Section Editor: **Thomas R. Vetter**

STATISTICAL GRAND ROUNDS

Segmented Regression and Difference-in-Difference Methods: Assessing the Impact of Systemic Changes in Health Care

Edward J. Mascha, PhD,*† and Daniel I. Sessler, MD†

Perioperative investigators and professionals increasingly seek to evaluate whether implementing systematic practice changes improves outcomes compared to a previous routine. Cluster randomized trials are the optimal design to assess a systematic practice change but are often impractical; investigators, therefore, often select a before–after design. In this Statistical Grand Rounds, we first discuss biases inherent in a before–after design, including confounding due to periods being completely separated by time, regression to the mean, the Hawthorne effect, and others. Many of these biases can be at least partially addressed by using appropriate designs and analyses, which we discuss. Our focus is on segmented regression of an interrupted time series, which does not require a concurrent control group; we also present alternative designs including difference-in-difference, stepped wedge, and cluster randomization. Conducting segmented regression well requires a sufficient number of time points within each period, along with a robust set of potentially confounding variables. This method compares preintervention and postintervention changes over time, divergences in the outcome when an intervention begins, and trends observed with the intervention compared to trends projected without it. Difference-in-difference methods add a concurrent control, enabling yet stronger inference. When done well, the discussed methods permit robust inference on the effect of an intervention, albeit still requiring assumptions and having limitations. Methods are demonstrated using an interrupted time series study in which anesthesiologists took responsibility for an adult medical emergency team from internal medicine physicians in an attempt to improve outcomes. (Anesth Analg 2019;129:618–33)

Large randomized controlled trials are generally considered the highest level of clinical evidence.[1,2] But individual subject randomization is impossible when entire health facilities or systems implement systemic changes—although such changes may be among the most consequential. For example, it is an enormous, expensive, and complicated process to switch from manual to electronic health records. Such changes are therefore usually simultaneously implemented throughout a facility. Other examples of changes that are typically implemented system-wide include enhanced care pathways, continuous ward monitoring, and alterations in federal reimbursement policies.

The analog of individual subject randomization that applies to system changes is cluster randomization, in which health care units, such as an entire hospital, are randomized to receive an intervention or not.[3,4] However, these sorts of trials are expensive and difficult to organize—and thus remain rare. Far more commonly, investigators seek to evaluate the effects of a system change within a single facility. Typically, they use a before–after approach in which summative results during a period before intervention are compared to those thereafter.[5–8]

The difficulty, though, is that simple comparisons of results before and after a systematic intervention are subject to major biases—all of which can make the intervention appear better than it actually is. For such studies, the 3 major types of bias are time-dependent confounding (especially, improvements over time), regression to the mean,[9,10] and the Hawthorne effect, which is a type of observation bias.[11,12] These and other biases intrinsic to before–after studies are discussed below, and in more detail by Ho et al.[13]

Although conventional before–after analyses are weak, there are relatively robust design and statistical approaches that can minimize major biases. Our goal is to elaborate various approaches for assessing systematic changes.[9–12] Salient features as well as advantages and disadvantages of each design are given, and designs are discussed from weakest to strongest. We begin by detailing the major threats to internal validity in before–after designs, and then discuss: the simple "after-only" design; designs without concurrent control groups including before–after (short time frame) and interrupted time series (long time frame); the same designs with concurrent controls (ie, difference-in-difference methods); stepped-wedge designs; and finally, cluster randomized trials.

We give particular attention to the interrupted time series design in which there is a substantial and stable period of repeated measurements before and after intervention so that the data in each period can be considered as part of a time series (ie, data measured over time from the same

From the Departments of *Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio; and †Department of Outcomes Research, Anesthesiology Institute, Cleveland Clinic, Cleveland, Ohio.

Accepted for publication February 25, 2019.

Funding: None.

The authors declare no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (www.anesthesia-analgesia.org).

Reprints will not be available from the authors.

Address correspondence to Edward J. Mascha, PhD, Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, 9500 Euclid Ave, P77-007, Cleveland, OH 44195. Address e-mail to maschae@ccf.org.

DOI: 10.1213/ANE.0000000000004153

entity, such as a hospital). In such designs, the changes in the time series starting at or after the new policy "interruption" can often be studied using segmented regression,[14] which effectively compares the slopes over time between preintervention and postintervention periods as well as assessing whether there was an outcome discontinuity ("jump") when the intervention started. It is also possible to test for deviations between the observed results after intervention compared to predicted "counterfactual" results expected had the preintervention trend continued. For example, after finding that the periods differ on slope, it may be of interest to investigate whether the point postintervention mean at a prespecified time point was actually different from what would have happened without the intervention.[14] We apply segmented regression methods to an observational study of the impact of an anesthesia-led rapid response team compared to the previous medicine-led team on rapid response usage in a large hospital.

KEY THREATS TO INTERNAL VALIDITY IN A BEFORE–AFTER DESIGN

We begin with a description of the key threats to internal validity in a before–after or interrupted time series design. Most of these threats can be at least partially controlled using the more advanced designs and statistical techniques we discuss in subsequent sections.

Confounding Bias

A confounding variable is one that is related to both the exposure of interest and outcome. Appropriate statistical analyses can adjust for potentially confounding factors that are known and accurately recorded. The difficulty, of course, is that some important confounding factors may not be known, or may be known but either not recorded or not recorded accurately.

Because before–after study periods are usually completely separated in time, a main challenge is trying to separate the effect of the intervention on outcome from the effect of time on outcome. By far, the most concerning potential confounding factor in studies comparing periods before and after a systematic intervention is that health care generally improves over time, so outcomes during later periods tend to be better than those during earlier periods. One can thus not assume that the exposure of interest "caused" the improvement. In fact, much of the improvement may be from subtle enhancements across many aspects of care, most of which are difficult or impossible to quantify. Or they may have resulted from an important practice change other than the one of interest. Many other factors related to the outcome, observed and unobserved, may change over time. Attributing all improvement to the single intervention of interest—ignoring other potentially important interventions—likely overestimates the effect of the intervention of interest. As we will explain, an appropriate method to assess the effect of an intervention in an appropriately designed before–after study will be to compare the trends over time between periods instead of comparing the means, and to do so while adjusting for as many potential confounding variables as possible.

Maturation Bias. A particular type of time-dependent confounding pertinent to perioperative medicine studies is maturation bias.[15] Providers may grow in general experience and expertise over the course of a study, and this improvement might result in improved patient outcomes, independent of the particular intervention of interest, and perhaps differentially between the time periods.

Instrumentation/Measurement Changes Between Periods

Instrumentation or measurement changes related to the primary outcome between the preintervention and postintervention periods can bias comparison between the periods.

As discussed below, it is possible to partially or largely adjust for time-dependent confounding by comparing rates of change before and after an intervention, rather than just comparing means or incidences between periods. But when comparing periods on the rate of change over time, it remains crucial to adjust for patient, provider, and institutional factors that might confound the analysis—as with any observational study.

Regression to the Mean

Systematic interventions are usually designed to improve quality and, therefore, often instigated in response to perceived problems. For example, antibiotic use might be restricted after an outbreak of *Clostridium difficile*. But the incidence of *C difficile* infections varies randomly within hospitals, usually for no apparent reason. A high-incidence period is thus naturally followed by a low-incidence period because the incidence returns to its long-term average. This process is called regression to the mean,[9,10] and it occurs independently of any intervention(s) because it is consequent to random variation. Even a completely useless intervention, which might nonetheless be expensive, will appear to have improved care when the outcome of interest reverts from the random high starting value to the mean value. The best protection against regression to the mean is to evaluate periods that are long compared to the natural fluctuations for particular outcomes.

Placebo and Hawthorne Effects

Considerable experience shows that outcomes of interest improve when people are simply informed that a topic is under study. This is known as the Hawthorne effect or observation bias. Improvements probably result from a myriad of small factors that improve under observation. Any major systematic intervention is automatically under observation just after being implemented. The positive response to observation will appear to improve outcomes even if the intervention itself is of no value. The difficulty is that the outcome improvement is likely to be (falsely) attributed to the intervention because it is all but impossible to quantify the Hawthorne effect or to adjust for it, especially in a before–after study.

Using long study periods (eg, many months or years) helps as the observation effect wanes with time. But even after adjusting for confounding bias and accounting for preintervention and postintervention temporal trends, placebo and Hawthorne biases may remain because there is rarely a “placebo” intervention in the preintervention period as there would be in a parallel-group randomized controlled trial.[11] A better approach to deal with these biases is to include a concurrent control group, when feasible, and to conduct a difference-in-difference analysis (see section: Before–After and Interrupted Time Series Designs: With a Concurrent Control).

Historical Events

When monitoring an outcome over time, it is natural that known historical events, including other interventions, may occur in the postintervention period to enhance the response, independent of the intervention of interest. Or events may occur in the preintervention period to make the outcomes there appear worse than would otherwise be expected. Obvious blips in an outcome over time with a known causes could theoretically be statistically adjusted for in either period using segmented regression, but the details are challenging.

AFTER-ONLY DESIGN: NO PREINTERVENTION DATA

The weakest design for attempting to assess the benefit of an intervention is an “after-only” study, that is, when there are no preintervention data and also no concurrent group without the intervention. This approach is essentially a case series, even if the series is large. Consequently, no “effect” or even association can be assessed at all. The study consists of data measured over a short or long period of time when all individuals receive the intervention or policy. In such a design, because there is no before intervention period and no control group without the intervention, researchers cannot make any inference about effectiveness of the intervention or even association between intervention and outcome.

Researchers can observe and summarize the trend over time in the intervention period, but it would be incorrect to conclude that a positive or negative slope is due to the intervention. Such studies can measure change over time but should take great caution to not equate “change” during that period with “effect of intervention,” no matter how convincing the results might appear.

BEFORE–AFTER AND INTERRUPTED TIME SERIES DESIGNS: WITHOUT A CONCURRENT CONTROL

The most commonly used approach to assess the benefit of a new policy/procedure outside of the traditional randomized trial or cross-sectional observational study is to compare outcomes after start of intervention with those before the intervention. In such designs, there is no concurrent control group, that is, no group that is similar to the preintervention period and that continued observation into the intervention time frame but without receiving the intervention. We will discuss 2 variations on this design: studies with short observation periods, and those with extended observation periods, which allow segmented regression analysis.

