

# A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation

Atle Fretheim<sup>a,b,c,\*</sup>, Fang Zhang<sup>a</sup>, Dennis Ross-Degnan<sup>a</sup>, Andrew D. Oxman<sup>b</sup>, Helen Cheyne<sup>d</sup>, Robbie Foy<sup>e</sup>, Steve Goodacre<sup>f</sup>, Jeph Herrin<sup>g,h</sup>, Ngaire Kerse<sup>i</sup>, R. James McKinlay<sup>j</sup>, Adam Wright<sup>k</sup>, Stephen B. Soumerai<sup>a</sup>

<sup>a</sup>Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 133 Brookline Avenue, Boston, MA 02215, USA

<sup>b</sup>Global Health Unit, Norwegian Knowledge Centre for the Health Services, PO Box 7004, St. Olavs pl, 0130 Oslo, Norway

<sup>c</sup>Department of Community Medicine, Institute of Health and Society, University of Oslo, PO Box 1130 Blindern, 0318 Oslo, Norway

<sup>d</sup>Nursing Midwifery and Allied Health Professions Research Unit, University of Stirling, Stirling FK9 4LA, UK

<sup>e</sup>Academic Unit of Primary Care, Leeds Institute of Health Sciences, University of Leeds, Charles Thackrah Building, 101 Clarendon Road, Leeds LS2 9LJ, UK

<sup>f</sup>School of Health and Related Research (SCHARR), University of Sheffield, Regent Court, Regent Street, Sheffield, S1 4DA, UK

<sup>g</sup>Division of Cardiology, Yale University School of Medicine, Yale University, 333 Cedar St, New Haven, CT 06510, USA

<sup>h</sup>Health Research & Educational Trust, 155 N Wacker, Suite 400, Chicago 60606, IL, USA

<sup>i</sup>School of Population Health, University of Auckland, Private Bag 9201, Auckland, New Zealand

<sup>j</sup>Health Information Research Unit, Department of Clinical Epidemiology & Biostatistics, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada

<sup>k</sup>Department of Medicine, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, USA

Accepted 17 October 2014; Published online 11 December 2014

## Abstract

**Objectives:** There is often substantial uncertainty about the impacts of health system and policy interventions. Despite that, randomized controlled trials (RCTs) are uncommon in this field, partly because experiments can be difficult to carry out. An alternative method for impact evaluation is the interrupted time-series (ITS) design. Little is known, however, about how results from the two methods compare. Our aim was to explore whether ITS studies yield results that differ from those of randomized trials.

**Study Design and Setting:** We conducted single-arm ITS analyses (segmented regression) based on data from the intervention arm of cluster randomized trials (C-RCTs), that is, discarding control arm data. Secondly, we included the control group data in the analyses, by subtracting control group data points from intervention group data points, thereby constructing a time series representing the difference between the intervention and control groups. We compared the results from the single-arm and controlled ITS analyses with results based on conventional aggregated analyses of trial data.

**Results:** The findings were largely concordant, yielding effect estimates with overlapping 95% confidence intervals (CI) across different analytical methods. However, our analyses revealed the importance of a concurrent control group and of taking baseline and follow-up trends into account in the analysis of C-RCTs.

**Conclusion:** The ITS design is valuable for evaluation of health systems interventions, both when RCTs are not feasible and in the analysis and interpretation of data from C-RCTs. © 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>).

**Keywords:** Evaluation methods; Randomized trials; Interrupted time-series; Quasi-experimental design; Impact evaluations; Health services research

**Funding:** A.F. conducted this work through the Harvard Medical School Fellowship Program in Pharmaceutical Policy Research, as a Harkness Fellow (Commonwealth Fund and the Norwegian Research Council) and Fulbright Scholar (US—Norway Fulbright Foundation). The research was performed without further external financial or material support.

**Conflict of interest:** R.F. reports grants from Reckitt Benckiser, outside the submitted work; no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years; and no other relationships or activities that could appear to have influenced the submitted work.

\* Corresponding author. Tel.: +47 23255000; fax: +47 23255010.

E-mail address: [atle.fretheim@nokc.no](mailto:atle.fretheim@nokc.no) (A. Fretheim).

## 1. Introduction

The randomized controlled trial (RCT) is widely regarded as the gold standard research design for measuring the impacts of interventions, and RCTs dominate effectiveness research in clinical medicine.

The field of health system and policy evaluation is very different: randomized trials are seldom carried out, despite substantial uncertainty about the impacts of health system interventions on the costs and outcomes of care. This is

**What is new?**

- Our findings support the position that concurrent control groups are important, but the single-arm interrupted time-series design, in which the preintervention period serves as control, yielded findings that were usually concordant with the randomized controlled trials (RCTs).
- If data from RCTs are analyzed without taking baseline and follow-up trends into account, our results indicate that the findings may sometimes be misleading.
- Those who commission or conduct impact evaluations of health system interventions should routinely use graphical displays of longitudinal data and time-series analysis methods when evaluating intervention effects, whether randomization is feasible or not.

due in part to practical difficulties encountered when conducting randomized trials of health system interventions, such as nationwide reforms (eg, introducing user fee exemptions for pregnant women and children). Therefore, other study designs are often used in this field. The simplest pre–post design uses single observations before and after an intervention to evaluate whether a change occurs. However, because factors other than the intervention (eg, secular trends) may cause an observed change (or lack of change), this is considered a weak method [1,2]. An extension of this approach is the single-arm interrupted time-series (ITS) design, where multiple measurements are carried out before and after an intervention, which can control for preintervention and postintervention trends [3–5]. The ITS method is widely recommended for impact evaluation of system and policy changes and has been promoted as “a particularly strong quasi-experimental alternative to randomized designs when the latter are not feasible” [6].

