International Journal of Public Health (2019) 64:637–642
https://doi.org/10.1007/s00038-018-1195-2

HINTS & KINKS

Evaluating the impact of health policies: using a difference-in-differences approach

Sahar Saeed[1] · Erica E. M. Moodie[1] · Erin C. Strumpf[1] · Marina B. Klein[2]

Received: 18 October 2018 / Revised: 27 November 2018 / Accepted: 19 December 2018 / Published online: 3 January 2019

Constrained healthcare resources worldwide have made evaluating the impact of population health interventions increasingly important to maximize health and equity, while minimizing costs. However, the effects of population-level exposures such as health policies can seldom be evaluated through randomized controlled trials (RCTs). The following article will examine how the difference-in-differences method can be used to estimate the causal effect of such interventions. While this method was formalized and is extensively used in the field of economics (Meyer 1995), its first application is believed to have originated in the field of public health in 1855 (Snow 1855). The difference-in-differences method emulates a randomized design by measuring changes in outcomes over time between exposed and control groups. But unlike an RCT where the researcher randomly assigns exposure status; in a difference-in-differences design, researchers use "natural experiments" to assign exposure status, thus known as a *quasi*-experimental model (Dimick and Ryan 2014; Ryan et al. 2015). Repeated outcome data are necessary to conduct a difference-in-differences analysis. The data can be in the form of longitudinal data (also known as panel data); sources may include payer/claims data, patient's electronic medical records or data from established cohort studies. Alternately, repeated cross-sectional data such as national surveys for example Demographic and Health Surveys (DHS) can be used.

In this paper, we describe the study design, how to parametrize regression models, and analytical considerations; we further provide published examples to illustrate the approach in practice. Like all methods, the difference-in-differences approach comes with strengths, assumptions and limitations, which we discuss and direct the reader to other resources.

Motivating example

We start with a hypothetical example. A researcher is interested in evaluating the impact of universal access to influenza vaccination (the intervention) on hospital admissions (the outcome). Designing an RCT would be expensive, take a considerable amount of time and may not be generalizable. Fortunately, a *natural experiment* was already underway; as of December 2012, one state modified their state-level healthcare coverage to include universal influenza vaccination. One approach would be to conduct a pre-/post-design, where changes in outcomes post-intervention (after December 2012) are compared to pre-intervention (before December 2012) (Fig. 1 panel I). If the unit of observation (aggregate hospital-level data or more granular individual-level data) was the same before and after the intervention, time-invariant confounders are controlled for by design (Saeed et al. 2018; Strumpf et al. 2017). However, for this study design to provide an unbiased association of the intervention, an implicit assumption is made that there are no time-varying confounders or underlying secular trends that may influence the outcome—a strong assumption that is rarely valid. For example, hospital admission rates could change regardless of the implementation of the new policy due to an aging population or an acute health event such as a particularly bad influenza season. Failing to account for these

✉ Erica E. M. Moodie
erica.moodie@mcgill.ca

Sahar Saeed
sahar.saeed@mail.mcgill.ca

Erin C. Strumpf
erin.strumpf@mcgill.ca

Marina B. Klein
marina.klein@mcgill.ca

[1] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, 1020 avenue des Pins Ouest, Montreal, QC H3A 1A2, Canada

[2] Division of Infectious Diseases/Chronic Viral Illness Service, Department of Medicine, Glen Site, McGill University Health Centre, Montreal, QC, Canada

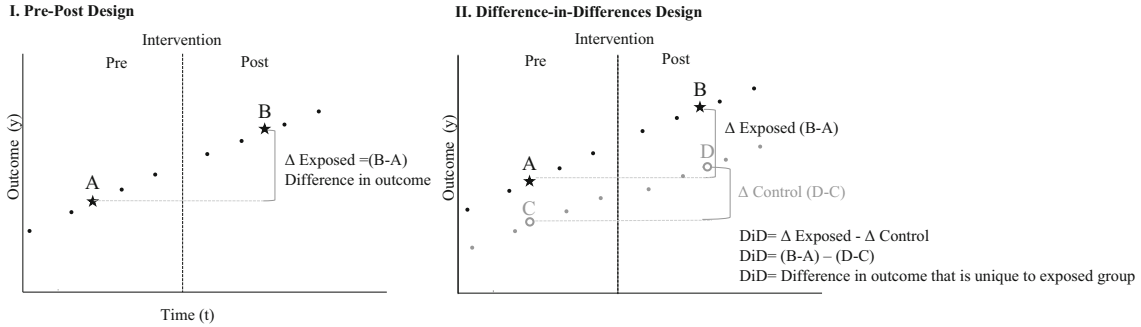

Fig. 1 Graphical representation of pre-post (I) and difference-in-differences (II) study designs. Stars indicate the outcome averaged over multiple times (black ·) among the exposed group [before and after the intervention (**A**, **B**)]; circles indicate the outcome averaged over multiple times (grey ·) among the control group [before and after the intervention (**C**, **D**)] (color figure online)