Short Before–After Periods

Sometimes there is only a very short period of observation before and even after start of a new policy, such as 1 month.[5–7] If there is not a concurrent control group, this design is especially weak because observed differences might be due to confounding variables, trends that were changing over time independent of the intervention, or regression to the mean. Confounding of the intervention effect in this setting can arise when there are variables (observed or unobserved) that are associated with both the time period and the outcome and are not adjusted for. Analysis would need to adjust for as many potentially confounding variables as possible, including characteristics of the patient, providers, and institution as appropriate.

Even after thorough adjustment for confounding, the potential for bias remains high. For example, differences between the time periods might be due to the Hawthorne effect, in which individuals modify aspects of their behavior in response to their awareness of being observed and a communal desire for particular outcomes. Improvement related to observation may thus stop or even reverse when monitoring stops, even though the intervention continues.

When an intervention is implemented at a time when outcomes were by chance poor, regression to the mean might occur in the new period, making the intervention look better than it really is. In addition, measurement methods, testing methods, or implementation might change between the time periods. Short observation periods reduce the potential contribution of time-dependent confounding, but increase the risk of regression to the mean. Short study periods also limit the number of patients and outcome events, which can compromise full adjustment for known confounding factors. For all these reasons, short observation periods weaken the analysis and increase the chance of various biases leading to incorrect conclusions.

Concurrent control groups would help here because one could directly compare the observed change from preintervention to postintervention with the change during the same time period (or an analogous one) in a setting in which the intervention was not implemented. (We discuss such “difference-in-difference” designs in the section: Before–After and Interrupted Time Series Designs: With a Concurrent Control.)

Extended Before and After Periods: Segmented Regression Analysis for Interrupted Time Series

Our main focus is the common situation in which researchers have data for a substantial period both before and after the intervention of interest, but where there is no concurrent untreated control group. These sorts of data are best analyzed as either an interrupted time series such as segmented regression,[14] change-point analysis,[16] or control charts.[17]

Comparing Means Is Not Valid. In a before–after design, independent of the length of time in each period, it is not appropriate to simply compare time periods on the mean of outcome (mean, proportion, or count) collapsed over time because doing so may mask important trends in either or both periods. Most important, there may be a trend in the preintervention period showing that the outcome is improving over time, which might well continue into the postintervention period. Comparing time periods on the means might then show an apparent change between the periods, but little or none of it may be caused by the intervention. Or suppose the preintervention trend has no change over time for a serious complication, while the postintervention trend shows either a steady decrease or an immediate drop followed by no further change (flat slope). Only showing the means for each period would not sufficiently describe the differences between these intervention effects. Figure 1 displays some of these possibilities, demonstrating why simply comparing means between periods is generally invalid for either a simple before–after or a longer interrupted time series design. Instead, trends over time need to be modeled and compared between periods.

Segmented Regression. We focus on segmented regression because it allows investigators to account for preintervention trends and confounding variables. In particular, it allows investigators to ask important questions including: Did the new intervention change outcomes? Was the change abrupt or gradual? When did the change begin? How much change was there? Using segmented regression, we can compare the slopes over time and compare intercepts (or mean of the outcome) from the end of preintervention with the beginning of postintervention periods. We can also test for deviations between the observed data after intervention compared to values predicted as if the preintervention trend had instead continued.[18]

Requirements for a Segmented Regression. In the design phase of a research project to assess the effect of a new health care policy or practice, investigators should carefully evaluate whether or not a segmented regression approach will be feasible. A segmented regression analysis requires a sufficient number of time points (preferably ≥10) both before and after the intervention and ideally includes at least a full calendar year in each period to reduce the risk of results being driven by seasonal trends. Sufficient data are also needed at each time point, which may limit the number of periods that can be considered.[14]

There should also be clear and visual evidence that the time series in the preintervention period was stable before the intervention—either without change (0 slope) or a fairly consistent positive or negative slope. If the change over time is clearly nonlinear in either period, a straightforward segmented regression approach that fits linear trends in each period would probably not be appropriate. However, there are other options for nonlinear data for which details are beyond the scope of this work. For example, threshold regression or change-point analysis can be used in nonlinear relationships to identify one or more values of the predictor associated with a change in the slope.[16,19,20] These approaches might be useful (eg, for determining the point at which an intervention became effective [or lost efficacy]) if the postintervention relationship with outcome is nonlinear. In a variation on segmented regression, a linear preintervention trend could be compared to various time points postintervention, even if the latter period were nonlinear by comparing predicted values under the preintervention trend with what was observed later.

Data for the model can either be aggregated (eg, using the mean of the outcome at each month [“analysis #1” below]) or on the subject/patient level (“analysis #2” below). In general, though, it seems preferable to analyze subject-level data because including more detail and variability in the

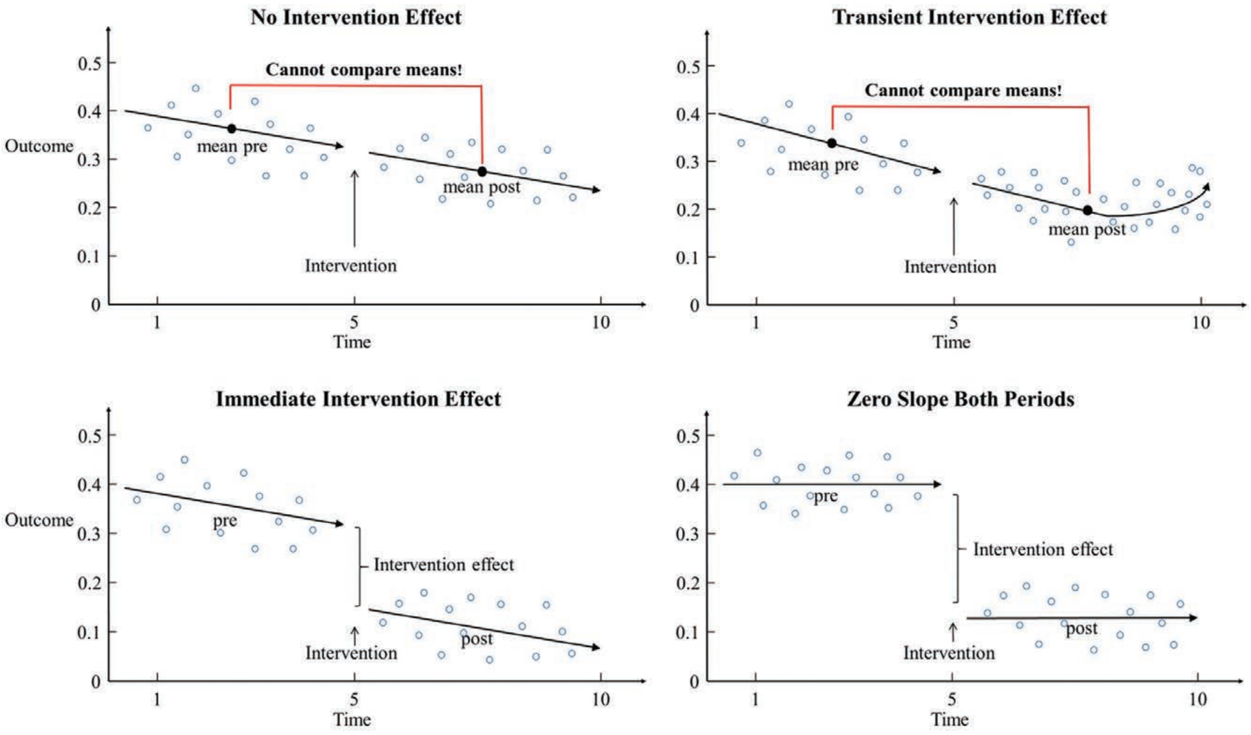

Figure 1. Possible preintervention and postintervention observed in a before–after study with sufficient time series data. The scenarios demonstrate why simply comparing means between periods is not generally valid for either a simple before–after or a longer interrupted time series design. Instead, trends over time need to be modeled and compared between periods.

outcomes and predictor variables enhances accuracy, and often power. However, aggregating or binning of data may help remove some of the effects of autocorrelation,[21] assist in visualizing the trends over time, and is the only option when individual data are unavailable (see Discussion).

Fitting a Segmented Regression Model. A basic segmented linear regression model using aggregated data for a continuous outcome can be specified as follows:

$$\begin{aligned} Y_t = \ & \beta_0 + \beta_1 \text{ time}_t + \beta_2 \text{ intervention}_t \ + \\ & \beta_3 \text{ time\_post}_t \ + ...+ \beta_4 \mathrm{M}_{t4} \text{ --- } \beta_p \mathrm{M}_{tp} \ + \varepsilon_t \end{aligned} \qquad (1)$$

where Y_t is the mean of the outcome for time t,

"time" is the value of time (from 1 to k) from start of preintervention to end of postintervention, and equals t,

"intervention" = 1 for received and 0 if not received at time t,

"time_post_intervention" equals 0 if preintervention and otherwise equals number of time periods from start of intervention to current time period

[ie, $\text{intervention } (1{,}0) \times (\text{time} - \text{intervention start time} + 1)$],

M_{t4} --- M_{tp} are the means of the (p − 3) covariates at time period t,

p is the total number of variables in the model, and

β_0 is the baseline level of the outcome at time 0

β_1 is the slope in the preintervention period

β_2 is the change in the mean of the outcome just after start of intervention

β_3 is the difference in slopes between post- and preintervention periods (ie, post − pre)

Using Equation 1, we estimate the slope in the preintervention period, β_1, and the difference in slope between the preintervention and postintervention periods (post − pre), β_3. A nonzero value of β_3 indicates a change in the temporal trends between the periods. For example, a stable zero-slope trend on the outcome in the preintervention period might have become either a positive or negative slope in the postintervention period, or a negative slope might have changed to zero slope or positive slope. Although not directly estimated in this model, the slope for postintervention is $\beta_1 + \beta_3$. Model 1 also estimates β_2, the change in the mean of the outcome from end of preintervention to start of intervention. An analogous model to 1 can easily be constructed if data are analyzed on the subject level instead of aggregated, where outcome Y_i would represent the outcome measured for a subject i at a particular time t. (Note: we do not imply that the same subject is measured over time [although that can be accounted for by incorporating a random effect for subject, for example].)