It is widely recognized that different study designs differ in internal validity and various study designs are sometimes placed in a hierarchy [7]. In these hierarchies, randomized trials are typically rated above nonrandomized studies, including ITS studies. Most systematic reviews published through the Cochrane Collaboration only include randomized trials. Among the exceptions are systematic reviews from the “Effective Practice and Organisation of Care” review group, which often include nonrandomized studies (including ITS studies) [8]. Findings from ITS studies are, however, generally considered to have a higher risk of bias than findings from RCTs [9]. This is based on the logical argument that only randomization is able to control for confounders that are not known or measured, whereas other study designs can only control for

confounders that are known and measured [10]. Studies have investigated the effectiveness of randomization in limiting selection bias (and thus ensuring comparable groups in effectiveness evaluations), but this work has mainly focused on clinical trials and less on health system and policy interventions [11]. Also, previous research comparing the results from randomized trials with those from other study designs has often lumped together many different types of nonrandomized studies. This may be inappropriate because all nonrandomized study designs are not equally prone to bias.

An overview of existing reviews addressed the issue of whether RCTs provide the same effect size and variance as nonrandomized studies of similar policies [12]. The authors reported that in many cases, the effect sizes from RCTs differed from nonrandomized studies. Consequently, they concluded that “policy evaluations should adopt randomized designs whenever possible.” However, ITS analyses were not considered separately in that report.

There are few empirical data from which to draw firm conclusions regarding the relative merits of different study designs for effectiveness evaluations. Debates on this topic are largely based on theoretical arguments. This is particularly the case for ITS because little has been done to compare findings from ITSs and RCTs in a systematic way.

In practice, randomized experiments of system interventions are almost invariably cluster randomized trials (C-RCTs) that randomize groups rather than individuals (eg, clinics, hospitals, communities). We recently conducted a re-analysis of one C-RCT and found that estimates from ITS analyses of the intervention arm only (single-arm ITS), and incorporating both the intervention and control groups (controlled ITS), were concordant with the C-RCT result [13]. Additional comparisons of the same sort would help to determine whether those findings can be generalized.

The aim of this article was to further explore whether ITSs yield results that differ from those of cluster randomized controlled trials (C-RCTs) and to identify possible explanatory factors for such differences. Our primary objective was to compare each trial result with the effect estimate based on the single-arm ITS (ie, only intervention group data, discarding the control group). In addition, we conducted ITS analyses incorporating data from both arms of each trial.

## 2. Methods

The full study protocol is found in the [Appendix](http://www.jclinepi.com) at [www.jclinepi.com](http://www.jclinepi.com).

### 2.1. Search for trials

We searched for C-RCTs of health system interventions where data were available for a series of time points before and after the interventions were implemented. The amount of data had to be sufficient to allow for meaningful ITS

analysis, that is, a minimum of six time points both before and after the intervention was implemented with a minimum of 50 observations per time point. In addition, we allowed for a maximum of six missing time points between the preintervention and postintervention periods (transition period).

As a secondary analysis, we included studies with only 3–5 time points before or after the intervention.

Our main source of candidate trials was an inventory of C-RCTs [14]. Every entry in the C-RCT inventory was coded according to the type of intervention that was evaluated. We retrieved full text reports of trials of interventions coded as financial arrangements, delivery arrangements, governance arrangements, or implementation strategies. We only included trials published in 2000 or later and reported in peer-reviewed publications written in English. As the trial inventory only included trials published up to 2010, we conducted a supplementary PubMed search for C-RCTs up to July 2012.

We concealed the results and discussion sections in the retrieved articles using 3M Post-it notes and attempted to remain blinded to the original results until after our analyses had been completed.

All full text reports were read by one of us (A.F.). If a trial was deemed potentially usable for our purpose, the corresponding author was contacted by e-mail and asked to share the trial data file with us. If we received no response after repeated attempts, we tried to contact one of the coauthors.

## 2.2. Analysis

We conducted our primary analysis on the intervention arm of the trials only, thus yielding effect estimates that the investigators would have found had they used a single-arm ITS (ie, no concurrent control group) rather than a C-RCT as their evaluation method. To the extent possible, we conducted our analyses on the trials' primary outcomes.

We reanalyzed the trial data using a basic segmented regression method [15,16]:

$$Y_t = \beta_0 + \beta_1 \times \text{Time}_t + \beta_2 \times \text{Intervention}_t + \beta_3 \times \text{Time after intervention}_t + \varepsilon_t$$

Here,  $Y_t$  is the dependent variable score at time  $t$ .  $\text{Time}_t$  is the value of the time variable "Time" at time  $t$ . "Intervention" is the value of the level-change variable (ie, 0 before the intervention and 1 after the intervention); for interventions that are gradually introduced, the time points during the transition phase are set to missing. "Time after intervention" is the value of the slope change variable defined as  $[\text{Time} - (n_1 + 1)] \times \text{Intervention}$ , where  $n_1$  is the total number of time points before the postintervention period.  $\beta_0$  is the intercept,  $\beta_1$  represents the modeled slope during the preintervention period,  $\beta_2$  the level change after the intervention, and  $\beta_3$  the change in slope from before to

after the intervention.  $\varepsilon_t$  is the error term representing the variability not explained by the model.

We ran ordinary least squares regressions and used the Durbin–Watson test to assess the degree of first-order autocorrelation [17]. We used the Prais–Winstone method to adjust for autocorrelation, when deemed necessary based on the Durbin–Watson test results.

