Methodological Considerations for Difference-in-Differences

Alyssa Bilinski, PhD, MS, AM; Ishani Ganguli, MD, MPH

In this issue of *JAMA Internal Medicine*, Apathy et al[1] present results of difference-in-differences (DiD) with electronic health record (EHR) metadata from Epic, examining changes in EHR use and visit volume after successful, voluntary adoption of team-based documentation support (eg, scribes). They found a decrease in documentation time and increase in visit volume, with larger effects for more intensive users.

Related articles pages 1212 and 1250

The study offers an opportunity to consider the merits and assumptions underlying DiD, a popular observational study design in which researchers estimate the average treatment effect on the treated by comparing pre-post differences between groups that were and were not exposed to a new treatment. DiD is based on a counterfactual parallel trends assumption: that absent the intervention, the treatment and comparison groups would have had parallel trajectories on average.[2,3] Although widely used and seemingly straightforward, DiD can be complex to apply and interpret, and in the past several years, many methodological studies have aimed to improve its reliability and transparency.[2,3]

The study by Apathy et al[1] incorporates several recent recommendations to strengthen DiD. For example, the authors present event study plots to provide evidence about the strength of the parallel trends assumption (Figure 2 in the study by Apathy et al[1]). Because the parallel trends assumption describes what would have happened to treated groups if there had not been an intervention, it cannot be directly tested. However, event study plots provide information about preintervention trends: if preintervention trends were parallel, this would increase the plausibility that trends would have remained parallel absent the intervention. In Figure 2,[1] each point on the event study plot shows a DiD estimate for the outcome at a given time (shown on the x-axis) relative to the last week prior to the intervention (x-axis equal to −1). For example, Figure 2A shows the difference in total weekly visits for treatment vs comparison groups at each time compared to the last preintervention week. Preintervention points (those to the left of the dotted vertical line at 0) can be thought of as placebo effect estimates. The DiD design is most reliable when these preintervention points are close to 0, have narrow error bars, and lack a discernible trend over time. Postintervention points on the event study plot show the distribution of treatment effects over time following treatment; effects that begin abruptly after intervention can lend credibility to a causal link between the treatment and the effect.

The authors also incorporated estimators designed to account for staggered treatment rollout. In this study, physicians adopted documentation support at different calendar times. In such cases, traditional DiD estimators can produce misleading or difficult-to-interpret estimates if treatment effects differ across adoption cohorts or are growing or shrinking over time.[2,4,5] To address this, the authors conducted sensitivity analyses using the Callaway and Sant'Anna estimator.[4] This, as well as other methods like that proposed by Sun and Abraham,[5] can appropriately account for staggered treatment timing.

Separately, this study prompts us to consider how we should interpret DiD studies that do not follow conventional practice in defining treatment (in this case, by defining the treatment group using a measure of that treatment's uptake). Specifically, the authors lacked information about documentation-support availability and therefore identified scribe adoption based on EHR metadata. They designated physicians as "treated" if they demonstrated a 1-time shift from no documentation support to consistent documentation support, each for a period of at least 4 weeks. This raises several nuances in interpreting the study results.

First, because treatment status was defined using uptake, this approach can only identify physicians who chose to adopt scribes and had some success in using them. Even within the sample meeting the inclusion criteria, there was significant heterogeneity in adoption, with the authors separately analyzing those with low adoption and those with high adoption. Though these details are not captured in this study, it is likely that some physicians deemed "untreated" were offered documentation support but chose not to use it. Indeed, prior research suggests that many physicians may not choose to adopt scribes when offered.[6-8] As a result, the effects shown in this study may be larger in magnitude than if the treated group had included all physicians who were offered or even briefly adopted scribes.