underlying secular changes would lead to erroneous (and counterintuitive) conclusions that the policy was associated with an increased rate of hospital admissions, when in fact other time-varying factors account for these changes. The difference-in-differences method takes the pre-/post-design one step further by including an external control group. Using our example including states that did not experience the policy change in 2012 could be defined as controls.

The design

As the name implies, this method compares the "difference" in outcomes of an exposed group, pre- and post-intervention, to the "difference" over the same time in a control group not subject to the intervention (Fig. 1 panel II) (Dimick and Ryan 2014; Strumpf et al. 2017). The crucial assumption to estimate an unbiased effect is that the only time-varying difference between the control and exposed groups is exposure to the intervention. Control groups should be chosen to be as similar in all ways to the treated group, except for the policy change, most importantly the trends in the outcome pre-intervention. By including a control group, the secular trends common to both groups are subtracted from the association between the intervention and the outcome. The double difference between pre- and post-intervention and between the control and exposed groups is known as the difference-in-differences (DiD) estimate, summarizing the average impact of the intervention. Figure 1 panel II illustrates the DiD estimate based on an average of multiple time points before and multiple time points after the policy implementation.

Assumptions of difference-in-differences approach

The strength of RCTs is that the randomization process should ensure that the exposed and control groups are exchangeable (both in terms of measured and unmeasured confounders) if the sample size is sufficient. Successful randomization is verified by comparing baseline characteristics of both groups (typically verified by results from a Table 1). By contrast, estimating an unbiased effect of an intervention with the difference-in-differences design requires the assumption that post-intervention trends of the control group provide an accurate counterfactual for "*what would have happened in the absence of the exposure*" (Little and Rubin 2000). Parallel pre-intervention trends (between the exposed and control groups) of the outcome are considered necessary and sufficient for this assumption to be reasonable. This assumption is verified by visually inspecting for parallel trends in the pre-intervention period (Abadie 2005), as illustrated in Fig. 2 panel I. In contrast, Fig. 2 panel II illustrates examples of how non-parallel trends can either result in counterfactual trends that either over-estimate or underestimate the impact of the interventions. Regression-based approaches can also be used to statistically test for parallel trends between the two groups in the pre-intervention period.

The validity of the difference-in-differences model also rests on the assumption that the intervention is *as good as random*, that is, independent of unobserved time-varying confounders, and that no other factors change differentially between the two groups over the study period. Supporting evidence for this assumption is provided by demonstrating minimal differences of observed characteristics between the exposed and control groups in the pre-intervention period, and by describing the motivation for, or the context of, the policy change. This is particularly important to rule out reverse causality—that is when changes in the outcome