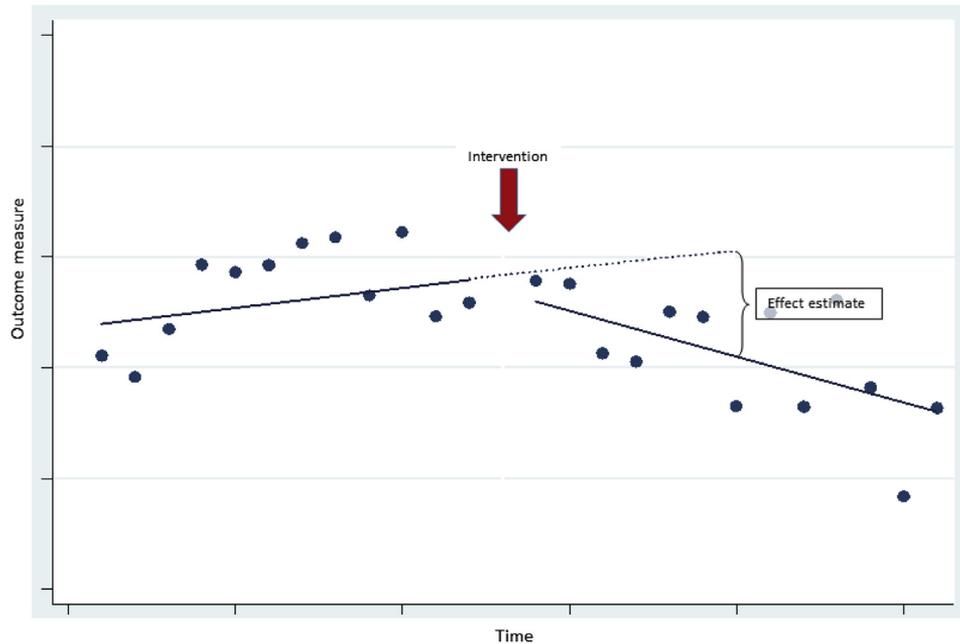
We prespecified intervention transition periods (ie, the period during which an intervention was being implemented before measurement of postintervention outcomes began) in the ITS analyses. We typically used months as the observation period, as is often done in ITS studies in our field [15]. However, we opted for other time points (eg, weeks) when we judged that to be more appropriate (ie, if there were sufficient data to allow for further disaggregation into more time points).

As is customary for ITS analyses, we computed two effect estimates: (1) the change in level and (2) change in trend (slope) from before to after the intervention. To include both effect estimates in one metric, for comparison against the C-RCT effect estimate, we modeled the ITS estimate halfway through the postintervention period (ie, the difference between the levels of the preintervention and the postintervention regression lines halfway through the postintervention period; see Fig. 1).

We also conducted our own C-RCT analyses, to ensure that the ITS and C-RCT estimates were as comparable as possible, for example, that they were based on the exact same data. In some cases, we discarded parts of the data set to make it amenable for time-series analysis, as specified in the Section 3. Consequently, our C-RCT estimates in some cases differ from the estimates in the original publications. Our C-RCT effect estimates were recalculated by comparing the postintervention observations in the intervention and control groups, using baseline levels for each cluster as covariates (analysis of covariance) in a generalized estimating equation (logistic regression using the logit link function for binary outcomes and the log link function for count outcomes) [18]. The model incorporated all data while controlling for clustering effects at the sampling unit level. We used the margins command to turn the resulting odds ratios into percentages, at population mean of other covariates [19,20]. Trends over time within the preintervention or postintervention periods were not taken into account in this model.

We also conducted a controlled ITS analysis of the difference between the intervention and control groups in the C-RCT. By subtracting the value of the outcome variable in the control group from the corresponding value in the intervention group at each time-series data point, we constructed a new time series of differences between the two groups. This time series was used to calculate the difference in slope and level changes between the intervention and control groups.

Finally, we consulted with authors of the included studies to consider possible explanatory factors for differences between the C-RCT and ITS findings.



**Fig. 1.** Example illustrating how level and trend changes were combined in one effect estimate: the difference between the level of the preintervention regression line and the postintervention line halfway through the postintervention period.

All statistical analyses were done in Stata 12.0 (College Station, TX) using the REGRESS, PRAIS, ESTAT DWATSON, and XTGEE commands.

### 2.3. Deviation from study protocol

We did not conduct ITS analyses on patient level data. In our previous study, such analyses added little to our main aggregated ITS approach [13]. Also, we decided to limit our analyses to absolute effect measures, as it proved technically difficult to estimate and compare relative effect estimates.

## 3. Results

We identified 524 potentially relevant trials for which we retrieved full text reports. From these, we identified 89 as potentially usable for ITS analysis. We managed to make contact with the majority of investigators, although in most cases, they responded that the data were not suitable for our purpose, not possible to share (eg, due to privacy concerns), or no longer available. Eventually, we received 19 data sets, of which eight met our inclusion criteria [21–28]. One of the eight trials [22] evaluated two interventions in two different patient groups, resulting in a total of nine ITS vs. C-RCT comparisons.

### 3.1. Description of included studies

Three of the trials evaluated the impact of system changes in health services [24,26,27]. The five remaining studies were trials of interventions aimed at improving

practices among midwives [21] and primary care physicians [22,23,25,28]. All studies were conducted in high-income countries. The number of observations varied considerably across the trials, see Table 1 for further details.

Only four trials included six or more time points before and after the interventions were included in our main analyses. Most of the remaining studies included five time points before and after the intervention. We assessed these studies separately.

### 3.2. Main findings

An overview of the results is shown in Fig. 2. In eight of the nine cases, the different analytical methods yielded overlapping 95% CIs. For a complete presentation of regression coefficients and autocorrelation statistics from the ITS analyses, see the Appendix at [www.jclinepi.com](http://www.jclinepi.com).

The four data sets with at least six time points before and after the intervention were from the studies by Wright et al. [27], Fretheim et al. [28], Goodacre et al. [24], and Foy et al. [23].