Hypothesis Testing. As in any study, hypothesis tests of interest using segmented regression should be specified in the design phase, or at least before analysis of the data. For example, investigators might decide that they will claim an intervention successful if there is either an immediate change in the mean, rejecting H0: $\beta_2 = 0$, or a change (in the "benefit" direction) in the slope, rejecting H0: $\beta_3 = 0$. In such a case, because there are 2 outcomes and either being significant would qualify the intervention as successful, a correction to protect type I error should be made. However, the above design might result in an improved mean at the start of intervention, but a worsening slope over time, for example. Therefore, a more robust design might include inference on both parameters at the same time in a "joint hypothesis test." For example, investigators might require an improved intercept ($\beta_2 \neq 0$) and/or improved slope ($\beta_3 \neq 0$), but also require that neither was "worse," or in the wrong direction. Thus, authors could first require noninferiority on both parameters and then superiority on at least one.[22]

Segmented Regression in Generalized Linear Model Framework. More generally, the most appropriate form of a segmented regression model for a specific application would be chosen based on the distribution of the outcome variable and model fit. As with standard regression modeling, a generalized linear model framework[23] assuming the exponential family can be considered, with the analyst choosing the appropriate distribution for the outcome variable and the corresponding "link function" (ie, relationship between the right side of model and the mean of the outcome on left) for the application at hand.

For example, with continuous and roughly normally distributed data (particularly, normally distributed residuals), a linear regression model would often be appropriate, that is, using the identity link function as in Equation 1. For binary data, a logistic regression model would be appropriate (ie, using a logit link), either for the aggregated data using the # events/# trials setup or for subject-level data using a traditional logistic regression model. Alternatively, aggregated binary data could be analyzed as a count assuming either a Poisson or negative binomial distribution with log link while adjusting for the total sample size for a given interval as an offset. Whatever model is chosen, it is incumbent on the analyst to assess model assumptions, as with any other analysis.

Technical Note on Segmented Versus Traditional Regression. A segmented regression differs from the traditional model with factors for group, time, and group × time interaction, particularly in the meaning and values for the β_2 parameter corresponding to the intervention effect. In a traditional (linear) model in which the interaction parameter is simply the product of a time variable (eg, from 1 to K, with K total time intervals) and group variable coded 1 and 0, β_2 is the mean of the intervention group at time = 0. In contrast, for segmented regression, β_2 is the difference between the predicted mean at the start of the intervention compared to at the end of preintervention. Otherwise, the parameters for intercept, time, and group × time have the same values and meanings as a traditional model.

Outcome variables in our application data below are binary, as is the case with many health care interventions. We therefore demonstrate a segmented regression model using logistic regression for a binary outcome on aggregated data, such that the outcome is expressed for a given time interval (say week, month, or quarter) as # events/# trials, equal to the proportion of subjects having the event in the given time interval. This formulation appropriately accounts for the potential variation in the denominator and the differing variances association with each estimated proportion over time because the variance of a proportion is a function of that proportion. Such models are better than trying to model aggregated binary outcome data using linear regression because the latter can generate nonsensical values such as proportions <0 or >1. A segmented regression model for aggregated binary data can be expressed as follows:

$$\text{logit}\ (P_t = \#\mathit{events} / \#\mathit{trials}) = \beta_0 + \beta_1\ \text{time}_t \\ + \beta_2\ \text{intervention}_t + \beta_3\ \text{time\_post}_t +\ \beta_4 \text{M}_{t4} \text{---}\ \beta_p \text{M}_{tp} \qquad (2)$$

where P_t is # events/# trials (proportion with the event) for time interval t, values for time, intervention and time_post are specified as in 1, and $\mathit{logit}\ (P_t) = \mathit{log}\big(P_t / [1 - P_t]\big)$. Parameters of interest have an analogous interpretation to 1 but in terms of log odds:

β_0 is logit of outcome at time 0 [back-transform to estimate probability of outcome]

β_1 is change in log-odds of outcome per increment in time interval in preintervention period; exponentiated β_1 is change in odds of outcome (ie, odds ratio) per increment in time period

β_2 is change in log-odds of outcome just after the start of intervention; exponentiated β_2 is change in odds of outcome (ie, odds ratio) just after the start of intervention

β_3 is difference in log-odds of outcome per unit time between post- and preintervention periods

For example, a value of 0.123 for β_1 would mean that the log odds of the outcome increases 0.123 for a difference of 1 in the time variable (eg, 1 month) in the preintervention period; exponentiating β_1 gives an odds ratio of $e^{0.123} = 1.13$, meaning that the odds of outcome is 13% higher for an increase of 1 time unit. A value of 0.4 for β_2, exponentiated to 1.49, would mean that the odds of the outcome increased about 50% from the end of preintervention to just after the start of intervention.

For analysis on the subject/patient level (analysis #2), differences from Equation 2 are that the outcome *Yi* is not a summary for a certain time interval, but rather the outcome value for the *i*th individual at a particular time; values for "time" would correspond to the relevant date for an individual's outcome measurement; and the covariates M are not means for an interval but rather patient or system/provider/seasonal values linked to the individual. An example is given in the applications section.

When conducting a segmented regression, researchers may find that some of the main parameters of interest (eg, β_1, β_2, or β_3 in Equation 1 or 2) are not significant, suggesting that the model could be simplified. For example, if the slopes do not differ between periods (ie, test of H0: $\beta_3 = 0$ in Equation 1 nonsignificant), one might fit a reduced model that only includes time, the intervention, and indicated confounders. But while such parsimony might seem intuitive from a general modeling perspective, it can be problematic in segmented regression. For example, the intervention effect in a model with only time and intervention effects would be interpreted as the difference between interventions at any point in time, even though it would not make sense to compare them at any point in time because the periods are completely separated by time. If both the time (β_1) and time after intervention (β_3) variables were nonsignificant, a model with only the intervention and confounding variables might be tempting. But it is nonetheless nearly always preferable to use a full model, which provides a more precise estimate of the effects of interest.

Predictions: With Versus Without Intervention. From the model in 2 (or similarly 1), we can also test whether the trend modeled with the observed data in the intervention period differs from what would be expected if the preintervention trend continued unabated. Such testing should be specified in the design phase of the study. For example, investigators might plan to test at 12 and 24 months after the start of intervention whether or not the resulting mean of the outcome, as predicted by the postintervention regression line, differs from the projection of the preintervention trend. Given a significant change in the slope, for example, it might be important to test whether there is evidence that means at certain time points are different from what was expected without the intervention. In fact a difference in slopes with no change in intercept implies that the true means with versus without intervention would in fact differ at postintervention times. However, the below testing is important because researchers may want to know at what point the differences are detected, taking into account the observed variability in the data and postintervention regression line as well as the uncertainty in the counterfactual projection.

Equations 3 and 4 below show the predicted values of the outcome at a particular time point (called "predict t") after the start of the intervention, with and without consideration of the intervention data, respectively. We use the ^ (or "hat") on top of model parameters here and throughout the manuscript to indicate estimates of the true population parameters, for example, when using the derived model for prediction.

$$\text{logit}(\hat{P}_{\text{predict}\ t\ \text{(with)}}) = \hat{\beta}_0 +\ \hat{\beta}_1\ \text{predict}\ t\ + \\ \hat{\beta}_2\ 1 + \hat{\beta}_3\ (\text{predict}\ t\ \text{-}\ \text{start}\ t) + \hat{\beta}_4\ \text{M}_{t4} \text{---}\ \hat{\beta}_p\ \text{M}_{tp} \qquad (3)$$

$$\text{logit}(\hat{P}_{\text{predict}\ t\ \text{(without)}}) = \hat{\beta}_0 + \hat{\beta}_1\ \text{predict}\ t\ + \\ \hat{\beta}_4\ \text{M}_{t4} \text{---}\ \hat{\beta}_p\ \text{M}_{tp} \qquad (4)$$

Then the difference between the predictions in Equations 2 and 3 is as follows:

$$\hat{\beta}_2 + \hat{\beta}_3\ (\text{predict}\ t\ \text{-}\ \text{start}\ t), \qquad (5)$$

where "Start t" is the time indicating end of the preintervention period, "Predict t" is the time point of interest, and where Equation 5 is an estimate of the absolute difference in the predicted logit of the outcome (assuming a logistic regression model) at the prediction time of interest, with versus without considering the effect of the intervention.

The standard error for the difference in predictions in Equation 5 is calculated using the formula for the variance of a difference, then taking the square root, and estimated by the following:

$$\widehat{\text{SE}}_{(5)\,=\,(3)\,-\,(4)} = \sqrt{\text{var}(\hat{\beta}_2) + (\text{predict}\ t\ \text{-}\ \text{start}\ t)^2\ \text{var}(\hat{\beta}_3)} \qquad (6)$$

We can then test whether the difference in predictions at the given time point with and without considering the intervention differs from 0 using a z test and comparing z

= difference/standard error = (5)/(6) to a standard normal distribution and rejecting if $P < .05$ in a 2-tailed test, for example.

Accounting for Autocorrelation Over Time. Segmented regression uses data measured over time from the same institution, department, or organization. In such time series, data observations that are set distances apart (eg, 1 week or 1 month) may be correlated with each other. Failing to account for this natural phenomenon in a time series analysis can give incorrect results.

