CASFM Methods Briefs

Difference-in-differences analysis and the propensity score to estimate the impact of non-randomized primary care interventions

Claire Godard-Sebillotte^{a,*}, Sathya Karunanathan^b and Isabelle Vedel^a

^aDepartment of Family Medicine, McGill University, Montreal, Canada and ^bOttawa Hospital Research Institute, Ottawa, Canada.

*Correspondence to Claire Godard-Sebillotte, Department of Family Medicine, McGill University, 5858 Côte-des-Neiges Road, 3rd Floor, Montreal, Quebec H3S 1Z1, Canada; E-mail: claire.godard-sebillotte@mail.mcgill.ca

Key words: Bias, biostatistics, confounding factors, difference-in-differences analysis, primary care, propensity score.

Introduction

Primary care delivery has been deeply transformed over the last couple of decades in high-income countries (1). For example, multidisciplinary primary care teams have become widespread in the USA [patient-centred medical homes (PCMH)] and in some Canadian provinces: i.e. Family Health Teams (FHTs) in Ontario or Family Medicine Groups (FMGs) in Quebec. Evaluating the impact, or causal effect, of these interventions is key to inform evidence-based policies (2). As most reforms and policies in primary care, these interventions were implemented outside of any randomized experiment (1, 3). Evaluating non-randomized interventions, reforms and policies (referred to as non-randomized interventions hereafter) necessitates overcoming pitfalls of observational studies like confounding and selection bias (2).

Advanced statistical methods such as the difference-in-differences analysis and propensity scores have been developed to assess the impact of non-randomized interventions (4, 5). These two methods are (i) increasingly used to measure non-randomized interventions impact, especially in primary care (6, 7) and (ii) particularly suited to provide the evidence needed to inform evidence-based policies: estimating the impact of policies on a population-wide scale, in a real-life context (8).

In this method brief, we illustrate the challenges of evaluating non-randomized interventions with an example: the evaluation of the impact of FMGs in Quebec (Canada) on health service use and costs (8).

Impact evaluation of FMGs on health service use and costs in Quebec, Canada

FMGs, multidisciplinary primary care teams, were first implemented in 2000 in Quebec, Canada, to enhance primary care access, continuity and care quality (1). These FMGs were intended to include 6–12 full-time equivalent family physicians receiving additional funding

for nurses, administrative support staff and computer equipment (9). Family physicians enrolment in these FMGs was voluntary and not associated with major changes in their payment system (9). Patient enrolment was also voluntary and occurred while registering with a physician practicing in an FMG (8).

There is evidence that physicians and patients that enrolled in FMGs differed from the general population of Quebec physicians and patients (10). For example, physicians who enrolled in FMGs had fewer years in medical practice and a more diverse clinical practice (especially in hospitals) (10). Patients who enrolled in FMGs had a lower socio-economic status, and greater use of hospital services for total and potentially avoidable causes (10). These differences between patients and physicians who enrolled in FMGs and those who did not could bias, or confound, the estimation of the impact of FMGs (intervention) on health service use and costs, such as emergency department visits (outcome) (10), if they are not accounted for (Fig. 1). The factors associated with both FMG enrolment and emergency department visits are known as confounding factors.

Impact estimation in randomized controlled trials

An unbiased estimate of the impact of an intervention can be drawn when the intervention group and the control group are comparable or exchangeable (11). Exchangeability refers to the control group being an adequate substitute for the outcome that would have occurred in the intervention group had the intervention group not received the intervention (12). Exchangeability between groups is achieved when groups are similar for all measured and unmeasured confounders (11, 12). The exchangeability condition is also known as ‘no unmeasured confounding’ (4) and is the focus of this methods brief. Other assumptions, detailed elsewhere, are necessary to estimate impact: positivity and consistency (3, 4).

KEY MESSAGES

- Primary care interventions are often non-randomized
- Lack of randomization may create bias in the estimation of their impact
- Advanced statistical methods allow estimation of their impact

Exchangeability is achieved in well-conducted randomized controlled trials, thanks to the randomization procedure. As shown in Figure 2A, as a result of the random assignment, all individuals have the same probability of receiving the intervention, and the intervention and control groups only differ from each other with respect to the intervention of interest. In such cases, the estimation of the impact is not biased by confounding factors. The impact of the intervention on the outcome can be directly estimated by measuring the difference in outcomes between the intervention and control group.