The graphical presentation of the study by Wright (Fig. 3) convincingly displays an immediate and substantial increase in physicians' registration of clinical problems in the medical records, following the introduction of an electronic alert system. No concurrent change among physicians in the control group is visible. The calculated effect estimates were very similar across the different methods (Fig. 2), but the observed waning of effect over time (Fig. 3) would not have been detected in a conventional aggregated C-RCT analysis. According to the study authors

**Table 1.** Included trials

| Trial (setting)  | Intervention   | Main outcome   | Events (time points) <sup>a</sup>  |
|--|--|--|--|
| Wright et al. [27] (clinics affiliated with an academic medical center, MA, USA) | A clinical alerting system that uses inference rules to notify providers of undocumented problems  | The number of study problems added <sup>b</sup>  | 13,551 Problems added (26 before and after)  |
| Fretheim et al. [28] (primary care practices, Norway)                            | Educational outreach visits, computer-based decision support and reminders, and patient educational material   | Prescribing of low-dose diuretics as first-line antihypertensive medication (proportion) | 966 Prescriptions for low-dose diuretics of 9,301 antihypertensive prescriptions (12 before and after)   |
| Goodacre et al. [24] (acute hospitals, UK)                                       | Establishment of chest pain unit   | Chest pain attendances resulting in admission (proportion)                               | 48,115 Admissions of 82,190 attendances (11 before and after)  |
| Foy et al. [23] (primary care practices, England, UK)                            | Brief educational messages added to article and electronic primary care practice laboratory test reports   | HbA1c below 6.35% (proportion)   | 11,882 Tests below 6.35%, of 68,007 tests (24 before and 35 after)   |
| Flottorp et al. [22] (primary care practices, Norway) <sup>c</sup>               | Patient educational material, computer-based decision support and reminders, an increase in the fee for telephone consultations, and interactive courses (for urinary tract infection or sore throat)                        | Use of antibiotics, for urinary tract infection or sore throat (proportion)              | Sore throat: 8,065 prescriptions of 16,939 consultations; urinary tract infection: 4,418 prescriptions of 9,887 consultations (5 before and after) |
| Kerse et al. [26] (residential care homes, New Zealand)                          | Residential care staff, using existing resources, implemented systematic individualized fall-risk management for all residents using a fall-risk assessment tool, high-risk logo, and strategies to address identified risks | Number of residents sustaining a fall, total falls                                       | 2,002 Falls (5 before and 12 after)  |
| Cheyne et al. [21] (maternity units, Scotland, UK)                               | Use of an algorithm by midwives to assist their diagnosis of active labor  | Use of oxytocin for augmentation of labor (proportion)                                   | 736 Uses of oxytocin in 2,195 births (5 before and 9 after)  |
| Haynes et al. [25] (primary care physicians, Ontario, Canada)                    | McMaster PLUS, an internet-based addition to an existing digital library, with quality- and relevance-rated medical literature to physicians   | Number of logins per month per user  | 3,841 Logins (3 before and 12 after)   |

<sup>a</sup> The total number (count) of outcomes included in our analyses and the number of time points included in the interrupted time series analyses, preintervention and postintervention periods, respectively.

<sup>b</sup> This was a prespecified secondary outcome. We used this because the main outcome (“acceptance of the alert”) was not amenable for time-series analysis (no baseline data).

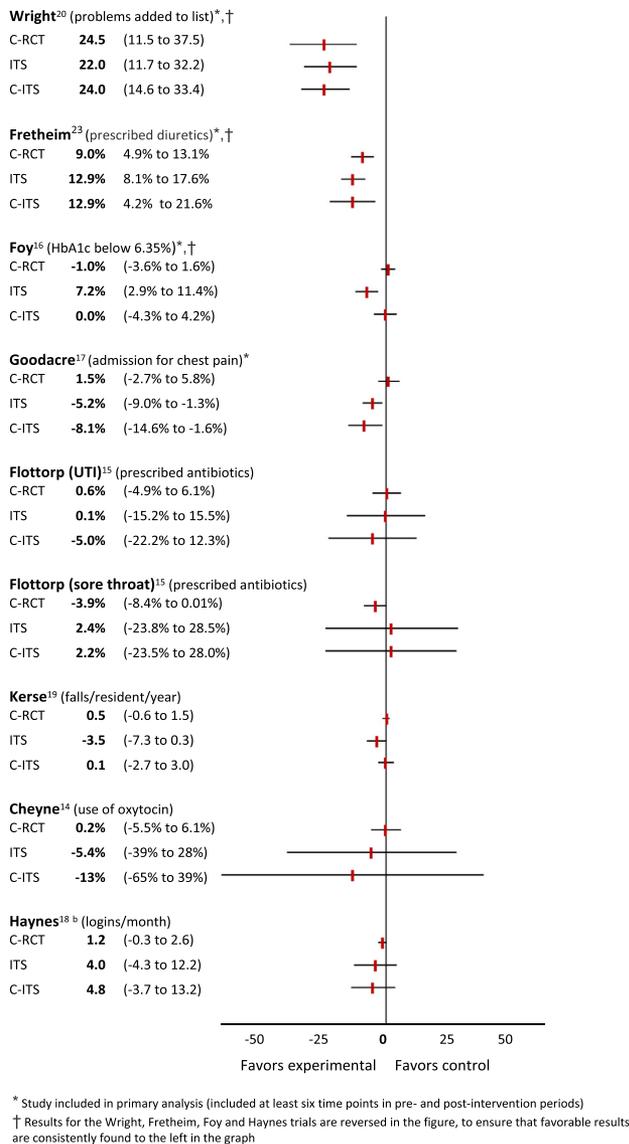
<sup>c</sup> In this trial, half the participating practices received interventions to implement guidelines for urinary tract infection and the other half received interventions to implement guidelines for sore throat, serving as controls for each other.

(A.W.), a likely explanation for the trend back toward baseline behavior is that the number of unregistered clinical problems had accumulated, so there were more problems to register at the beginning of the intervention period.

