COVID-19

Evidence-based COVID-19 policy-making in schools

New research can help policymakers make evidence-based decisions about the risks and benefits of in-person schooling; strategic use of the available data will be key to getting this right.

Elizabeth A. Stuart and David W. Dowdy

Policy questions such as how best to re-open schools for in-person learning during a pandemic are incredibly important — but also incredibly hard to answer in an evidence-based fashion. As with many other policy decisions related to COVID-19, strategies about the reopening of schools were initially crafted with minimal direct evidence. Recently, however, an increasing number of empirical studies are shaping an evidence base about the risks and benefits of in-person schooling, including two studies in this issue of *Nature Medicine*.

First, Ertem et al.[1] use data from the United States to examine the effects of different schooling models on COVID-19 case rates by comparing counties with in-person schooling to those with hybrid or virtual modes of education. In a similar vein, Fukumoto et al.[2] compare COVID-19 case rates in Japanese municipalities where schools opened with similar municipalities that kept schools closed. Neither study found a consistent relationship between school reopening and COVID-19 case rates — but findings varied across contexts, especially within the United States. As school policymakers must now weigh evidence from several studies (often with results that appear to conflict), it is important to keep three considerations in mind: the causal question being asked, the comparisons being made, and the context to which the findings pertain.

For policy decisions, we are almost always interested in a causal question — that is, one that compares outcomes (for example, case rates) under two different possible states of the world. In one state, a well-defined group experiences the intervention of interest (for example, school reopening); in the other, that same group experiences a comparator condition (such as continued virtual learning). This comparison immediately raises the 'fundamental problem of causal inference'[3] — that, for any given school at any point in time, we can only observe outcomes under one state (for example, school reopening), whereas the other state is unobserved, or 'counterfactual'. Thus, we are forced to use data from different groups — or in this case, different schools — to estimate what would be seen in the same group under the intervention state versus the comparator state; in other words, a 'causal contrast'. Appropriate causal inference therefore requires strong study designs such as randomization, longitudinal evaluation of communities with schools that did and did not reopen (as in Ertem et al.[1]), and/or well-selected comparison groups (as in Fukumoto et al.[2]). Robust designs allow for reasonable estimation of what would have happened in the communities with schools that re-opened, had they actually stayed closed.

Although others have appropriately highlighted the importance of study design in answering pandemic-related causal policy questions[4–6], we argue that policymakers should also 'keep it simple'. Specifically, most causally-focused studies can be evaluated in terms of their question, comparison and context. By asking whether these three components of a given study seem reasonable — and the degree to which they apply to a current decision — policymakers without extensive methodological expertise can make a rapid assessment as to the relevance of a particular study.

For example, consider a county school board evaluating the results of the study by Ertem et al.[1] to decide whether to restrict in-person learning in the face of a new pandemic wave. This study relates to county-level decisions on opening or closing in-person learning — not, for example, decisions at the state or national level. In addition, the specific comparisons made in

Credit: Prostock-studio/Alamy Stock Photo

this study would only be directly applicable to an 'all-or-nothing' closure decision, as few counties adopted an approach of keeping elementary schools open but middle and high schools closed. In terms of context, different results were seen in the South than in other regions — and results in the United States might not necessarily be generalizable to other countries. But by focusing on the question, comparison and context, non-expert decision-makers could reasonably assess the relevance of this study to their policy decision. We should encourage this sort of thinking — and make it more accessible by highlighting these three elements in any analysis that seeks to estimate a policy-relevant causal effect.

Sadly, there is often a disconnect between the questions, comparisons and contexts addressed in research studies, and those that policymakers must consider. Regarding the question being asked, Ertem et al.[1] and Fukumoto et al.[2] both consider area-level policy decisions; other studies of in-person schooling have focused on the behaviors of individual households[7]. Some studies have compared 'school reopening' with 'school closure' overall, whereas others sought to estimate the effects of specific mitigation strategies. But these may not be the questions that local policymakers need to answer; even randomized trials in schools are not always immediately relevant for local decision making if the study population is too different from the population of policy interest[8] or the strategies being studied differ substantially from the policy options on the table. For example, a recent study compared daily testing with isolation for close contacts of individuals with COVID-19[9] — but many school systems might be interested in less frequent testing, or different strategies for children versus staff members.