The good news is that much of the autocorrelation in a segmented regression analysis may be removed by appropriate adjustment for confounding variables.[18] Nevertheless, researchers should assess the autocorrelation in the model residuals after adjusting for confounding. This can be done by creating a correlogram, which is a graphical method to assess the degree of autocorrelation for successive data that are various increasing distances apart (eg, lag1, lag2, etc). In a simple analysis, evidence of autocorrelation with the previous observation would typically be indicated by a Durbin–Watson statistic of <2. Sometimes autocorrelation can be removed by simply analyzing the difference between successive data points as opposed to the data points themselves. Still, the actual autocorrelation may be complex and depend on multiple factors that are beyond the scope of this article.[24]

Lagarde[25] gives an example in which there is noticeable first-order autocorrelation (ie, errors in one period correlated with errors in the next consecutive period), with Durbin–Watson statistic of 1.23, and where estimates of the parameters of interest vary considerably depending on whether the adjustment is made or not. A challenge is that adjusting for autocorrelation can lead to a variance that is too small if not done correctly. Basic SAS code (SAS Institute Inc, Cary, NC) for making the adjustment using linear regression can be found at the end of Supplemental Digital Content 1, Document, http://links.lww.com/AA/C788.[26] Different ways of coding the time variables are explained by Ostrom.[27] Analyzing aggregated data may help remove some of the autocorrelation, with the best way to do this varying by application.[21]

BEFORE–AFTER AND INTERRUPTED TIME SERIES DESIGNS: WITH A CONCURRENT CONTROL

Adding a concurrent control group that did not receive the intervention in either period allows researchers to compare their before–after experience with a control. This approach is a substantial improvement over a design with no concurrent control.

Difference-in-Difference Approach

If a before–after design also has a concurrent control group that did not receive the intervention in either period, inference can often be improved by using the "difference-in-difference" analytic approach.[28–34] This approach compares between-period changes in cohorts that received the intervention with changes in similar cohort(s) that did not, over a similar time frame. It can be used with either short or long preintervention and postintervention time periods.

As in any cohort comparison, a key aspect of such a design is choice of an appropriate reference group or groups. The control(s) should be as similar as possible to the setting in which the intervention was implemented, including complete or considerable overlap in time, and ideally differ from the experimental setting only by the intervention. For example, a concurrent control group for assessing the impact of a new policy on cardiac surgery patients in one hospital might be a similar hospital in the same city, or better, another floor in the same department in the same hospital (assuming internal contamination can be avoided).

Simple Difference-in-Difference Method. Analysis for the simple difference-in-difference approach involves comparing the preintevention-to-postintervention change (eg, difference in means) in a unit (eg, hospital or department) in which some subjects were exposed to the intervention in the post period with the preintervention-to-postintervention change during the same time period in a unit that was not exposed. While numerous time points can be used preintervention and postintervention, the simple difference-in-difference method does not compare the trends over time from before and after intervention versus concurrent trends without intervention. Instead, it compares the difference in means before and after intervention with the difference in means for similar time period(s) without the intervention.

A main assumption for this simple difference-in-difference approach is the "parallel trend" assumption, meaning that the trends over time in the groups being compared are expected to be the same if it were not for the effects of the intervention in one of the groups. Therefore, it is important to confirm that preintervention trends were similar in the exposed and control populations. Then, in the difference-in-difference approach, researchers should adjust for as many potentially confounding variables as possible because the treated and concurrent control groups might differ in important ways that change over time. A key assumption is that unmeasured variables are not true confounders; that is, they either differ between groups but do not change over time (ie, are time invariant) or are time-varying factors that do not differ between groups. This implies that a graph of the data plotted over time should resemble a set of parallel lines.[33]

A basic difference-in-difference analysis can be formulated as a direct extension of the simple before–after comparison of means. The basic approach is to fit a regression model of the outcome Y as a function of cohort type (Exp = 1 for experimental cohort and 0 for concurrent control), time period (ie, Intervention = 0 for pre [or "no"] and 1 for post [or "yes"]), cohort-by-time period interaction = Exp × Intervention (the effect of interest), and a set of potentially confounding variables M (modeled here as a vector), as follows:

$$\begin{aligned} Y_{gt} = \beta_0 + \beta_1\ \text{Exp}_g + \beta_2\ \text{Intervention}_t \\ + \beta_3\ (\text{Exp}_g \times \text{Intervention}_t) + \boldsymbol{\beta}\mathbf{M} \end{aligned} \tag{7}$$

where Y_{gt} is, for example, the outcome for a subject in the gth cohort in time period t. For this simple model, subjects are assumed to have been measured once, so data are independent. We assess whether the difference between time periods is consistent for cohorts that did and did not receive the intervention. The parameter of interest is thus β_3, the cohort-by-time period interaction, which estimates the difference between experimental and control cohorts on the difference in means between periods 0 and 1.

The simple 2 × 2 difference-in-difference design in Equation 7 can be generalized to include multiple time periods so that an entity such as a state or hospital could be modeled as being on or off the intervention (Exp = 1 or 0) during any number of periods, as in the following example:

$$\begin{gathered} Y_{igt} = \beta_0 + \beta_1 \text{ Exp}_g + \beta_2 \text{ Period}_t + \beta_3 \\ (\text{Exp}_g \ \times \ \text{Period}_t) + u_i \beta_{4*} + \boldsymbol{\beta}\mathbf{M} \end{gathered} \tag{7.5}$$

Where Y_{igt} is an outcome within the ith unit (eg, hospital or state) in a certain period and either on or off the intervention, and where u_i is a fixed-effect indicator for the ith unit (eg, hospital or state), with separate indicator and β coefficient (*) for each unit. This might be a linear mixed-effects model, which would also consider unit as a random effect or otherwise account for the potential within-unit correlation across multiple periods and exposures.

Models can also be generalized to account for secular trends, to assess further interactions, and to accommodate a continuous intervention. For example, Sun et al[35] used a difference-in-difference approach to assess the effect of "opt-out" regulation on access to surgical care for urgent cases in the United States using the National Inpatients Sample. They modeled the outcome of interest, whether a particular surgery was received by a patient, as a function of whether a particular state was in opt-out status for the given year, a year indicator, state indicator, confounding variables (patient, state, and national), and trends over time within state.

Segmented Regression With Difference-in-Difference. Interrupted time series designs can be strengthened by adding a concurrent control unit or group that did not receive the intervention.[36] That is, segmented regression methodology can be expanded to include units or institutions for which no intervention was implemented during the study time period, but for which all other variables of interest have been collected. The segmented regression pre-to-post differences in slopes and intercepts in settings that did receive intervention can then be compared to the analogous differences in the contemporaneous unexposed cohort(s). This approach is expected to remove more bias due to changes in secular trends compared with segmented regression without a concurrent control, thereby improving the reliability of inferences.

A segmented regression model using difference-in-difference might expand on Equation 1 as follows:

$$\begin{aligned} Y_t = {} & \beta_0 + \beta_1 \text{ time}_t + \beta_2 \text{ intervention}_t + \beta_3 \text{ time\_post}_t \\ & + \beta_4 \text{ Exp}_t + \beta_5 \text{ Exp} \times \text{time}_t + \beta_6 \text{ Exp} \times \text{intervention}_t \ + \\ & \beta_7 \text{ Exp} \times \text{intervention} \times \text{time\_post}_t + \varepsilon_t \end{aligned} \tag{8}$$

where Exp = 1 for the experimental cohort and 0 for the concurrent control, and β_5, β_6, and β_7 compare cohorts (Exp = 1 vs 0) on the segmented regression parameters in model 1. For example, β_5 measures the difference between cohorts on preintervention slope, β_6 compares cohorts on change in intercept at start of intervention, and β_7 compares cohorts on the change in slope between periods. An intervention might be considered effective in this model if either or both parameters β_6 or β_7 were found to differ from 0, following a joint hypothesis setup discussed earlier after Equation 1.

Advanced Difference-in-Difference Methods. Difference-in-difference methods can be expanded to directly adjust for additional factors when estimating the treatment effect, such as known confounding variables, in what is called a difference-in-difference-in-difference method. For example, one can first estimate the difference-in-difference to compare hospitals that did versus did not experience the intervention across similar time periods, and then subtract off the within-hospital difference for a factor that may be confounding the comparison of interest.[33]

Stepped-Wedge Cluster Randomized Trials

A new policy or practice can also be tested in a stepped-wedge cluster randomized design. In this design, clusters such as hospitals, hospital units, or practices all start without the intervention and are randomly started on the intervention at predefined intervals through the study period until all are receiving it. Once an intervention is begun in a cluster, it remains in place for the study duration.[37,38] In the first period, no clusters receive intervention, and by the last period, all receive it, thus creating "steps" when enrollment is graphed over time. In the analyses, adjustment should be made for the intracluster correlation, including repeated measurements on units within a cluster.

A stepped-wedge design is practical when logistical or ethical consideration preclude all clusters starting the intervention at the same time as would happen in a cluster randomized trial, and when individual randomization and intervention concealment are not practical. More steps are better, and each step should be long enough to observe outcomes on patients. Delayed rollout and delayed treatment effects within a step can decrease power. Careful planning is required to avoid this issue altogether, although statistical approaches can be used to adjust for the fraction of the optimal follow-up time that was experienced in a step. There may also be problems with contamination due to long periods between initial recruitment and intervention start. A stepped-wedge design usually requires fewer clusters than a parallel design but more measurements per cluster. The magnitude and variance of treatment effect heterogeneity between clusters should be considered in sample size calculations.

While a before–after design is completely confounded by time, stepped-wedge designs have overlap and are thus only partially confounded by time.[39] Because introduction of the intervention is staggered over time with a higher percentage of clusters under the intervention toward the end, and because many potentially confounding factors also change over time, the analysis must include adjustment for trends over time, usually accomplished by including a variable indicating the timing of each "step." Trends over time within each step may also be assessed, accounted for, and compared to preintervention periods as with an interrupted time series. Thus, modeling can be used to separate time and intervention effects.[39–41]

Traditional Cluster Randomized Trials

Arguably, the strongest approach to evaluation of systematic changes that preclude individual randomization is a cluster randomized trial in which all clusters are randomized in a parallel fashion to either intervention or control.[3,4] This design removes bias associated with changing trends over time, so that all intervention groups experience the same changes, at least within the randomized strata. Sample size for a cluster randomized trial is based largely on the number of clusters and the expected within-cluster correlation; adequately powered trials thus typically require many clusters.[42,43] Cluster randomized trials are challenging to organize and execute; thus, they remain rare. Nonetheless, cluster randomization is the preferred approach for quantifying effects of system changes when they are logistically practical.[44]

Sample Size and Power Considerations

See the section on requirements for a segmented regression for basic data requirements to appropriately fit a segmented regression model. As with any study, power for the primary outcome should be estimated when designing before–after or interrupted time series studies. For segmented regression, authors should specify the change in slope and/or intercept between the preintervention and postintervention periods that they want to be able to detect with sufficient power. Existing monthly summary data might be used, for example, to estimate a slope and standard error in hypothesized preintervention and postintervention periods. When practical, the best approach is to use existing data for the preintervention period that will provide an accurate estimate of the preintervention slope, and then, for example, calculate the difference in slopes (post − pre) that they would have sufficient power to detect under the circumstances.