In the study by Fretheim, an immediate increase in the prescribing of recommended first-line medication (low-dose diuretic for hypertension) was observed in the intervention group when the quality improvement intervention was implemented (Fig. 3). The effect estimates were concordant across methods (Fig. 2).

In the study by Goodacre, we discarded data from the last month before the start of the intervention and from the first month after, because of missing data from these months. The graphical presentation (Fig. 3) shows that among patients presenting at the emergency room, there

was a slight increase in the proportion admitted to hospital in the intervention group during the baseline period, whereas the control hospitals showed a minor trend in the opposite direction. After dedicated chest pain units were introduced, there was a small immediate reduction in admissions, followed by a downward trend in the intervention hospitals. Neither the baseline difference in trends nor the change in trend after the intervention could be detected in the C-RCT analysis, which resulted in a highly uncertain estimate of 1.5 percentage points (95% CI: −2.7, 5.8). In contrast, the single-arm ITS and the controlled ITS analyses found a clear impact of the intervention: −5.2 percentage points (95% CI: −9.0, −1.3) and −8.1 percentage points (95% CI: −14.6, −1.6), respectively. The 95% CIs from all three analyses overlap.



**Fig. 2.** Effect estimates expressed as absolute differences (95% CIs).

The graphical presentation of the data set from the study by Foy revealed that the proportion of patients with diabetes who had achieved an HbA1c goal of 6.35% stopped decreasing when brief educational messages were introduced to physicians (see Fig. 3). Surprisingly, the same pattern occurred among physicians who were not exposed to this intervention. Although the single-arm ITS yielded a discrepant result, the controlled ITS produced a result similar to the C-RCT.

The four data sets that included five time points in one or both periods were from the studies by Flottorp et al. [22], Kerse et al. [26], and Cheyne et al. [21]. Not surprisingly, the effect estimates based on ITS analyses of such few data points were imprecise (ie, wide 95% CIs, see Fig. 2).

The findings from the two data sets from the study by Flottorp were concordant across the various analytical methods (ie, the CIs overlapped, see Fig. 4). However,

the imprecise effect estimates from the ITS analyses limit the usefulness of the comparison.

The graphical presentation of the Kerse data set raises several questions. Why did the rate of falls in the residential homes increase so rapidly during the baseline period? Why did the increase in falls cease in the control group when the intervention was launched in the intervention group but continue to increase in the intervention group? Why did the rate of falls start to decrease in the intervention group several months later? We do not know the answers to these questions. Nonetheless, a simple, conventional comparison of changes from before to after the intervention, as is usually done in C-RCT analyses, reveals none of the confusing underlying dynamics in the occurrence of study outcomes and would be misleading. In this case, breaking the data down into shorter periods offers advantages in drawing inferences from the study. In their original report, the authors presented a simple graph showing the initial rise and subsequent reduction in fall rate associated with the intervention and proposed several possible explanations [26].

The graph illustrating the findings from the study by Cheyne shows a high degree of variation in the proportion of deliveries with oxytocin from 1 month to the next (Fig. 4). This probably reflects random variation due to low numbers of patients (see Table 1). Again, comparing the C-RCT estimate with the ITS estimates is of limited use because of the wide CIs around ITS estimates (see Fig. 2).

Finally, we included one study with only three preintervention time points: the trial by Haynes et al. of an Internet-based medical literature service for physicians. The graphical presentation shows wide variation of rates across time points (Fig. 5), likely due to the low number of events per month (see Table 1) and apparently diverging trends in the two groups in the postintervention period. The high variability and the few time points mean that time-series analysis on these data is unlikely to yield reliable results. Nevertheless, the results were relatively consistent across the different analyses (see Fig. 2).

#### 4. Discussion

We identified eight C-RCTs of health system interventions where the available data enabled us to use time-series methods to estimate the effect of the intervention. However, only four of the trials had sufficient data to allow for six or more time points before and after the intervention, which we considered a minimum threshold for reliable results. The findings were largely concordant, yielding similar results across different analytical methods. However, our analyses revealed limitations with both the conventional approach to analyzing C-RCTs and with the ITS approach. These findings may have implications for the design of future impact evaluations of health system interventions.

First, the value of having a concurrent control group was clearly shown in the analyses based on the trial by Foy et al.