Impact estimation in non-randomized interventions

In non-randomized interventions, the mechanism by which patients come to receive the intervention or not can be much more complex and is often poorly understood. As a result, the intervention and control groups may not be exchangeable (Fig. 2B). As shown in Figure 2B, when intervention and control groups are not exchangeable, estimating the impact of the intervention without taking into account these differences leads to a biased estimate.

In Quebec, the patients and physicians who enrolled in FMGs systematically differed from those that did not, creating two non-exchangeable groups. A comparison of emergency department visits before and after the implementation of FMGs between the persons enrolled in FMGs (intervention group) and the persons not enrolled in FMGs (control group) does not allow disentangling what is caused by the intervention (FMG) and (i) what is related to permanent differences between the control and the intervention groups (i.e. potential confounding factors) or (ii) what is due to temporal trends (Fig. 1). For example, higher emergency department visits in the FMG patient population may be attributed to their worse health status, rather than the intervention itself (10). Moreover, quality of care may have improved over time in Quebec, independently of the intervention, reducing emergency department visits (temporal trend). Taking into account these permanent differences between the intervention and control groups (potential

confounding factors) and temporal trends in the outcome is essential to accurately estimate the impact of FMGs.

The difference-in-differences analysis

The difference-in-differences analysis can be used to estimate the impact of non-randomized interventions. The difference-in-differences analysis allows disentangling the impact of the intervention from (i) permanent differences between control and intervention groups (potential confounding factors) and (ii) temporal trends in the outcome unrelated to the intervention (2). The impact of the intervention is estimated through the difference between two differences in the outcomes: (i) D1: difference between the pre- and post-intervention periods in the control group and (ii) D2: difference between the pre- and post-intervention periods in the intervention group (Fig. 3). This estimate is computed from a regression model, which includes two dichotomous variables: the time (before or after the intervention), the group (intervention or control group) and an interaction term between the time and the group (2, 6). The impact is estimated through the coefficient of this interaction term (2, 6). A model, compared with a simple subtraction, allows adjusting for potential confounders and thus reduces residual confounding. In addition, it allows estimating whether the impact of the intervention is significant or, in other words, whether the coefficient of the interaction term is significantly different from zero (2, 6). More detailed explanations are provided elsewhere (2, 12).

The difference-in-differences analysis relies on two assumptions:

- The common “shock” assumption’, stating that any events occurring during or after the time the intervention was implemented will equally affect the intervention and control groups (6).
- The ‘common trend assumption’, stating that the intervention and control group trends would have varied in a similar fashion after the intervention, the intervention group not having received the intervention (6).

Figure 1. Factors associated with both the intervention Family Medicine Group (FMG) and the outcome (emergency department visits): confounding factors of the impact of the intervention.

Figure 2. Impact estimation in (A) randomized controlled trials and (B) non-randomized intervention. (A) In a randomized controlled trial, random assignment (R) of the intervention creates exchangeable groups (FMG and non-FMG). Comparison of outcomes between the two groups provides an unbiased estimate of the impact of the intervention on the outcome. Red circles represent patients who experience the outcome: emergency department visits. (B) In a non-randomized intervention (Non-R), patients and physicians decide whether to enroll in FMGs, which may create groups that are not exchangeable. Comparison of outcomes between the two groups results in a biased estimate of the impact of the intervention on the outcome. Red circles represent patients who experience the outcome: emergency department visits.

Figure 3. Impact estimation with the difference-in-differences analysis. In the difference-in-differences analysis, the impact of the intervention is estimated through two differences in the outcomes: (i) D1: difference between the pre- and post-intervention average outcomes in the control group and (ii) D2: difference between the pre- and post-intervention average outcomes in the intervention group.

These two assumptions cannot be formally tested. However, they guide the choice of an appropriate control group that meets the two assumptions.

The common “shock” assumption should be questioned through discussion with stakeholders knowledgeable of the context. In our example, the absence of differential exposure between the intervention and control groups to other provincial policies or major

disease-modifying events has to be questioned. For example, did all patients enrolled in FMGs also belong to the same geographic region that was affected by a heat wave that increased emergency department visits?