Table 1 Limitations and solutions specific to difference-in-differences models

Potential limitations	Explanation	Solutions/sensitivity analyses	Published examples [study name, country and year of study]
Inappropriate control group	The critical limitation to implementing a difference-in-differences design is finding the right control group, which can be difficult in practice An inappropriate control group can include one where the parallel trends assumption is violated or if there is confounding by indication (the group that received the intervention was differentially chosen)	When appropriate control groups are not available, alternative methods using "synthetic controls" can be used to overcome this barrier. Synthetic controls aim to estimate treatment effects by constructing a weighted combination of control units, which represents what the treated group would have experienced in the absence of receiving the treatment (Noémi et al. 2016) Alternatively, researchers can carefully select a subset of control units that have an average pre-exposure outcome trend that is parallel to the exposed group	Examples of how researchers used synthetic controls: Estimating the effect of California's Tobacco Control Program [Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program, USA, 2010] (Abadie et al. 2010) Using financial incentives for kidney donations [Financial incentives for kidney donation: a comparative case study using synthetic controls, USA, 2015] (Bilgel and Galle 2015) Example of selecting a subset of control countries and averaging pre-exposure outcome trends [Removing user fees for facility-based delivery services: a difference-in-differences evaluation from ten sub-Saharan African countries, 10 sub-Saharan African Countries, 2015] (McKinnon et al. 2015)
Lead time effect	It is possible that outcomes may begin to change in anticipation of the intervention. If so, this suggests that changes in outcomes may have preceded the intervention, which can result in a biased estimate of the intervention effect (either attenuation or augmentation)	Lead time effects can be assessed by evaluating whether changes already started to occur during the pre-intervention period	Researchers interested in evaluating the association between same-sex marriage policies and adolescent suicide assessed changes in suicide attempts two years before the policy changes, by including a lead time indicator variable (available in author's supplemental materials) [Difference-in-differences analysis of the association between state same-sex marriage policies and adolescent suicide attempts, USA, 2017] (Raifman et al. 2017)
Lagged effect	Interventions such as policies are not always implemented immediately. System-level changes such as policy implementations may have greater reach to populations; however, their impact on individual health outcomes may not be immediate. This will result in a dilution, or underestimate, of the true effect of the intervention	A priori, based on substantive knowledge, a lagged impact model can be parameterized if it is known how long the latency period should last	In a study evaluating the association between user fees for facility-based delivery services and the portion of births delivered by cesarean section and neonatal mortality rates, researchers ran a series of regression models to test both lead and lagged effects and illustrated the impact visually at − 3, − 2, − 1 years (pre-intervention) and + 1 post-intervention [Removing user fees for facility-based delivery services: a difference-in-differences evaluation from ten sub-Saharan African countries, sub-Saharan African countries, 2015] (McKinnon et al. 2015)
Residual confounding	Also known as a "shock" in the economics literature. A time-varying confounder that differentially affects either the exposed or control group and coincides with the timing of the policy change of interest. This can result in a biased estimate of the intervention effect (either attenuation or augmentation)	Residual confounders need to be assessed based on substantive knowledge of the setting in which the policy is being assessed *Sensitivity analysis:* The use of a *placebo outcome* or *negative control test (Lipsitch* et al. 2010*)* (an outcome you would not expect to be affected by the exposure) may be used	To identify whether uterine rupture impacted hospital's vaginal birth after cesarean delivery rate, researchers used gestational diabetes as a placebo outcome since there was no possibility of it being affected by the occurrence of uterine rupture. This was to test if there were any differential changes in outcome rates among the exposed hospitals at the time of the rupture, such as a change in hospital protocol procedures [Effect of uterine rupture on a hospital's future rate of vaginal birth after cesarean delivery, Canada, 2014] (Riddell et al. 2014)

Table 1 continued

Potential limitations	Explanation	Solutions/sensitivity analyses	Published examples [study name, country and year of study]
Spillover effect	Similar to clustered-RCT, there may be a possibility of a "spillover" effect. That is when outcomes in the control group (not exposed to the intervention) change due to proximity to the exposed group	Alternative control groups can be assessed to test this hypothesis. Of course, finding an alternative control group can be difficult in practice *Sensitivity analysis* can be performed to evaluate whether spillovers have occurred (as described)	While studying the effect of a new law that required property owners of abandoned buildings to install working doors and windows in all structural openings on crime rates in Philadelphia, investigators were concerned that crimes may spill over to neighboring (unexposed) neighborhoods. They evaluated potential spillovers by applying three varying degrees of contiguous radii expanding from each exposed site and then comparing the difference-in-differences estimates of the crime rates at each geographic level. A decrease in crimes immediately surrounding project sites, but increase in crimes in surrounding areas, would suggest a spillover or displacement of crimes. [A difference-in-differences study of the effects of a new abandoned building remediation strategy on safety, USA, 2015] (Kondo et al. 2015)