Heterogeneity in adoption also suggests limitations to the generalizability of study results. Physicians who were offered or adopted scribes likely differed from those without team-based documentation in several important ways, including specialty, clinical workflows, clinical full-time equivalence, ability to delegate, or comfort with technology. The authors note that because DiD estimates the average treatment effects on the treated (ie, physicians who used support), their results may not generalize to other physicians. As a result, further work is needed to understand whether it would be worthwhile for clinical leaders to offer team-based documentation support to physicians who would have been less likely to independently choose this option, and what it would take to ensure meaningful use of this support for these individuals.[6-8]

A final, more subtle consideration for readers relates to the use of physicians who chose not to adopt scribes as a comparison group. Beyond information about preintervention trends, such as that provided by event study plots, it can be difficult to provide specific evidence in favor of the parallel trends assumption. This requires explaining why groups with different outcome levels would have been expected to have similar outcome trajectories and why treatment and comparison groups were likely to respond similarly to shocks (defined as events that alter the trajectories of treatment and comparison groups). In this case, although the authors show that the preintervention trends were generally similar, known differences between physicians who do and do not choose to adopt scribes suggest that they might have reacted differ-

ently to such shocks (eg, the introduction of new EHR tools or other clinical workflows) postintervention, risking violation of the parallel trends assumption.

Overall, this study[1] applies DiD to highlight promising capacity for documentation support to reduce physician EHR usage time and allow for more visits. Even as more research is needed to understand the best ways to implement and evaluate such programs, it provides important insights about potential benefits of documentation support for physicians and health care systems.

ARTICLE INFORMATION

Author Affiliations: Brown University School of Public Health, Providence, Rhode Island (Bilinski); Harvard Medical School, Boston, Massachusetts (Ganguli); Associate Editor, *JAMA Internal Medicine* (Ganguli).

Corresponding Author: Alyssa Bilinski, PhD, MS, AM, 121 South Main St, 8th Floor, Providence, RI 02903 (alyssa_bilinski@brown.edu).

Published Online: August 26, 2024. doi:10.1001/jamainternmed.2024.4132

Conflict of Interest Disclosures: Dr Bilinski reported grants from the Council of State and Territorial Epidemiologists and the National Center for HIV, Viral Hepatitis, STD, and Tuberculosis Prevention outside the submitted work. Dr Ganguli reported consulting fees from FPrime and grants from the National Institute on Aging, the Commonwealth Fund, Arnold Ventures, National Institute on Minority Health and Health Disparities, and the Agency for Healthcare Research and Quality outside the submitted work. No other disclosures were reported.

REFERENCES

1. Apathy NC, Holmgren AJ, Cross DA. Physician EHR time and visit volume following adoption of team-based documentation support. *JAMA Intern Med*. Published online August 26, 2024. doi:10.1001/jamainternmed.2024.4123

2. Roth J, Sant'Anna PHC, Bilinski A, Poe J. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *J Econom*. 2023;235(2):2218-2244. doi:10.1016/j.jeconom.2023.03.008

3. Wing C, Dreyer M. Making sense of the difference-in-difference design. *JAMA Intern Med*. Published online August 26, 2024. doi:10.1001/jamainternmed.2024.4135

4. Callaway B, Sant'Anna PHC. Difference-in-differences with multiple time periods. *J Econom*. 2021;225(2):200-230. doi:10.1016/j.jeconom.2020.12.001

5. Sun L, Abraham S. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. arXiv. Preprint revised September 22, 2020. https://arxiv.org/abs/1804.05785

6. Ghatnekar S, Faletsky A, Nambudiri VE. Digital scribe utility and barriers to implementation in clinical practice: a scoping review. *Health Technol (Berl)*. 2021;11(4):803-809. doi:10.1007/s12553-021-00568-0

7. Florig ST, Corby S, Devara T, Weiskopf NG, Gold JA, Mohan V. Variable impact of medical scribes on physician electronic health record documentation practices: a quantitative analysis across a large, integrated health-system. *J Am Board Fam Med*. 2024;37(2):228-241. doi:10.3122/jabfm.2023.230211R2

8. Micek MA, Arndt B, Baltus JJ, et al. The effect of remote scribes on primary care physicians' wellness, EHR satisfaction, and EHR use. *Healthc (Amst)*. 2022;10(4):100663. doi:10.1016/j.hjdsi.2022.100663