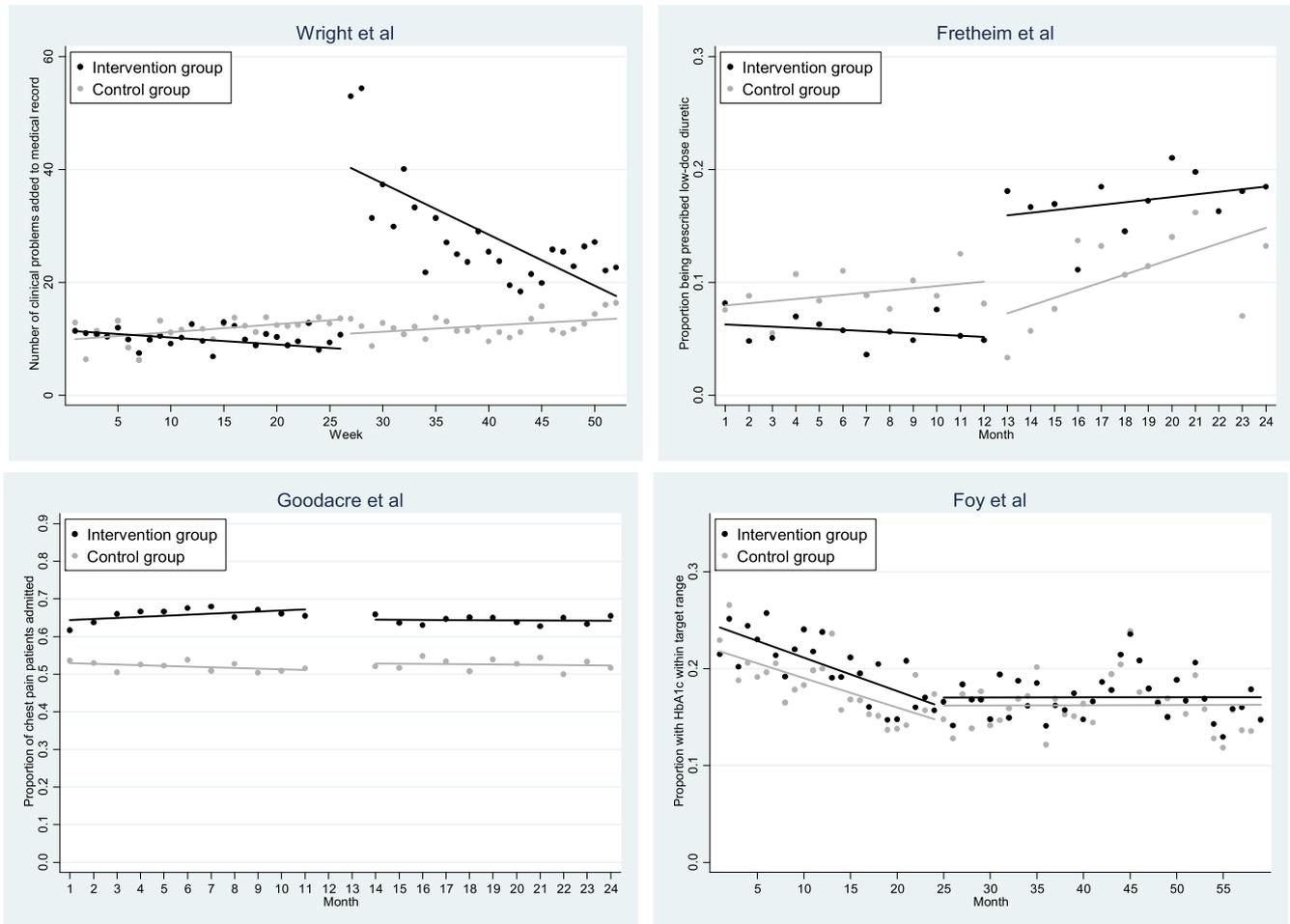


Fig. 3. Linear regression of preintervention and postintervention time series, from studies with a minimum of six time points before and after intervention.

[23]. Had we not known that the observed changes were similar in the intervention and control groups, we would have been misled to believe that the intervention was effective. Thus, the single-arm ITS analysis did not provide a valid effect estimate in this case. Two common threats to the validity of single-arm ITS analyses are “history” (ie, cointerventions—“the possibility that forces other than the treatment under investigation influenced the dependent variable at the same time when the intervention was introduced”) and “instrumentation” (ie, changes in how an outcome is being recorded at the same time as an intervention is being implemented) [6]. We have not been able to identify a plausible explanation as to why both the intervention and control groups changed simultaneously when the intervention was implemented for the study by Foy. However, the benefits of a control group in detecting such anomalous effects are clear, and the use of control groups is to be recommended whenever possible when assessing the effects of health system interventions.

Second, our analyses of the trials by Goodacre et al. [24], Wright et al. [27], and Kerse et al. [26] demonstrate how conventional C-RCT analysis may conceal important dynamics in the preintervention and the postintervention

periods. The most striking example is the Kerse trial. Notable changes that took place during both the baseline and postintervention periods would not have been noticed without disaggregating the data sets into time series. Such dynamics are of key importance in understanding the effects over time and should trigger further qualitative exploration. The Goodacre and the Wright data illustrate the same phenomenon, although in a less striking way: important information may be lost when changes within the preintervention and postintervention periods remain unexplored; such exploration is usually not done in C-RCT analyses but should be strongly recommended.

Third, our findings support the notion that fewer than six time points in the preintervention and postintervention periods are probably too few for reliable ITS analyses.

#### 4.1. Strengths and weaknesses of the study

Despite a thorough search for eligible trials, we ended up with a small number of data sets. Nevertheless, we believe we have identified some key issues that are important to consider when impact evaluations of health system interventions are conducted and interpreted. Although ITS