Sample size calculations for a study using segmented regression to analyze time series data should include total sample size for each period, number of intervals in each period, effect size (difference in slopes and/or intercepts between periods), and the expected autocorrelation over time.[45] In their 2011 article, Zhang et al[45] detail approaches with examples and offer a SAS program on request. Kontopantelis[46] wrote a STATA module for simulation-based power calculations for interrupted time series designs, but it only tests for level changes and not changes in slope. Penfold and Zhang[47] gives an example of the database needed to conduct an interrupted time series, as well as SAS code to implement a difference-in-differences model. For stepped-wedge and cluster randomized trials, Hemming and Taljaard[48] provide excellent guidance on sample size calculations.

For a cluster randomized trial, the total required sample size is a function of the number of subjects per cluster, the number of clusters, and the expected intracluster correlation. The total required sample size can thus be calculated as the product of the sample size required for a traditional parallel-group randomized trial and what is called the "design effect" or variance inflation factor, $1+\ (n\ -\ 1)\rho$, where n is the average cluster size and ρ is the intraclass correlation. In general, lower cluster size with more clusters, along with lower intraclass correlation, will decrease the total sample size.[49,50]

APPLICATION OF SEGMENTED REGRESSION TO THE ADULT MEDICAL EMERGENCY TEAM STUDY

Starting in 2012, a group of clinicians at the Cleveland Clinic restructured the rapid response team to be anesthesiologist led, based on the hypothesis that the change would improve the AMET results by decreasing cardiopulmonary attacks and in-hospital mortality. For our application of these methods, we assess the tertiary hypothesis that the change would increase the number and percent of rapid response team calls among all eligible patients ("rapid response").

We used data from this currently unpublished study to conduct a segmented regression analysis that assessed the relationship between the restructuring of care and the 3 outcome variables, especially rapid response. The preintervention period ran from January 1, 2009, to December 31, 2011 (3 years), and the intervention period from January 1, 2012, to December 31, 2016 (4 years). The analysis included full data on 151,145 cases on 93, 970 hospitalized patients, of whom 44% were in the preintervention period and 56% thereafter. SAS code for the main analyses is given in Supplemental Digital Content 1, Document, http://links.lww.com/AA/C788.

In this application, we stipulate that the intervention will be deemed effective if there is either an improvement in the slope (ie, $\beta_3 \neq 0$ in Equation 2 in the "benefit" direction) and/or a sudden improvement in the outcome (ie, $\beta_2 \neq 0$ in the "benefit" direction) corresponding with the intervention, and that the intervention is at the same time not found to be worse (inferior) on either outcome. Because significance on either of the 2 tests for benefit would allow a conclusion that the intervention was effective, a Bonferroni correction should be applied to the significance criterion (eg, 0.05/2 = 0.025).[22]

Crude Results (Ignoring Time Trends)

We begin with a naïve comparison between the periods, as discussed above in the section Comparing Means Is Invalid. Over the study period, the percentage of all cases in the preintervention and postintervention periods experiencing rapid response was 1.7% vs 3.6% (P < .001). Results were similar using the first record for each patient.

While these results are in the expected direction and statistically significant, we cannot accurately conclude that the difference is due to the intervention because we ignored the preintervention and postintervention trends over time. For example, clear trends over time in the rapid response outcome are evident from Figure 2. We also ignored confounding due to other factors. A more appropriate analysis is to consider 2 distinct segmented regression analyses to compare the periods—first using aggregated monthly data and then using data from individual patients.

Analysis 1: Aggregated Data

A sample of the monthly aggregated data in the Adult Medical Emergency Team study is given in Table 1, and raw (unadjusted) data for the outcome over time for each time period are given in Figure 2. We aggregated the data by month mainly to be able to display the trend and its linearity and stability in the outcome over time. We chose 1-month intervals because that gave sufficient data for each

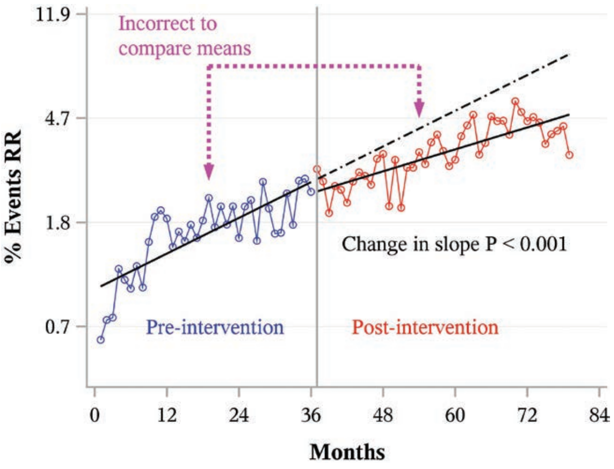

Figure 2. Segmented regression of application study data—unadjusted. Monthly data from our application study in which investigators at the Cleveland Clinic sought to evaluate whether modifying the adult medical emergency team to be anesthesiologist led would increase the percentage of candidate patients for whom an RR call was made compared to the previous period during which the team was led by general physicians. Data are 151,145 records on 93,970 patients. Solid lines through the data are the estimated preintervention and postintervention lines estimated from Equation 2, a logistic segmented regression model on the outcome of proportion of RR encounters for a given month (ie, # encounters/# available patients). The dotted line in the postintervention period is the projected preintervention trend. The represented slopes and intercepts are not adjusted for confounding variables, but the *P* value of <.001 for difference in slopes between the periods is adjusted, indicating a decreased slope in the postintervention period compared to preintervention. RR indicates rapid response.

interval, and also a sufficient number of intervals within each period to allow visualization and modeling of the relationship over time.

We focus our exposition of methods on the rapid response outcome variable—the proportion of available patients in the given month in whom the rapid response team was called. Data were modeled using logistic regression with the outcome defined as # events/# trials for each month, as in Equation 2, where "# events" was the number of rapid response calls, and "# trials" was the number of available patients for a given month. In the final model, we adjusted for a robust list of potential confounding variables (as the mean for a given month) including patient age, body mass index, surgical versus medical status, and 8 comorbidities including history of renal failure, liver disease, solid tumor without metastasis, obesity, unintended weight loss, anemia, drug above, and depression. The model was derived by including as many biologically plausible potential confounding variables as possible without overfitting. The entire list of potential confounding variables is summarized between periods in Supplemental Digital Content 2, Appendix, Table A1, http://links.lww.com/AA/C787.

All Data, Confounder Adjusted. Results at the top of Table 2 indicate that: (1) the preintervention log-odds ratio for the rapid response outcome over time ("months" variable), or slope, was significantly greater than 0 (odds ratio [95% CI] of outcome for a 1-month increase of 1.04 [1.03–1.05], $P <$ .001 testing whether β_1 in Equation 2 equals 0); (2) there was not a statistically significant increase or jump in the odds of the outcome from just before to just after the start

Table 1. Adult Medical Emergency Team Study Monthly Data[a] From Cleveland Clinic (2009–2016)

	Outcome Variable		Time (mo)	Intervention (1 = Anesthesia Led)	Time Postintervention (–36 mo)	Monthly Means for Confounders		
N	No. of Rapid Responses	% Rapid Responses				Age	Body Mass Index	Chronic Heart Failure
1867	11	0.59	1	0	0	56.42	28.72	0.08
1686	12	0.71	2	0	0	55.88	28.63	0.08
1922	14	0.73	3	0	0	56.05	29.50	0.09
1899	22	1.16	4	0	0	56.11	29.18	0.09
1821	19	1.04	5	0	0	56.27	28.60	0.09
1978	19	0.96	6	0	0	56.03	28.67	0.09
2020	24	1.19	7	0	0	56.20	28.74	0.08
1958	19	0.97	8	0	0	56.19	29.08	0.10
1940	29	1.49	9	0	0	56.45	28.98	0.08
2058	39	1.90	10	0	0	55.36	29.20	0.09
…		…	…	…	..	…	…	
1956	58	2.97	37	1	1	56.26	28.99	0.11
1779	47	2.64	38	1	2	56.00	28.94	0.10
1934	38	1.96	39	1	3	55.87	29.50	0.10
1817	46	2.53	40	1	4	56.05	29.65	0.09
1884	46	2.44	41	1	5	55.61	29.16	0.12
1757	38	2.16	42	1	6	55.94	29.19	0.11
1856	49	2.64	43	1	7	55.56	28.71	0.09
1945	56	2.88	44	1	8	55.10	29.26	0.10
1869	52	2.78	45	1	9	54.57	28.64	0.10
…		…	XX…	…	…	…	…	

Y variables and monthly means for confounders represent the mean across all cases for each month of the study. "Intervention" is 1 for the intervention period (anesthesia-led adult medical emergency team, started in month 37) and 0 for preintervention period (general physician led).

[a]Sample of monthly data used in the section Application of Segmented Regression to the Adult Medical Emergency Team Study.

