How much can you trust the results of this health policy evaluation:

A pragmatic guide for state policymakers

Megan S. Schuler[1], Elizabeth A. Stuart[2], Beth Ann Griffin[1], Rosanna Smart[1], Rosalie Liccardo Pacula[3], David Powell[1], Stephen Patrick[4], Shelby M. Hockenberry[5], Bradley D. Stein[1]

Policymakers, regulators, and other stakeholders need reliable evidence about which state policies are truly effective to help them make informed decisions that affect people's health. Published scientific studies that evaluate the impact of state policies can answer important questions about policy effectiveness. For example: "Did state-level Medicaid expansion lead to more people getting substance use treatment?" and "Do naloxone access laws help to reduce fatal opioid overdoses?"

However, accurately determining the extent to which specific outcomes changed because of a policy is fundamentally challenging. Not all policy evaluation studies are equally rigorous—for example, some have major flaws in their analytic approach or limitations due to the data used. Findings from rigorous, high-quality evaluation studies are more trustworthy and provide a more reliable foundation for decisionmaking.

This Perspective is intended to help policymakers and other stakeholders assess policy evaluation studies, with a focus on health policy studies at the state level. Our focus is on state policies because much of the public health-related policy action occurs at the state level (e.g., policy responses to the opioid epidemic or COVID-19 pandemic) and note that studies evaluating policies at different levels (e.g., federal, local) may involve some distinct analytic considerations. We offer a few rules of thumb to identify policy evaluation studies that provide stronger evidence about whether there is a *cause and effect* relationship between a policy and an outcome. We highlight aspects of a study's design and data measures to help assess how much to trust study findings regarding policy effectiveness. We also highlight some *red flags* indicating that a study's findings should be viewed with caution.

1: RAND Corporation
2: Johns Hopkins Bloomberg School of Public Health
3: USC Price School of Public Policy and Schaeffer Center for Health Economics & Policy
4: Vanderbilt University Medical Center
5: National Governors Association

Key Questions about Study Design of State Policy Evaluation Studies

We focus on evaluation studies that examine if a given state policy (e.g., Medicaid expansion; naloxone access laws) systematically impacts a given outcome (e.g., improved access to treatment; fatal opioid overdoses). Conceptually, the central questions are (1) whether the outcome changed from before to after the policy was implemented and (2) whether this change was due to the policy of interest.

In order to determine whether policy implementation resulted in changed outcomes, an evaluation study would *ideally* compare the outcomes a state(s) experienced during a period when the policy was in place to the outcomes that the *same* state(s) would experience during the *same* period but without the policy in place. This difference represents the true *causal effect* of the policy. Obviously, this is impossible: states cannot be both exposed and not exposed to a policy during the same period. As such, different study designs use various analytic approaches to estimate the (fundamentally) unobservable result of what would have happened to outcomes in states with the policy had they not had the policy.

For example, policy evaluation studies often use outcomes observed in a *comparison group* of states that were not exposed to the policy to infer *what would have happened* if states that did enact the policy had not done so. Not all evaluation studies use a comparison group, however. Some policy evaluation studies only examine states that enacted the policy and strictly use the pre-policy period as a point of comparison. Such studies assume that the outcome trend observed among these states during the pre-policy period would have continued in the same manner in the post-policy period *had the policy not been enacted*. While certain study design choices (e.g., comparison group versus no comparison group) can provide stronger evidence of a policy effect than others, the overall strength of evidence a study provides reflects the totality of analytic choices made. As such, as detailed by this Perspective, it is important to comprehensively assess multiple aspects of policy evaluation studies.

Below we review some key characteristics of interest regarding state policy evaluation studies, including: What is the policy of interest? What is the outcome(s) of interest? How are the pre-policy and post-policy time periods defined? Does the study design use a comparison group of states that were not exposed to the policy? How does the study adjust for co-occurring factors and policies that may have impacted the outcome? How consistent is the evidence of a policy effect across studies? We discuss these questions as they relate to two different aspects of a policy evaluation study: (1) calculating the appropriate numerical change in outcomes and (2) asserting that this change is actually due to the policy.

I. Determining whether the outcome changed after the policy was implemented: Analytic considerations

Does the study examine meaningful outcomes among a relevant population?

Given that policy evaluation studies seek to assess policy-related change in the outcome, analytic choices related to the outcome(s) are key. Studies vary widely in terms of what outcomes are being measured, how the outcomes are measured, and the populations among whom outcomes are measured. To determine a study's relevance and strength of evidence, it is important to assess the *why, how,* and *who* dimensions of the study.

