian Institutes of Health Research. Dr Karmouzian is supported by a Vanier Canada Graduate Scholarship and a Pierre Elliott Trudeau Foundation Doctoral Scholarship. Dr Law received salary support through a Canada Research Chair in Access to Medicines and a Michael Smith Foundation for Health Research Scholar Award.

Contributors CH conceived the idea and designed the study. MRL provided guidance throughout the research process. CH, MK and MRL contributed to the search strategy. CH and HR performed screening of titles, abstracts and full texts. CH and HR extracted data from eligible studies and performed risk of bias assessment. CH analysed the data and drafted the manuscript. All authors reviewed the manuscript and contributed to the revision of the manuscript and approved its final version.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

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

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD
Celestin Hategeka http://orcid.org/0000-0001-7808-4652

REFERENCES
BMJ Global Health

10 Soumerai SB, Ceccherelli R, Rujogim K. False dichotomies and health policy research designs: randomized trials are not always the answer. JAMA 2017;32:204–9.
19 Polus S, Pieper D, Burns J, et al. Heterogeneity in application, design, and analysis characteristics was found for controlled before-after and interrupted time series studies included in Cochrane reviews. J Clin Epidemiol 2017;91:56–69.
23 López Bernal J. The use of interrupted time series for the evaluation of public health interventions. London School of Hygiene & Tropical Medicine, 2011.
41 Cochrane Effective Practice and Organisation of Care (EPOC). EPOC resources for review authors, 2017.
58 Wagner AD, Mugo C, Blumer-Mirote S, et al. Continuous quality improvement intervention for adolescent and young adult HIV