Analyses can, and should, evaluate differences in estimated effects across contexts, but these explorations are often limited by the available data. Although Ertem et al.[1] highlight interesting variation in the estimated effects of school closures across US regions, they also note an inability to accurately pinpoint explanations for this variation, which might include inconsistent mitigation strategies, weather-related factors, or differences in underlying community transmission rates of SARS-CoV-2. It is also worth noting that research studies use retrospective data, whereas policy decisions must be made in the present. Together, these challenges highlight the importance of performing research that is as close as possible in question, comparison and context to actual policy decisions that are being considered. If these diverge substantially, policymakers will default to decision-making in the absence of evidence, thus invalidating the considerable efforts to bring an evidence base to bear in this process. Rarely are blanket conclusions — for example, that reopening schools does not fuel SARS-CoV-2 transmission — appropriate.

It is therefore crucial that, in informing evidence-based decision-making, researchers clearly state the causal questions, comparisons and contexts — while making use of the most appropriate data and study designs available. In the social sciences, the UTOSTi (units, treatment, outcomes, settings and times) framework has helped to articulate some of these considerations[10]; we now need a similarly simple guide for scientists and decision-makers asking policy-relevant questions about the COVID-19 pandemic. No single study will be relevant to all policy questions; therefore, we must urgently build a diverse evidence base that mirrors those most likely to be encountered — and then communicate those results to decision-makers in real-time, using language that can be broadly understood. ❐

Elizabeth A. Stuart ✉ and David W. Dowdy
Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MA, USA.
✉*e-mail: estuart@jhu.edu*

Published online: 25 November 2021
https://doi.org/10.1038/s41591-021-01585-2

References

1. Ertem, Z. et al. *Nature Med*. https://doi.org/10.1038/s41591-021-01543-y (2021).
2. Fukumoto, K., McClean, C. T. & Nakagawa, K. *Nature Med*. https://doi.org/10.1038/s41591-021-01571-8 (2021).
3. Holland, P. W. *J. Am. Stat. Assoc.* **81**, 945–960 (1986).
4. Bonvini, M., Kennedy, E., Ventura, V. & Wasserman, L. Preprint at https://arxiv.org/abs/2103.04472 (2021).
5. Goodman-Bacon, A. & Marcus, J. *Survey Res. Methods* **14**, 153–158 (2020).
6. Haber, N., Clarke-Deeder, E., Salomon, J., Feller, A. & Stuart, E. A. *Am. J. Epidemiol*. (in the press).
7. Lessler, J. et al. *Science* **372**, 1092–1097 (2021).
8. Orr, L. L. et al. *J. Policy Anal. Manage.* **38**, 978–1003 (2019).
9. Young, B. C. et al. *The Lancet* **398**, 1217–1229 (2021).
10. Cook, T. D. *J. Policy Anal. Manage.* **33**, 527–536 (2014).

Acknowledgements

D.W.D. was supported by a Hopkins Business of Health Initiative pilot grant and a Johns Hopkins University Catalyst Award. E.A.S. was supported by a Johns Hopkins University Discovery Award and by National Institutes of Health award P50MH115842 (PI: Daumit).

Author contributions

Both authors contributed to the development of the ideas, wrote the text, and edited the final version.

Competing interests

The authors declare no competing interests.

MACHINE LEARNING

Check for updates

Rising to the challenge of bias in health care AI

AI-based models may amplify pre-existing human bias within datasets; addressing this problem will require a fundamental realignment of the culture of software development.

Mildred K. Cho

In artificial intelligence (AI)-based predictive models, bias—defined as unfair systematic error—is a growing source of concern, particularly in healthcare applications. Especially problematic is the unfairness that arises from unequal distribution of error among groups that are vulnerable to harm, historically subject to discrimination or socially marginalized.

In this issue of *Nature Medicine*, Seyyed-Kalantari and colleagues[1] examine three large, publicly available radiology datasets to demonstrate a specific type of bias in AI-based chest X-ray prediction models. They found that these models are more likely to falsely predict that patients are healthy if they are members of underserved populations, even when using classifiers