According to the ‘common trend assumption’, outcomes in the control and intervention groups vary in a similar fashion over time.

Figure 4. Impact estimation in non-randomized intervention with propensity score: propensity scores assignment (A), matching (B), data pre-processing (C), impact estimation (D), compared with impact estimation in non-randomized intervention without propensity score (E). Preprocessing creates intervention and control groups that are exchangeable so that the comparison of outcomes between the groups provides an unbiased estimate of the impact of the intervention on the outcome. Red circles represent patients who experience the outcome. Red crosses represent patients who are removed from the analytical sample.

As such, trends in outcomes observed in the control group correspond to what would have been observed in the intervention group, had they not received the intervention (13). Plausibility of the ‘common trend assumption’ should be assessed through discussion with stakeholders knowledgeable of the context. In addition, comparability of the trends in the outcome before the intervention, between intervention and control groups, can be assessed graphically and with a formal statistical test.

Several sensitivity analyses may be performed to assess whether the impact estimation is unbiased and thus whether the impact measured can be attributed to the intervention. The impact of the intervention on placebo outcomes, which should not be affected by the intervention, may be measured. A null effect should be found. The estimated impact of the intervention at a calendar date before the intervention was implemented should also be null. The impact may also be measured using several suitable control groups and should be comparable. Restriction of the study population to the continuously enrolled participants may be considered, taking into account its potential influence on generalizability and study power to detect an impact (13).

In non-randomized interventions where the control and intervention groups are exchangeable and the above assumptions are met, difference-in-differences provides a reasonably unbiased estimate of the intervention’s impact. If however, exchangeability of the control and intervention groups is questionable, one may resort to the use of propensity scores.

Table 1. Propensity score (PS) use and impact estimation	
PS use	Impact estimation
PS matching	Compare outcomes of matched pairs of comparable participants while taking into account the matched nature of the data
Stratification on PS	Compare participant’s outcome within each strata or pooled across strata; strata of participants based on PS values
Covariate adjustment on PS	Include PS as a covariate in impact estimation model
Inverse probability of treatment weighting using the PS	Compare outcomes from intervention and control groups where participants have been assigned weights according to their PS

Propensity scores

A propensity score is a balancing score used to construct an analytical sample in which intervention and control groups are exchangeable. By creating exchangeable groups, the impact of non-randomized interventions can be estimated (14). A propensity score represents each participant’s probability of receiving the intervention based on their baseline characteristics, regardless of whether or not they actually received it. In our example, the probability of every participant to enrol in an FMG is modelled. The underlying assumption of using propensity scores is the absence of unmeasured confounders, as the propensity score can only

create balance between groups on measured variables. More detailed explanations are provided elsewhere (2, 7, 14–18).

Propensity scores can be used in different ways (Table 1). Propensity score matching is illustrated in the FMG impact evaluation example (Fig. 4A–E).

Specification of the propensity score follows an iterative approach until balance of baseline characteristics between intervention and control groups is achieved (14, 15). If balance is not achieved, specification of the propensity score can be improved by using additional predictors (14). The model is a multiple regression (i.e. logistic regression modelling the probability of being in the intervention group versus control group). Predictors included in the propensity score model are (i) potential confounding factors of the association between the intervention and the outcome or (ii) predictors of the outcome (14). Only variables measured at baseline should be included. Variables affected by the intervention are not included (14). There is no way to verify that the predicted probabilities reflect the true probabilities. The propensity score is considered correctly specified when balance of baseline characteristics between intervention and control groups is achieved. It is to be noted that balance can only be assessed in measured characteristics. If unmeasured confounders are not balanced between groups, groups will remain non-exchangeable. Presence of unmeasured confounders should be assessed through discussion with stakeholders knowledgeable of the context. The impact of unmeasured confounders can be explored through sensitivity analyses, testing the robustness of the impact estimation to unmeasured confounders (14, 16, 19, 20). Assessing balance in measured baseline characteristics is usually done with standardized differences (14). Standardized differences compare the differences in means in units of the pooled standard deviation. A threshold of less than or equal to 0.1 is recommended to consider the difference between the two groups negligible (14).