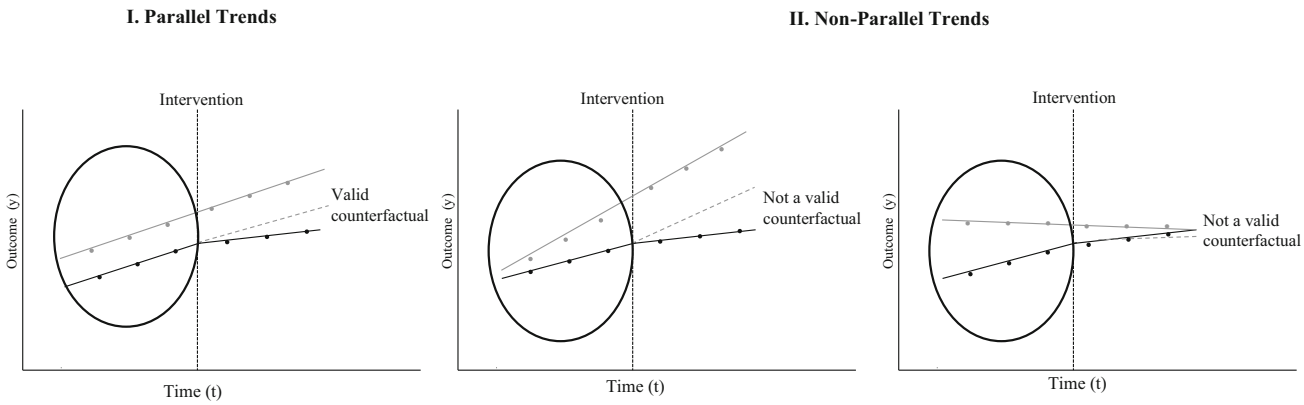

Fig. 2 Illustration of the parallel trends assumption of difference-in-differences method. Black line represents the exposed group, grey line represents the control group and dashed grey line represents the counterfactual trend (mirroring the control group). Panel I, illustrates parallel pre-intervention trends, resulting in a valid counterfactual. Panel II, illustrates two examples of non-parallel trends resulting in invalid counterfactual trends (color figure online)

during the pre-period may have motivated the policy change.

Modeling and statistical considerations

While arithmetic can be used to calculate average changes in outcomes between two groups, regression-based modeling is commonly used to control for time-varying confounders and efficiently calculate standard errors. Basic difference-in-differences models include three main variables in the regression models: an indicator variable for exposure (exposed group = 1, control group = 0), an indicator variable for time between pre- and post-intervention (post = 1, post = 0 (pre-intervention)) and the DiD estimator, which is the interaction between exposure and time:

$$Y = \beta_0 + \beta_1(\text{exposure}) + \beta_2(\text{post}) + \beta_3(\text{exposure}) * (\text{post}).$$

Figure 3 illustrates and provides interpretation of each coefficient of the regression model contrasting the pre-post and difference-in-differences designs. For linear regression models (used for continuous outcomes), the DiD estimator (β_3) describes the "excess" in outcome, controlling for secular trends (β_2) and time-invariant differences between the exposed and control groups (β_1) (see references (McCormick et al. 2015; Raifman et al. 2017; Riddell et al.

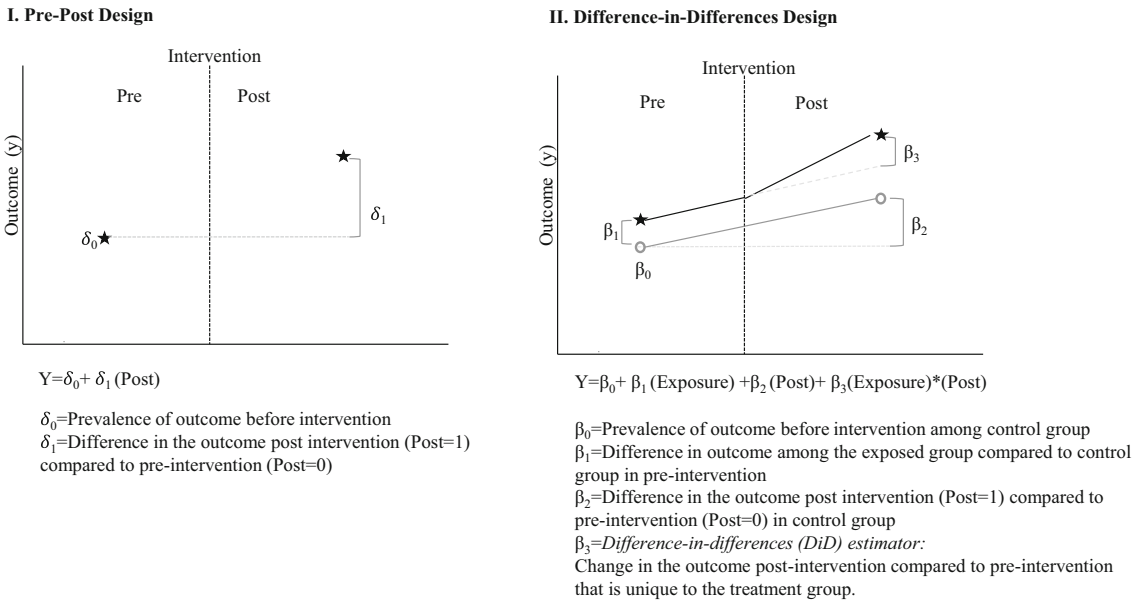

Fig. 3 Graphical representation of regression models for pre-post (I) and difference-in-differences (II) study designs. Stars indicate the outcome averaged over multiple time points among the exposed group (before and after the intervention); circles indicate the outcome averaged over multiple time points among the control group (before and after the intervention)