- *The "why."* An essential question is why a study chose the outcome(s) it examined, and whether these outcomes are relevant to policy decisionmaking. High-quality studies evaluate meaningful outcomes based on a presumed link between the policy and the outcome—e.g., the hypothesis that policies to curb over-prescribing of opioids will reduce opioid misuse. The outcomes being evaluated may be desired ones (e.g., less opioid prescribing) or an *unintended consequence* (e.g., increased use of heroin due to restricted access to prescription opioids). Whether examining an intended or unintended consequence, stronger studies explain why they evaluate outcomes in relation to the policy of interest.
- *The "how."* It is important to understand how outcomes are measured, because not all data are equally informative. The strongest studies will use objective outcome data that are measured the same way across all states at all time points. Collecting new data is costly and time consuming, so researchers often rely on data routinely collected by state agencies or other organizations, such as insurance billing data. While these data are generally standardized and consistently collected within a given agency, there may be important variation in how data are collected across states. For example, when reporting emergency department (ED) use, some states include all ED visits

whereas other states exclude ED visits that result in hospital admission (only reporting these as part of hospital admissions data). If these data are combined in an analysis, differences across states could be erroneously interpreted as meaningful, whereas they simply reflect differences in data collection. As another example, the number of overdoses treated in emergency departments does not capture all overdoses because not everyone who overdoses is taken to a hospital. Relying only on ED data on overdoses would underestimate the true magnitude of all overdoses. Overall, it is essential to consider the advantages and limitations of how the data used to measure outcomes were collected.

- *The "who."* It is important to determine for whom outcomes are assessed. Studies that evaluate outcomes among the full population the policy is intended to affect provide the strongest, most generalizable evidence. Notably, studies focusing on population subgroups are critically important to assessing health disparities: It is essential to understand who does not benefit (or may even be harmed) by a given policy. If population subgroups are examined, rigorous studies will clearly explain why they selected a particular population subgroup (e.g., adults over 65). Be wary of studies that, after finding no overall policy effect, conduct many subgroup analyses without clear justification, which is known as "fishing" for an effect. Conducting numerous statistical tests raises the chance of a false positive finding (incorrectly concluding there's a policy effect when there's no true effect). Finally, an important caveat regarding subgroup analyses: evidence that a policy is effective (or harmful) in a certain subgroup does not mean that the policy would have the same effect for other population subgroups.

Does the study examine meaningful outcomes among a relevant population?

Key Questions:

Does the study clearly explain how the policy is likely to affect the outcome(s) being evaluated?

Are outcome(s) measured consistently across all states and over time?

Does the study evaluate outcome(s) for the total population that the policy is likely to affect? If the study focuses on a population subgroup, does the study clearly explain why?

Red flags:

Studies that examine many subgroups without clear rationale may be at risk of identifying "false positive" effects of the policy, simply due to chance.

Does the study evaluate the policy over a meaningful time period?

The timing of policy implementation relative to when the outcome is measured is one factor determining whether a study can provide evidence of a causal effect. Demonstrating cause-effect ordering is key to establishing causality—that is, a true "cause" must occur prior in time to the "outcome." In the context of state policy evaluation, it is essential that pre-policy outcome(s) are measured strictly *before* policy implementation and post-policy outcome(s) are measured strictly *after* implementation to ensure cause-effect ordering.

Suppose the policy of interest went into effect in June 2021 and an evaluation study assessed the policy's impact on rates of opioid-exposed newborns by comparing the annual rates for 2020 and 2021. While the 2020 annual rate is measured entirely prior to policy implementation, the 2021 annual rate spans both before and after policy implementation. If rates have declined in 2021, this does not provide clear evidence of causality because changes in the 2021 annual rate may have occurred partially or entirely before policy implementation. A more appropriate study design would compare annual rates for 2020 (strictly before the policy) and 2022 (strictly after the policy).

Additionally, the post-policy period should be long enough for policy effects to appear, as policy effects are rarely instantaneous. Note that the size of the policy effect may change over time (for example, due to increased public awareness or gains in implementation efficiency). Studies that examine policy effects over a longer period after implementation and assess whether the policy effect changes over time can provide more information regarding the sustained impact of a policy.

For example, suppose that a study was investigating how state Medicaid expansion affected treatment for opioid use disorder. In terms of pre-policy data, repeated measurements of treatment rates spanning multiple years before Medicaid expansion are needed to clearly characterize treatment trends that existed before the policy was implemented. However, there's a balancing act—pre-policy periods that go too far back in time may not be comparable due to historical trends (e.g., an economic recession that impacted treatment utilization). Similarly, repeated measurement of treatment rates across multiple years after expansion ensures that post-policy trends can be rigorously characterized. If the policy had an increasing impact on treatment access over the first three years after implementation, studies that only considered the first one or two years would not accurately characterize the overall impact of the policy.