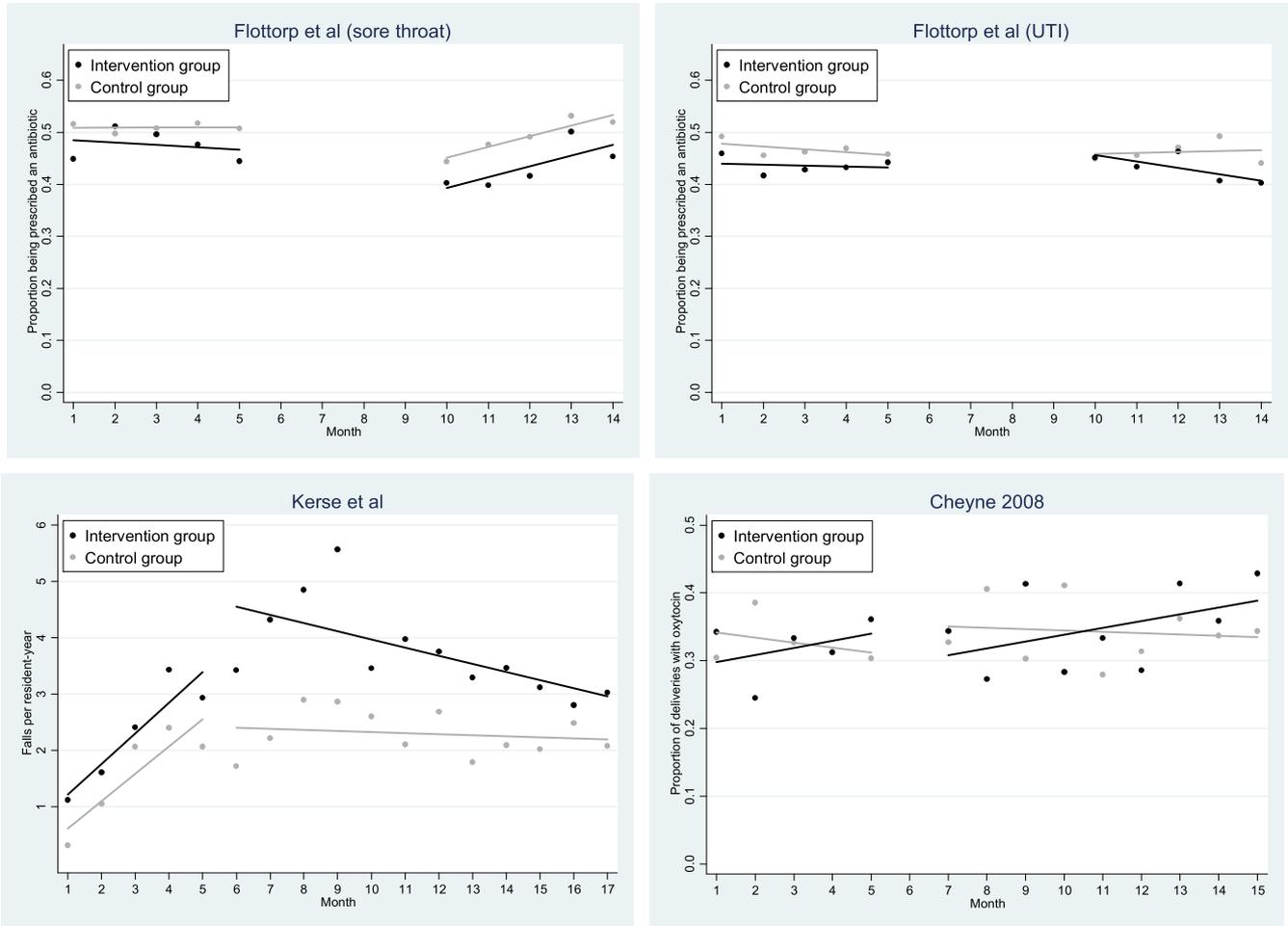


Fig. 4. Linear regression of preintervention and postintervention time series, from studies with five time points per period.

analyses fared well in our comparisons with C-RCTs, this was not always the case and these findings should be considered suggestive until more data are available. Unfortunately, it may not be possible to predict in advance when a control group is necessary to avoid misleading findings from a single-arm ITS study, as was the case for one of the trials included in this study.

Our extensive effort to search for and retrieve trials for our study means that we probably have identified most of the relevant and accessible studies. To expand the number of trials, we could have included a broader range of interventions, including studies from other fields than health systems and health policy research. For example, trials conducted in the fields of economics, education, social welfare, and development might have provided relevant data for the exploration of the relative merits of the various study designs. Future work may determine if our results are replicated in these fields.

Our ITS analyses may not represent a fair comparison with the C-RCTs, for two reasons. First, our ITSs contain fewer time points than typically recommended for these types of studies, and results may be less stable than ITS analyses with long baseline series. Second, we used linear regression

modeling in all our ITS analyses, but in some cases, other models (eg, quadratic curves) may fit the data better.

#### 4.2. Implications

Our findings support the position that concurrent control groups are important when trying to evaluate the effectiveness of health system interventions; failure to use control groups can sometimes lead to erroneous inference about intervention effects. On the other hand, the single-arm ITS design, where the preintervention period serves as control, yielded findings that were mostly consistent with controlled analyses. Of note, if data from RCTs are analyzed without taking into account trends over time, we have shown that the findings may also sometimes be misleading. Thus, those who commission or conduct impact evaluations of health system interventions should routinely use graphical displays of longitudinal data and time-series analysis methods in evaluating intervention effects whether randomization is feasible or not.

The objective of our study was to compare different evaluation methods—not to reassess or replicate previously published effect estimates. In general, our C-RCT estimates

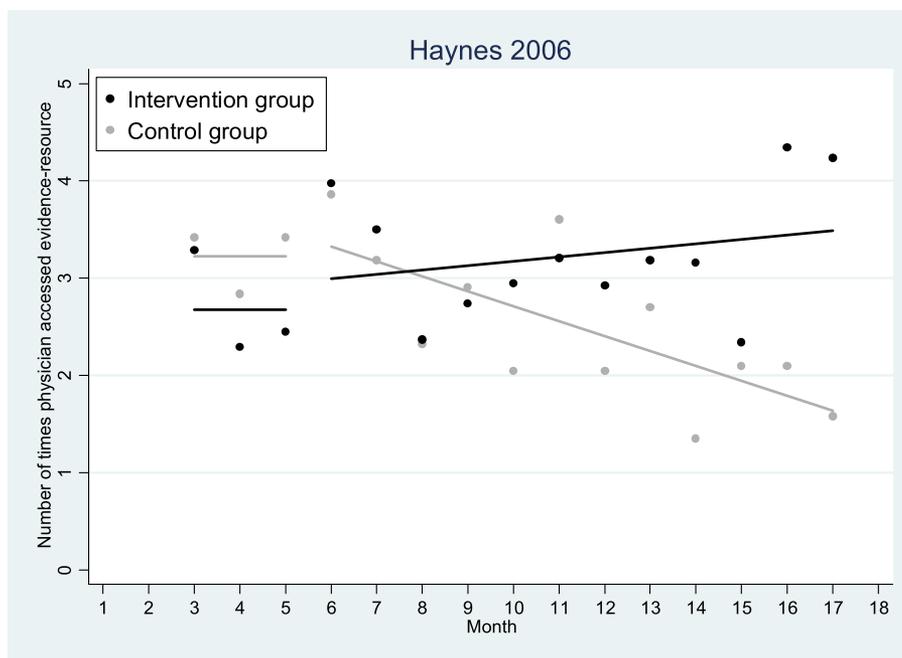


Fig. 5. Linear regression of preintervention and postintervention time series, from studies with three data points before the intervention.

are in agreement with the results from the original C-RCT publications. One exception is the Haynes et al. trial where we, contrary to the original authors, found no convincing effect. For our analyses, we discarded a large part of the original data set to enable an ITS analysis, which probably explains the discrepancy.