2014) for examples of linear DiD models). When outcomes are binary or counts, logistic or Poisson models are used, respectively, see references (Bendavid et al. 2012; Jena et al. 2015; King et al. 2013; Wharam et al. 2007). In contrast to the linear models, here β_3 describes a relative change as expressed as a change in odds, risk or rate ratios (Puhani 2012). While the interpretation of the DiD estimator is similar in both linear and binary models, particular attention should be made when evaluating the parallel trends assumptions when using nonlinear models. Specifically, outcome measurements should be plotted on the log scale to assess for parallel trends. Furthermore, since the same units of observation are repeatedly measured over time, outcomes will be correlated, violating the independence assumption of standard regression models (Liang and Zeger 1993). Modeling decisions on the nature of the correlation matrix are described elsewhere (Kmenta 2010) and should be taken into consideration.

Difference-in-differences models are flexible in the sense that they allow researchers to add multiple exposure/control groups, time periods, time-varying confounders and further evaluate effect modification (Bertrand et al. 2002; Strumpf et al. 2017). Other extensions can include using a categorical variable indicating varying "intensities" of the intervention instead of binary exposures. For example, researchers evaluated the impact of varying degrees of seatbelt laws (none, primary and secondary enforcement) on traffic fatal accident rates (Harper and Strumpf 2017). While the difference-in-differences approach has the strength of controlling for secular trends and fixed differences between groups to provide unbiased estimates of the impact of the intervention, there are limitations specific to this study design, summarized in Table 1. Published work demonstrating solutions to these limitations or examples of sensitivity analyses are referenced.

Conclusion

Quasi-experimental designs such as the difference-in-differences approach can provide an alternative to evaluating the impact of interventions such as public health policies when RCTs are not feasible. By including control group(s) to act as the counterfactual, time-varying confounders are controlled by design. In addition, this study design may be more intuitive and accessible to a diverse audience with the use of graphics to depict results. However, as we have reviewed in this article, there are specific assumptions and analytical considerations that need to be made when conducting a difference-in-differences design to accurately estimate the impact of the intervention of interest.

Funding This study was funded through support by Doctoral Awards funded to SS by the Canadian Institutes of Health Research and the Canadian Hepatitis C Network. ECS and EEMM are supported by a Chercheur boursier Junior 2 from the Fonds de Recherche Santé

(FRQ-S). The Canadian HIV-HCV Coinfection Cohort Study is supported by the Fonds de recherche du Québec-Santé (FRQ-S); Réseau SIDA/maladies infectieuses, the Canadian Institutes of Health Research (CIHR FDN 143270) and the CIHR Canadian HIV Trials Network (CTN222).

Compliance with ethical standards

Conflict of interest Authors SS, EEMM and ECS declare that they have no conflicts of interest. None of the authors feel in conflict of interest with regard to this study, and there was no pharmaceutical industry support to conduct this study although MBK has received research grants for investigator-initiated trials from Merck and ViiV Healthcare and consulting fees from ViiV Healthcare, Bristol-Meyers Squibb, Merck, Gilead and AbbVie.

References

Abadie A (2005) Semiparametric difference-in-differences estimators. Rev Econ Stud 72:1–19

Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J Am Stat Assoc 105:493–505. https://doi.org/10.1198/jasa.2009.ap08746

Bendavid E, Holmes CB, Bhattacharya J, Miller G (2012) HIV development assistance and adult mortality in Africa. JAMA 307:2060–2067. https://doi.org/10.1001/jama.2012.2001

Bertrand M, Duflo E, Mullainathan S (2002) How much should we trust differences-in-differences estimates? The National Bureau of Economic Research, 8841

Bilgel F, Galle B (2015) Financial incentives for kidney donation: a comparative case study using synthetic controls. J Health Econ 43:103–117. https://doi.org/10.1016/j.jhealeco.2015.06.007