Table 2. Analysis 1: Segmented Regression on Aggregated Monthly Data

Parameter	Estimate	Standard Error	95% CI	Odds Ratio (95% CI)	*P* Value
All data, confounder adjusted: 151,145 records on 93,970 patients					
Intercept, β_0	−0.264	3.09			.93
Time, β_1	0.041	0.006	0.029, 0.053	1.04 (1.03, 1.05)	<.001
Intervention, β_2	−0.158	0.113	−0.379, 0.063	0.85 (0.68, 1.06)	.16
Time_post, β_3	−0.024	0.006	−0.036, −0.011	0.98 (0.96, 0.99)	<.001
First record per patient, confounder adjusted: N = 93,970					
Intercept, β_0	2.40	2.11			.26
Time, β_1	0.044	0.005	0.034, 0.054	1.045 (1.035, 1.055)	<.001
Intervention, β_2	−0.432	0.129	−0.684, −0.179	0.65 (0.51, 0.84)	.001
Time_post, β_3	−0.022	0.006	−0.034, −0.011	0.98 (0.97, 0.99)	<.001
All data, unadjusted: 151,145 records on 93,970 patients					
Intercept, β_0	−4.548	0.068			<.001
Time, β_1	0.025	0.003	0.019, 0.031	1.025 (1.019, 1.031)	<.001
Intervention, β_2	−0.047	0.068	−0.180, 0.087	0.95 (0.84, 1.09)	.49
Time_post, β_3	−0.008	0.003	−0.015, −0.002	0.992 (0.985, 0.998)	.014

Outcome is the proportion of rapid response calls for given month. Logistic regression model with # events/# trials setup assessed change in log odds (slope) and intercept after start of adult medical emergency team anesthesiologist-led intervention. Results: All 3 analyses show an increasing trend before intervention ($\beta_1 \neq 0$), which decreased during the intervention period ($\beta_3 \neq 0$). Only the second analysis showed a significant change (a drop) in the outcome at the start of intervention ($\beta_2 \neq 0$). Only the first analysis, using all data, confounder adjusted, shows significant change (improvement) in slope postintervention. Time: months since start of study (1–36 preintervention and 37–79 postintervention). Intervention: 0 for preintervention, 1 for post. Time_post: equals 0 in preintervention period and starts at 1 for first month in postintervention.

β_0: intercept: covariable-adjusted log odds of outcome at time 0 with intervention = 0.

β_1: preintervention slope (change in log odds of outcome per month). β_2: log-odds ratio of outcome at the start of intervention compared to end of preintervention. β_3: difference between periods in slope of outcome over time (postintervention minus preintervention).

of intervention (odds ratio [95% CI], 0.85 [0.68–1.06], *P* = .16 testing whether $\beta_2 = 0$); and (3) the trend over time (the slope, or log odds of outcome over time) was decreased after the intervention compared to before, with an estimated change (95% CI) in log-odds ratio per month of 0.98 (0.96–0.99), *P* < .001 testing whether $\beta_3 = 0$. Because the change in slope was in the wrong direction (ie, indicating "inferior" to preintervention trend), the intervention was not deemed effective by this analysis.

A change in log-odds ratio (ie, the slope) can be visualized with the unadjusted results in Figure 2 in which there appears to be a positive trend in the preintervention period that is slightly attenuated in the postintervention period. Note that the vertical axis for Figure 2 is appropriately on the logit scale to coincide with the logistic regression model. If the data were instead plotted on the actual scale by plotting the actual percent or proportion with the outcome, the conclusion from the plot and a corresponding analysis on that raw scale (eg, using linear regression on the percent with outcome each month) might well be different—and incorrect.

We assessed the goodness of fit of the above model using the Hosmer–Lemeshow goodness-of-fit test, with a resulting *P* value of .03, indicating some evidence of lack of fit. A plot of standardized Pearson residuals showed several residuals with absolute value >2, and one of them (the last observation) >3. Removing the data point with the largest residual (−3.3, the final month in the study) improved the fit of the model (*P* = .26); this gave an almost identical estimate of the difference in slopes (*P* value still <.001), but now shows a statistically significant drop at the start of intervention (*P* = .034 compared to *P* = .16). But because there is no particular reason to believe the last month's data were erroneous, we retain the original model as the primary result. From a plot of the confounder-adjusted residuals over time, there did not appear to be evidence of strong autocorrelation over time (data not shown). Residuals also appeared to be normally distributed, both visually and passing the Shapiro–Wilks test for normality.

First Record Per Patient. Table 2 also gives results on the aggregated data using only the first record per patient, including 93,970 records. Results on the preintervention trend (*P* < .001) and change in trend postintervention (*P* < .001) are very close, with the same conclusion, compared to using the full dataset at the top of Table 2.

But in this analyses, there is also evidence of a reduction in outcome events (ie, rejecting the null hypothesis that $\beta_2 = 0$) at the start of the intervention, with an odds ratio (95% CI) of 0.65 (0.51–0.84) indicating lower odds of the outcome. Very similar results were found when using the last instead of the first observation for each patient (results not shown). Finally, the unadjusted results at the bottom of Table 2 for this same events/trials logistic segmented regression give the same conclusions as the adjusted results, both using all data.

Predictions "With" Versus "Without" Intervention. From this model, we can also test whether, for example, the trend modeled with the observed data in the intervention period differs from what would have been expected if the previous trend would have continued. We did this by applying Equations 3, 4, and 5 below (* indicates application of previously defined equations of the same number):

$$\text{logit}(\hat{P}_{60\,(\text{with})}) = \hat{\beta}_0 + \hat{\beta}_1\,60 + \hat{\beta}_2\,1 + \hat{\beta}_3\,24 + \text{covariates}, \quad (3)^*$$

$$\text{logit}(\hat{P}_{60\,(\text{w/o})}) = \hat{\beta}_0 + \hat{\beta}_1\,60 + \qquad + \text{covariates}, \quad (4)^*$$

Then the difference between (3)* and (4)* is

$$\hat{\beta}_2 + \hat{\beta}_3 \times 24, \quad (5)^*$$

which gives an estimate of the absolute difference in the predicted logit of the outcome at 60 months (ie, 24 months after the start of intervention) between the preintervention trend and the postintervention trend.

Our estimates give $\text{logit}(\hat{P}_{60\,[\text{with intervention}]}) - \text{logit}(\hat{P}_{60\,[\text{without intervention}]}) = 1.493 - 2.196 = -0.703$ or using 5*, equals $\hat{\beta}_2 \times 1 + \hat{\beta}_3 \times 24 = -0.1579 - 0.0237 \times 24 = 0.0198 - 0.1512 = -0.727$ which differs only by rounding error.

Then using Equation 5, we estimate the standard error of the difference in 5* by:

$$\begin{aligned}\text{SE}_{(5)=(3)-(4)} &= \sqrt{\text{var}(\hat{\beta}_2) + (\text{predict } t - \text{start } t)^2\,\text{var}(\hat{\beta}_3)}\\ &= \sqrt{\text{var}(\hat{\beta}_2) + 24^2\,\text{var}(\hat{\beta}_3)}\\ &= \sqrt{0.1128^2 + 24^2\,(0.0064^2)} = \sqrt{0.034} = 0.185\end{aligned}$$

We then conducted a z-test assessing whether the difference in the predictions (considering the postintervention data or not) at 24 months into the intervention period equals zero by comparing z = difference/standard error = −0.703/0.185 = −3.79 to a standard normal distribution, for which the 2-sided test yields $P < .001$. Results indicate a statistically significant lower predicted logit (and corresponding lower probability of the outcome) with versus without the intervention at 24 months into the intervention period. Comparing predictions at 36 and 48 months after the start of intervention were even more significant, in line with the estimated decrease in log odds (or slope) of the outcome in the postintervention period reported in Table 2. Such an association could further be expressed as a percent difference in the predicted values.[14] Note that Equations 3 and 4 and their difference are given on the logit scale above because we are using a logistic regression model and the logit, or $\log(p/[1-p])$, where p is the probability of outcome, is a normally distributed variable, enabling inference with standard statistical tests. The estimated probability of outcome for either equation can be calculated by $\exp\left(\text{logit}\right)/\left(1+\exp\left[\text{logit}\right]\right)$.

Analysis 2: Case-Level Data

Case-level analyses for the outcome of rapid response activation (yes/no) were conducted on cases for which all listed potential confounding variables were available. We conducted 3 separate segmented regression analyses, with results given in Table 3.

All Data, Confounder Adjusted. For the primary analysis (top of Table 3), multiple records were retained for patients with >1 encounter. Segmented regression was conducted using a generalized estimating equation logistic model for 93,970 patients with 151,145 observations to adjust for confounding and within-subject correlation (exchangeable correlation assumed) on the repeated measurements for some subjects. This within-subject correlation is a different entity from the possible autocorrelation across data in a time series that was discussed earlier. The preintervention slope was significantly positive (ie, rejecting H0: $\beta_1 = 0$, $P < .001$). However, there was little evidence of a change in the outcome at the start of intervention to reject H0: $\beta_2 = 0$ ($P = .87$). Finally, there was insufficient evidence to reject H0: $\beta_3 = 0$ ($P = .051$) of the slope changing in the postintervention period compared to preintervention at the .05 significance level, although it was close being deemed a more negative slope. In this analysis, therefore, the intervention was not deemed effective (ie, neither β_2 nor β_3 differed from 0).

First Record Per Patient, Confounder Adjusted. A second analysis used only the first observation for each patient, resulting in N = 93,970 observations. Using segmented regression and adjusting for confounding in a logistic regression model again showed a significantly positive slope in the preintervention period ($P = .001$), not sufficient evidence of change in intercept at the start of intervention ($P = .42$), and a decline in the slope (ie, in the unexpected direction) in the postintervention period compared to preintervention ($P = .003$).

First Record for Each Patient: Crude/Unadjusted. Finally, analysis was done on the first observation per patient, but with no adjustment for confounding. Conclusions were the same as in Analysis B, so that confounding adjustment had little noticeable effect.

Application Summary

The conclusion from crude analyses ignoring the time trends was that the anesthesiologist-led Adult Medical Emergency Team was associated with better outcomes. In contrast, segmented regression analyses did not identify a benefit.

From the aggregated data on the rapid response activation outcome, our first conclusion was that the outcome increased over time in the preintervention period in each analysis, that is, when using all data, when using either the first or last record per patient, and whether or not adjusting for confounding. There was also a consistent, unexpected decrease in the log odds or slope of the outcome in postintervention compared to preintervention. Finally, there was no evidence of an abrupt change in the outcome just after the start of intervention compared to the end of preintervention when using all data, although a reduction was detected when using either the first or last observation per patient.