As a result of using propensity scores, the analytical sample of the non-randomized study now mimics a randomized controlled trial, where participants in the intervention and control groups only differ with respect to whether or not they received the intervention but had the same probability of receiving the intervention (Fig. 4D and E).

Conclusion

Most primary care interventions, reforms and policies are implemented outside of randomized experiments. The difference-in-differences analysis and propensity score can provide an unbiased estimation of the impact of these non-randomized interventions. These methods are therefore key to the evaluation of interventions to inform health policy makers and future policies.

Acknowledgements

The authors would like to thank Dr. Erin Strumpf for her helpful advice on the content and organization of this paper and Nadia Sourial, Dr. Geneviève Arsenault-Lapierre and Marine Hardouin for their assistance with the figures.

Declaration

Funding: This methods brief was funded by the Canadian Institutes of Health Research—Vanier Canada Graduate Scholarship.

Ethical approval: none.

Conflict of interest: none.

References

- Strumpf E, Levesque J-F, Coyle N *et al.* Innovative and diverse strategies toward primary health care reform: lessons learned from the Canadian experience. *J Am Board Fam Med* 2012; 25 (suppl 1): S27–33. doi:10.3122/jabfm.2012.02.110215.
- Gertler PJ, Martinez S, Premand P, Rawlings L, Vermeersch C. *Impact Evaluation in Practice*. Washington DC: The International Bank for Reconstruction and Development/The World Bank. http://siteresources.worldbank.org/EXT/HDOffice/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf (accessed on 26 April 2018).
- Sourial N, Longo C, Vedel I, Schuster T. Daring to draw causal claims from non-randomized studies of primary care interventions. *Fam Pract* 2018; 35: 639–43.
- Hernán MA. *Causal Inference Book*. <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/> (published 19 October 2012; accessed on 26 April 2018).
- Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; 11: 550–60.
- Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA* 2014; 312: 2401–2.
- Haukoos JS, Lewis RJ. The propensity score. *JAMA* 2015; 314: 1637–8.
- Strumpf E, Ammi M, Diop M, Fiset-Laniel J, Tousignant P. The impact of team-based primary care on health care services utilization and costs: Quebec's family medicine groups. *J Health Econ* 2017; 55: 76–94.
- Diop M, Fiset-Laniel J, Provost S *et al.* Does enrollment in multidisciplinary team-based primary care practice improve adherence to guideline-recommended processes of care? Quebec's Family Medicine Groups, 2002–2010. *Health Policy* 2017; 121(4): 378–88. doi:10.1016/j.healthpol.2017.02.001.
- Coyle N, Strumpf E, Fiset-Laniel J, Tousignant P, Roy Y. Characteristics of physicians and patients who join team-based primary care practices: evidence from Quebec's Family Medicine Groups. *Health Policy* 2014; 116: 264–72.
- Craig P, Katikireddi SV, Leyland A, Popham F. Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. *Annu Rev Public Health* 2017; 38: 39–56.
- Strumpf EC, Harper S, Kaufman, Jay S. Fixed effects and difference-in-differences. In: Oakes JM, Kaufman JS (Eds). *Methods in Social Epidemiology*. San Francisco: Jossey-Bass, 2nd edn. 2017.
- Stuart EA, Huskamp HA, Duckworth K *et al.* Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Serv Outcomes Res Methodol* 2014; 14: 166–82.
- Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011; 46: 399–424.
- Rosenbaum PR, Rubin DB. *The Central Role of the Propensity Score in Observational Studies for Causal Effects*. 1983. <https://dash.harvard.edu/handle/1/3382855> (accessed on 14 January 2018).
- Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. *J Econ Surv* 2008; 22(1): 31–72. doi:10.1111/j.1467-6419.2007.00527.x.
- Rassen JA, Shelat AA, Myers J *et al.* One-to-many propensity score matching in cohort studies. *Pharmacoepidemiol Drug Saf* 2012; 21(Suppl 2): 69–80. doi:10.1002/pds.3263.
- Brookhart MA, Schneeweiss S, Rothman KJ *et al.* Variable selection for propensity score models. *Am J Epidemiol* 2006; 163: 1149–56.
- Lash TL, Fox MP, MacLehose RF *et al.* Good practices for quantitative bias analysis. *Int J Epidemiol* 2014; 43: 1969–85.
- Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* 2011; 22: 42–52.