Dimick JB, Ryan AM (2014) Methods for evaluating changes in health care policy: the difference-in-differences approach. JAMA 312:2401–2402. https://doi.org/10.1001/jama.2014.16153

Harper S, Strumpf EC (2017) Primary enforcement of mandatory seat belt laws and motor vehicle crash deaths American. J Prev Med 53:176–183. https://doi.org/10.1016/j.amepre.2017.02.003

Jena AB, Goldman DP, Seabury SA (2015) Incidence of sexually transmitted infections after human papillomavirus vaccination among adolescent females. JAMA Intern Med 175:617–623. https://doi.org/10.1001/jamainternmed.2014.7886

King M, Essick C, Bearman P, Ross JS (2013) Medical school gift restriction policies and physician prescribing of newly marketed psychotropic medications: difference-in-differences analysis. BMJ (Clin Res Ed) 346:f264. https://doi.org/10.1136/bmj.f264

Kmenta J (2010) Mostly harmless econometrics: an empiricist's companion. Springer, Berlin

Kondo MC, Keene D, Hohl BC, MacDonald JM, Branas CC (2015) A difference-in-differences study of the effects of a new abandoned building remediation strategy on safety. PLoS ONE 10:e0129582. https://doi.org/10.1371/journal.pone.0129582

Liang KY, Zeger SL (1993) Regression analysis for correlated data. Annu Rev Public Health 14:43–68. https://doi.org/10.1146/annurev.pu.14.050193.000355

Lipsitch M, Tchetgen Tchetgen E, Cohen T (2010) Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 21:383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb

Little RJ, Rubin DB (2000) Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Annu Rev Public Health 21:121–145. https://doi.org/10.1146/annurev.publhealth.21.1.121

McCormick D, Hanchate AD, Lasser KE, Manze MG, Lin M, Chu C, Kressin NR (2015) Effect of Massachusetts healthcare reform on racial and ethnic disparities in admissions to hospital for ambulatory care sensitive conditions: retrospective analysis of hospital episode statistics. BMJ (Clin Res Ed) 350:h1480. https://doi.org/10.1136/bmj.h1480

McKinnon B, Harper S, Kaufman JS, Bergevin Y (2015) Removing user fees for facility-based delivery services: a difference-in-differences evaluation from ten sub-Saharan African countries. Health Policy Plan 30:432–441. https://doi.org/10.1093/heapol/czu027

Meyer BD (1995) Natural and quasi-experiments in economics. J Bus Econ Stat 13:151–161. https://doi.org/10.2307/1392369

Noémi K, Richard G, Dominik H, James TA, Silviya N, Matt S (2016) Examination of the synthetic control method for evaluating health policies with multiple treated units health. Economics 25:1514–1528. https://doi.org/10.1002/hec.3258

Puhani PA (2012) The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. Econ Lett 115:85–87. https://doi.org/10.1016/j.econlet.2011.11.025

Raifman J, Moscoe E, Austin SB, McConnell M (2017) Difference-in-differences analysis of the association between state same-sex marriage policies and adolescent suicide attempts. JAMA Pediatr 171:350–356. https://doi.org/10.1001/jamapediatrics.2016.4529

Riddell CA, Kaufman JS, Hutcheon JA, Strumpf EC, Teunissen PW, Abenhaim HA (2014) Effect of uterine rupture on a hospital's future rate of vaginal birth after cesarean delivery. Obstet Gynecol 124:1175–1181. https://doi.org/10.1097/aog.0000000000000545

Ryan AM, Burgess JF Jr, Dimick JB (2015) Why we should not be indifferent to specification choices for difference-in-differences. Health Serv Res 50:1211–1235. https://doi.org/10.1111/1475-6773.12270

Saeed S, Moodie EEM, Strumpf EC, Klein MB (2018) Segmented generalized mixed effect models to evaluate health outcomes. Int J Public Health 63:547–551. https://doi.org/10.1007/s00038-018-1091-9

Snow J (1855) On the mode of communication of cholera, 2nd edn. John Churchill, London

Strumpf EC, Harper S, Kaufman JS (2017) Fixed effects and difference in differences. In: Oakes JM and Kaufman JS (eds) Methods in social epidemiology. Jossey-Bass, San Francisco

Wharam JF, Landon BE, Galbraith AA, Kleinman KP, Soumerai SB, Ross-Degnan D (2007) Emergency department use and subsequent hospitalizations among members of a high-deductible health plan. JAMA 297:1093–1102. https://doi.org/10.1001/jama.297.10.1093