Results from the case-level data were similar to those of the aggregated data in having an increasing preintervention slope that decreased postintervention, but there was no evidence of an abrupt reduction after the start of intervention. In this study, more credence might be given to the subject-level data in Table 3 because we were able to use all patient data and adjust for within-subject correlation using generalized estimating equations, whereas the aggregated data analysis in Table 2 was not adjusted for within-subject correlation.

Our overall conclusion was that there was no evidence that changing the Adult Medical Emergency Team to being anesthesiologist led had a positive effect on the utilization of rapid response. Rather, the intervention might have slowed the progress in rapid response usage being made before the intervention.

DISCUSSION

Major health system processes are changed under the assumption that the changes will enhance care, reduce cost, or otherwise improve health care. But just as with drugs and

Table 3. Analysis 2: Case-Level Segmented Regression Analyses for Adult Medical Emergency Team Data on Binary Rapid Response Variable

Parameter	Estimate	Standard Error	Odds Ratio (95% CI)	*P* Value
All data, confounder adjusted: 151,145 records on 93,970 patients generalized estimating equation logistic model to adjust for within-subject correlation				
Intercept, β_0	−5.25	0.14	…	…
Time, β_1	0.0198	0.0028	1.020 (1.014–1.026)	<.001
Intervention, β_2	−0.0112	0.0669	1.011 (0.867–1.127)	.87
Time_post, β_3	−0.0063	0.0032	0.994 (0.988–1.000)	.051
First record per patient, confounder adjusted. Logistic regression model. N = 93,970				
Intercept, β_0	−5.35	0.15	…	…
Time, β_1	0.028	0.004	1.028 (1.021–1.035)	<.001
Intervention, β_2	−0.07	0.09	0.93 (0.79–1.10)	.42
Time_post, β_3	−0.012	0.004	0.988 (0.980–0.996)	.003
First record for each patient: crude/ unadjusted. Logistic regression model. N = 93,970				
Intercept, β_0	−4.727	0.0819	…	…
Time, β_1	0.0308	0.0035	1.031 (1.024–1.038)	<.001
Intervention, β_2	−0.1332	0.0854	0.875 (0.740–1.035)	.12
Time_post, β_3	−0.0129	0.0040	0.987 (0.979–0.995)	.001

Same outcome aggregated by month in Table 2. Parameters explained in Table 2, Equation 1.

Time: months since start of study (1–36 preintervention and 37–79 postintervention). Intervention: 0 for preintervention, 1 for postintervention. Time_post: equals 0 in preintervention period and starts at 1 for first month in postintervention. Results: All 3 case-level analyses show a positive odds ratio for the preintervention slope ("time" variable), and a nonsignificant odds ratio at the start of intervention with the "intervention" variable. Using all data did not show a change in the slope between pre and post (*P* = .051), while both of the "first record per patient" analyses showed a change in the counterintuitive direction (ie, odds ratio significantly <1.0).

devices, the history of health care is that practice or system changes that seem likely to be beneficial often are not. Novel drugs and devices are always formally tested—and often fail. Changes in health system processes should similarly be tested; even if there is little doubt about benefit, the magnitude of the benefit is of considerable interest and needs to be quantified for accurate economic analyses.[51,52]

When cluster randomized trials are not feasible, the effect of a new policy or practice can sometimes be validly assessed using segmented regression[14,53–56] or difference-in-difference[29] methods. However, the ability to make valid inference in such studies depends on how well they are planned and conducted, the underlying phenomenon being studied, and rigor of the statistical methods. For example, a before–after design is strengthened considerably by longer time periods (eg, ≥1 year in each preintervention and postintervention), stable trends within time periods, availability of confounding variables, and sufficient data at multiple time points within each period.[14,18,57] Inference is further strengthened when concurrent controls are included, facilitating a difference-in-difference analysis on the means[29] or, if the periods are sufficiently long, on the slopes.

An interrupted time series design using segmented regression might not require adjustment for confounding variables if patient, provider, and institution characteristics do not change over time and are therefore similar before and after intervention.[18] Similarly, many of the perceived threats to validity in an interrupted time series design are diminished to the extent that they result in a consistent change (or lack of change) in the outcome measured over time, and similarly in both periods. For example, one might assume that there were continual improvements in medicine over the preintervention and postintervention periods, but that the effects of these improvements on the outcome(s) being studied were consistent across the entire study period. If so, then a non–confounder-adjusted comparison of the slopes and intercepts between the periods of interest in a segmented regression could be a valid analysis. However, such condition are unlikely and difficult or impossible to verify. We, therefore, argue for a more conservative and less "assuming" approach—adjusting as thoroughly as possible for available potentially confounding variables.

We conducted segmented regression on aggregated data as well as on subject-level data in our example of anesthesiologist-led versus medicine-led medical response teams, and generally reached the same conclusions from each. With either method, it is important to display aggregated data, so investigators can confirm that there are stable trends in each period, and whether the trends appear linear. Analyzing raw data generally improve power and allow inclusion of more detail as well as the natural variability in the outcomes and predictor variables. However, sometimes only the aggregated data are available. As well, aggregating data may help remove some effects of autocorrelation[21]; for example, there may be complex autocorrelations in the subject-level serial data (eg, seasonal/holiday and provider effects) that are difficult to identify and completely remove by standard autocorrelation adjustments.[24] However, when the main goal is to compare linear trends over extended periods of time, as in segmented regression, we argue that the presence of some residual autocorrelation and local perturbances may not be as relevant as in other settings (eg, when comparing concurrent groups on the mean).

When conducting a segmented regression researchers may find that some of the main parameters of interest, such as slope, difference in slopes, or the intervention effect are not significant, suggesting that the model could be simplified. However, such simplifications can lead to unnatural interpretations; it is thus usually preferable to use a full model, which has the added benefit of providing a more precise estimate of the effects of interest.

Threats to validity remain even after careful implementation of a segmented regression. First, there may have been competing interventions during the study. While known interventions can be included in a segmented regression analysis fairly easily, there are often many smaller interventions that are either unknown or cannot be measured easily or accurately. Another threat to validity is change in instrumentation or in the ability to measure the outcome of interest over the course of the study. Finally, even after thoroughly adjusting for patient, provider, and institution characteristics at each time point, selection bias may remain, especially if the composition of the population changes during the study period.

As with all observational studies, sensitivity analyses for segmented regression are encouraged as they help assess robustness of the chosen design and statistical methods.[58] For example, robustness to the length of time intervals within each period could be assessed, as well as the effect of artificially introducing periodic perturbations to assess the robustness of the model to unknown events. In segmented regression, a good sensitivity analysis to assess how well confounding and other biases have been accounted is to assess the intervention effect on a variable that should not be affected by the intervention, in a so-called falsification or placebo analysis.[14,59] If the intervention effect on such a purportedly unrelated variable is similar to the actual outcome variable of interest, and particularly if both show either a positive or negative "effect," the main results should be considered suspect. We conducted such an analysis for the current study on the outcome length of hospital stay. Results (only presented here) indicated nonsignificant findings for preintervention slope (P = .39), change in mean at the start of intervention (P = .58), and pre–post difference in slope (P = .74), showing quite different trends over time and tests of pre–post differences from the outcome of interest, and thus somewhat bolstering the validity of our conclusions.

Segmented regression can easily be extended to include multiple intervention "events," such as a progressive set of quality improvement initiatives.[14] In such cases instead of modeling and testing trends in a single postintervention period compared to preintervention, the model can be extended to test several either adjacent or even overlapping intervention period. One would include the relevant segmented regression parameters for each new intervention (eg, binary indicator for each specific intervention, time variable starting at 1 for beginning of each new intervention, and intervention-by-time interaction variable).[14] Segmented regression can similarly handle "breaks" between the preintervention and intervention periods, or between subsequent interventions, by properly coding the time variables. Along these lines, stepped-wedge designs need to space the "steps" between randomized inclusion of clusters to allow sufficient time for the intervention to be implemented and the outcomes measured.

Control charts can also be used to assess whether perturbations in a steady preintervention period coincide with the start of the intervention or with other known events.[17,60,61] Change-point analyses[62,63] for time series can be applied in a before–after design to estimate the point(s) in time associated with a change in the trend, and compare to timing of known events. Traditional segmented regression assumes a linear trend (including linearity in the logit for binary outcomes) at least in the preintervention period, but nonlinear interrupted time series analyses can be used when appropriate.[64] Segmented regression can also be used to make inference on the nature of an intervention effect (eg, immediate versus gradual and transient versus persistent).

When using segmented regression or difference-in-difference, we recommend emphasizing the limitations of the design and analyses even more than in a traditional cross-sectional or observational study with a concurrent control group, where the main issues might "only" be residual confounding, measurement bias and ascertainment bias.

In summary, researchers and quality improvement professionals should appropriately and prospectively design studies to adequately assess the impact of new interventions, policies, and processes. Cluster randomized trials are the strongest approach, but are often impractical. When before–after or interrupted time series designs are selected, robust and well-designed segmented regression or difference-in-difference analyses can help make appropriate inferences. ■

DISCLOSURES

Name: Edward J. Mascha, PhD.
Contribution: This author helped design the study, write and revise all sections, and analyze the data.
Name: Daniel I. Sessler, MD.
Contribution: This author helped write and revise the manuscript.
This manuscript was handled by: Thomas R. Vetter, MD, MPH.