## 5. Conclusion

ITS design is a valuable approach in the evaluation of health systems interventions, both when RCTs are not feasible and in the analysis and interpretation of data from randomized trials.

## Acknowledgments

The authors thank Signe Flottorp, MD PhD, and Brian Haynes, MD PhD, with the McMaster PLUS team for accommodating our use of their trial data.

## Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.jclinepi.2014.10.003>.

## References

- [1] Barber S. Health system strengthening interventions: making the case for impact evaluation. Geneva: The Alliance for Health Policy and Systems Research; 2007.
- [2] Fretheim A, Oxman AD, Lavis JN, Lewin S. SUPPORT tools for evidence-informed policymaking in health 18: planning monitoring and evaluation of policies. *Health Res Policy Syst* 2009;7(Suppl 1): S18.
- [3] Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. *Int J Technol Assess Health Care* 2003;19(4):613–23.
- [4] Campbell DTR, Ross HL. The Connecticut Crackdown on Speeding: time-series data in quasi-experimental analysis. *Law Soc Rev* 1968; 3(1):33–54.
- [5] Simonton DK. Cross-sectional time-series experiments: some suggested statistical analyses. *Psychol Bull* 1977;84(3):489–502.
- [6] Shadish W, Cook T, Campbell D. Quasi-experiments: interrupted time-series designs. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company; 2002:171–206.
- [7] Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches the GRADE Working Group. *BMC Health Serv Res* 2004;4(1):38.
- [8] Effective Practice and Organisation of Care (EPOC). What study designs should be included in an EPOC review and what should they be called? EPOC resources for review authors. Oslo: Norwegian Knowledge Centre for the Health Services; 2013.
- [9] Reeves B, Deeks J, Higgins J, Wells G. Including non-randomized studies. In: Higgins J, Green S, editors. *Cochrane handbook for systematic reviews of interventions*. Chichester: John Wiley & Sons; 2008.
- [10] Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;317:1185–90.
- [11] Odgaard-Jensen J, Vist G, Timmer A, Kunz R, Akl E, Schünemann H, et al. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Syst Rev* 2011. Issue 4. Art. No.: MR000012. <http://dx.doi.org/10.1002/14651858.MR000012.pub3>
- [12] Oliver S, Bagnall AM, Thomas J, Shepherd J, Sowden A, White I, et al. Randomised controlled trials for policy interventions: a review of reviews and meta-regression. *Health Technol Assess* 2010;14:1–165. iii.

- [13] 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;66:883–7.
- [14] Oxman AD, Chalmers I. Fair tests of health-care policies and treatments: a request for help from readers. *Bull World Health Organ* 2009;87(6):407.
- [15] 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.
- [16] Gillings D, Makuc D, Siegel E. Analysis of interrupted time series mortality trends: an example to evaluate regionalized perinatal care. *Am J Public Health* 1981;71:38–46.
- [17] Huitema B. Simple interrupted time-series designs. The analysis of covariance and alternative: statistical methods for experiments, quasi-experiments, and single-case studies. 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc; 2011:367–402.
- [18] Ukoumunne O, Thompson S. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. *Stat Med* 2001; 20:417–43.
- [19] Willlliams R. Using the margins command to estimate and interpret adjusted predictions and marginal effects. *STATA J* 2012;12(2):308–31.
- [20] Wooldridge JM. *Econometric analysis of cross section and panel data*. 2nd ed. Cambridge, MA: MIT Press; 2010.
- [21] Cheyne H, Hundley V, Dowding D, Bland JM, McNamee P, Greer I, et al. Effects of algorithm for diagnosis of active labour: cluster randomised trial. *BMJ* 2008;337:a2396.
- [22] Flottorp S, Oxman AD, Havelrud K, Treweek S, Herrin J. Cluster randomised controlled trial of tailored interventions to improve the management of urinary tract infections in women and sore throat. *BMJ* 2002;325:367.
- [23] Foy R, Eccles MP, Hrisos S, Hawthorne G, Steen N, Gibb I, et al. A cluster randomised trial of educational messages to improve the primary care of diabetes. *Implement Sci* 2011;6:129.
- [24] Goodacre S, Cross E, Lewis C, Nicholl J, Capewell S. Effectiveness and safety of chest pain assessment to prevent emergency admissions: ESCAPE cluster randomised trial. *BMJ* 2007; 335:659.
- [25] Haynes RB, Holland J, Cotoi C, McKinlay RJ, Wilczynski NL, Walters LA, et al. McMaster PLUS: a cluster randomized clinical trial of an intervention to accelerate clinical use of evidence-based information from digital libraries. *J Am Med Inform Assoc* 2006;13: 593–600.
- [26] Kerse N, Butler M, Robinson E, Todd M. Fall prevention in residential care: a cluster, randomized, controlled trial. *J Am Geriatr Soc* 2004;52(4):524–31.
- [27] Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, McLoughlin KS, et al. Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial. *J Am Med Inform Assoc* 2012;19:555–61.
- [28] Fretheim A, Oxman AD, Havelrud K, Treweek S, Kristoffersen DT, Bjorndal A. Rational prescribing in primary care (RaPP): a cluster randomized trial of a tailored intervention. *PLoS Med* 2006; 3(6):e134.