REFERENCES

1. Devereaux PJ, Yusuf S. The evolution of the randomized controlled trial and its role in evidence-based decision making. *J Intern Med*. 2003;254:105–113.
2. Sessler DI, Imrey PB. Clinical research methodology 3: randomized controlled trials. *Anesth Analg*. 2015;121:1052–1064.
3. Glynn RJ, Brookhart MA, Stedman M, Avorn J, Solomon DH. Design of cluster-randomized trials of quality improvement interventions aimed at medical care providers. *Med Care*. 2007;45:S38–S43.
4. Cherniak W, Anguyo G, Meaney C, et al. Effectiveness of advertising availability of prenatal ultrasound on uptake of antenatal care in rural Uganda: a cluster randomized trial. *PLoS One*. 2017;12:e0175440.
5. Takasaki H. Mechanical diagnosis and therapy enhances attitude toward self-management in people with musculoskeletal disorders: a preliminary evidence with a before-after design. *SAGE Open Med*. 2017;5:2050312117740986.
6. Costantini M, Di Leo S, Beccaro M. Methodological issues in a before-after study design to evaluate the Liverpool care pathway for the dying patient in hospital. *Palliat Med*. 2011;25:766–773.
7. Wright DB. Comparing groups in a before-after design: when t test and ANCOVA produce different results. *Br J Educ Psychol*. 2006;76:663–675.
8. Solberg BC, Dirksen CD, Nieman FH, et al. Introducing an integrated intermediate care unit improves ICU utilization: a prospective intervention study. *BMC Anesthesiol*. 2014;14:76.
9. Stephen MS. Regression towards the mean, historically considered. *Stat Methods Med Res*. 1997;6:103–114.
10. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol*. 2005;34:215–220.

11. McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M, Fisher P. The Hawthorne Effect: a randomised, controlled trial. *BMC Med Res Methodol.* 2007;7:30.
12. Franke RH, Kaul JD. The Hawthorne experiments: first statistical interpretation. *Am Sociol Rev*. 1978;43:623–643.
13. Ho AMH, Phelan R, Mizubuti GB, et al. Bias in before-after studies: narrative overview for anesthesiologists. *Anesth Analg.* 2018;126:1755–1762.
14. 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.
15. Slack MK, Draugalis JR. Establishing the internal and external validity of experimental studies. *Am J Health Syst Pharm.* 2001;58:2173–2181.
16. Fong Y, Di C, Huang Y, Gilbert PB. Model-robust inference for continuous threshold regression models. *Biometrics.* 2017;73:452–462.
17. Fretheim A, Tomic O. Statistical process control and interrupted time series: a golden opportunity for impact evaluation in quality improvement. *BMJ Qual Saf.* 2015;24:748–752.
18. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int J Epidemiol.* 2017;46:348–355.
19. Muggeo VMR. Testing with a nuisance parameter present only under the alternative: a score-based approach with application to segmented modelling. *J Stat Comput Simul*. 2016;86:3059–3067.
20. Fong Y, Huang Y, Gilbert PB, Permar SR. chngpt: threshold regression model estimation and inference. *BMC Bioinformatics.* 2017;18:454.
21. Dexter F, Marcon E, Epstein RH, Ledolter J. Validation of statistical methods to compare cancellation rates on the day of surgery. *Anesth Analg.* 2005;101:465–473.
22. Mascha EJ, Turan A. Joint hypothesis testing and gatekeeping procedures for studies with multiple endpoints. *Anesth Analg.* 2012;114:1304–1317.
23. McCullagh P, Nelder J. *Generalized Linear Models*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC; 1989.
24. Moore IC, Strum DP, Vargas LG, Thomson DJ. Observations on surgical demand time series: detection and resolution of holiday variance. *Anesthesiology.* 2008;109:408–416.
25. Lagarde M. How to do (or not to do). Assessing the impact of a policy change with routine longitudinal data. *Health Policy Plan.* 2012;27:76–83.
26. SAS Institute Inc. *SAS/ETS 9.1 User's Guide.* Cary, NC: SAS Institute Inc; 2004.
27. Ostrom CW Jr. *Time Series Analysis: Regression Techniques. 2nd ed. Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07–009*. Newbury Park, CA: Sage Publications Inc; 1990.
28. Zhou H, Taber C, Arcona S, Li Y. Difference-in-differences method in comparative effectiveness research: utility with unbalanced groups. *Appl Health Econ Health Policy.* 2016;14:419–429.
29. Buckley J, Shang Y. Estimating policy and program effects with observational data: the "differences-in-differences" estimator. *Pract Assess Res Eval*. 2003;8:1–8. Available at: http://PAREonline.net/getvn.asp?v=8&n=2. Accessed January 12, 2016.
30. Cataife G, Pagano MB. Difference in difference: simple tool, accurate results, causal effects. *Transfusion.* 2017;57:1113–1114.
31. Hyldgård VB, Laursen KR, Poulsen J, Søgaard R. Robot-assisted surgery in a broader healthcare perspective: a difference-in-difference-based cost analysis of a national prostatectomy cohort. *BMJ Open.* 2017;7:e015580.
32. Goodman RM. Effect of cost efficiency reporting on utilization by physician specialists: a difference-in-difference study. *Health Serv Manage Res.* 2012;25:173–189.
33. Wing C, Simon K, Bello-Gomez RA. Designing difference in difference studies: best practices for public health policy research. *Annu Rev Public Health.* 2018;39:453–469.
34. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA.* 2014;312:2401–2402.
35. Sun E, Dexter F, Miller TR. The effect of "opt-out" regulation on access to surgical care for urgent cases in the United States: evidence from the national inpatient sample. *Anesth Analg.* 2016;122:1983–1991.
36. MacBride-Stewart S, Marwick C, Houston N, Watt I, Patton A, Guthrie B. Evaluation of a complex intervention to improve primary care prescribing: a phase IV segmented regression interrupted time series analysis. *Br J Gen Pract.* 2017;67:e352–e360.
37. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ.* 2015;350:h391.
38. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials.* 2007;28:182–191.
39. Hughes JP, Granston TS, Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. *Contemp Clin Trials.* 2015;45:55–60.
40. Zhan Z, van den Heuvel ER, Doornbos PM, et al. Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study. *J Clin Epidemiol.* 2014;67:454–461.
41. Matthews JNS, Forbes AB. Stepped wedge designs: insights from a design of experiments perspective. *Stat Med.* 2017;36:3772–3790.
42. Gao F, Earnest A, Matchar DB, Campbell MJ, Machin D. Sample size calculations for the design of cluster randomized trials: a summary of methodology. *Contemp Clin Trials.* 2015;42:41–50.
43. Deke J. Design and analysis considerations for cluster randomized controlled trials that have a small number of clusters. *Eval Rev*. 2016;40:444–486.
44. Cook AJ, Delong E, Murray DM, Vollmer WM, Heagerty PJ. Statistical lessons learned for designing cluster randomized pragmatic clinical trials from the NIH Health Care Systems Collaboratory Biostatistics and Design Core. *Clin Trials.* 2016;13:504–512.
45. Zhang F, Wagner AK, Ross-Degnan D. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. *J Clin Epidemiol.* 2011;64:1252–1261.
46. Kontopantelis E. *ITSPOWER: Stata Module for Simulation Based Power Calculations for Linear Interrupted Time Series (ITS) Designs, Statistical Software Components S458492*. Chestnut Hill, MA: Boston College Department of Economics; 2018.
47. Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. *Acad Pediatr.* 2013;13:S38–S44.
48. Hemming K, Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *J Clin Epidemiol.* 2016;69:137–146.
49. Ribeiro DC, Milosavljevic S, Abbott JH. Sample size estimation for cluster randomized controlled trials. *Musculoskelet Sci Pract.* 2018;34:108–111.
50. van Breukelen GJ, Candel MJ. Calculating sample sizes for cluster randomized trials: we can keep it simple and efficient! *J Clin Epidemiol.* 2012;65:1212–1218.
51. Jarl J, Desatnik P, Peetz Hansson U, Prütz KG, Gerdtham UG. Do kidney transplantations save money? A study using a before-after design and multiple register-based data from Sweden. *Clin Kidney J.* 2018;11:283–288.
52. Picton P, Starr J, Kheterpal S, et al. Promoting a restrictive intraoperative transfusion strategy: the influence of a transfusion guideline and a novel software tool. *Anesth Analg.* 2018;127:744–752.
53. Leahy I, Johnson C, Staffa SJ, Rahbar R, Ferrari LR. Implementing a pediatric perioperative surgical home integrated care coordination pathway for laryngeal cleft repair. *Anesth Analg*. 2018 [Epub ahead of print].
54. Said ET, Sztain JF, Abramson WB, et al. A dedicated acute pain service is associated with reduced postoperative opioid requirements in patients undergoing cytoreductive surgery with hyperthermic intraperitoneal chemotherapy. *Anesth Analg.* 2018;127:1044–1050.
55. Bhutiani M, Jablonski PM, Ehrenfeld JM, McEvoy MD, Fowler LC, Wanderer JP. Decision support tool improves real and perceived anesthesiology resident relief equity. *Anesth Analg.* 2018;127:513–519.

56. Shah AC, Nair BG, Spiekerman CF, Bollag LA. Process optimization and digital quality improvement to enhance timely initiation of epidural infusions and postoperative pain control. *Anesth Analg*. 2019;128:953–961.
57. Karkouti K, McCluskey SA, Callum J, et al. Evaluation of a novel transfusion algorithm employing point-of-care coagulation assays in cardiac surgery: a retrospective cohort study with interrupted time-series analysis. *Anesthesiology.* 2015;122:560–570.
58. Vetter TR, Mascha EJ, Kilgore ML. Physician supervision of nurse anesthetists: to opt in or to opt out? *Anesth Analg.* 2016;122:1766–1768.
59. Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? *JAMA.* 2013;309:241–242.
60. Mohammed MA. Using statistical process control to improve the quality of health care. *Qual Saf Health Care.* 2004;13:243–245.
61. Thor J, Lundberg J, Ask J, et al. Application of statistical process control in healthcare improvement: systematic review. *Qual Saf Health Care.* 2007;16:387–399.
62. Schluter PJ, Hamilton GJ, Deely JM, Ardagh MW. Impact of integrated health system changes, accelerated due to an earthquake, on emergency department attendances and acute admissions: a Bayesian change-point analysis. *BMJ Open.* 2016;6:e010709.
63. Texier G, Farouh M, Pellegrin L, et al. Outbreak definition by change point analysis: a tool for public health decision? *BMC Med Inform Decis Mak.* 2016;16:33.
64. Lamberson WR, Firman JD. A comparison of quadratic versus segmented regression procedures for estimating nutrient requirements. *Poult Sci.* 2002;81:481–484.