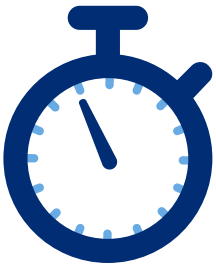
Finally, when determining which time periods to compare outcomes across, it is essential that evaluation studies are informed by real-world understanding of state policy actions. The date a policy is passed (*enactment date*) is often not the same date it goes into effect (*implementation date*), and policy enforcement may start at an even later date. States may experience unplanned delays in policy implementation or enforcement (e.g., due to lack of funding) that need to be considered when defining an appropriate study period. Rigorous studies will define pre-policy and post-policy periods based on accurate data regarding implementation and/or enforcement.

Does the study assess the policy over a meaningful time period?

Key Questions:

Timing matters. Is the study period long enough for the policy to have a detectable effect on outcomes?

Does the study define the policy "start date" and identify the data source for policy dates?



Red flags:

Some policies may take time to "ramp up" to full effectiveness.

Studies may assess effectiveness over a period that is too short to detect the policy's full effect.

Do states in the policy group truly have similar policies?

Some evaluation studies examine the impact of a policy in a single state—often these studies are evaluations of an early-adopter state. However, many studies examine the impact of a policy across multiple states. When studies combine multiple states together as a group all with the policy of interest, it is imperative that each state's policies are indeed very similar. High quality studies clearly define how their policies of interest are defined and describe their data source for determining policy details and dates. This is essential because policies that have the same name may be quite different in practice.

As an example, consider prescription drug monitoring programs (PDMPs). PDMPs are state-wide databases that track prescriptions of controlled substances, including opioids. All 50 states and the District of Columbia now have implemented a PDMP, but the actual operation of PDMPs varies substantially across states. For example, some states require that clinicians check patient records in the PDMP before prescribing (*provider mandatory access*); other states only require that pharmacists access the PDMP before dispensing (*pharmacist mandatory access*). Some states include only schedule II and III drugs in the PDMP, while other states include all scheduled drugs. Overall, there is a lot of variation across states in what comprises a PDMP.

When policies vary significantly across states in ways that may impact outcomes differently, different analytic approaches can provide more helpful evidence. For example, studies that classify states with respect to specific policy components (e.g., mandatory access PDMP versus non-mandatory access PDMP) can provide insights about which components are most important for changing outcomes. However, there are tradeoffs to consider. In general, studies that have a larger number of states in the policy group(s) provide stronger, more generalizable evidence of a policy effect. Creating too many policy groups, and thus picking up smaller differences in policies across states, means that there will be fewer states in each group. As a result, the study will have weaker statistical ability to identify effects of each policy.

Do states in the policy group truly have similar policies?

Key Questions:

Policies that have the same name may be quite different in practice. Does the study give specific details of the policy of interest, noting any variations across states?

If policy details vary substantially across states, does the study examine or account for variations in policies?

Red flags:

Studies that do not discuss or address potential policy variability across states.

II. Determining whether a change in outcomes was due to the policy of interest: Analytic considerations

Again, an evaluation study would *ideally* compare the outcomes a state(s) experienced during a period when the policy was in place to the outcomes that the *same* state(s) would experience during the *same* period but without the policy. Since the latter is fundamentally unobservable, statistical analyses use different methods to estimate this unobservable result of what would have happened to outcomes in states with the policy *had they not had the policy*. Each analytic approach requires its own set of assumptions regarding under what conditions the method will be able to provide a reliable estimate of a policy effect. High quality studies will clearly state these underlying assumptions and describe the extent to which they are believable in their study. Many of these assumptions are related to whether the study can disentangle the effect of the policy from that of other factors that may have also impacted the outcome. As discussed below, many studies tackle this challenge by using a comparison group of states that were not exposed to the policy.

Does the study account for other co-occurring factors and policies that may have impacted the outcome?

Just because a study shows that outcomes changed before and after policy implementation does not mean that the policy *caused* these changes. A key challenge is determining how much of the change in the outcome was due to the policy of interest as opposed to other co-occurring factors (such as economic events, sociodemographic shifts, or other policy or regulatory efforts). Studies may do this by measuring and controlling for these factors directly (e.g., as variables in a regression analysis) or through the choice of study design (e.g., by selecting a comparison group that experienced very similar co-occurring factors). For example, some analytic methods estimate the policy effect by "subtracting out" the extent to which the outcome changed in the comparison group (due to factors other than the policy) from the amount the outcome changed in the policy states (due to both the policy and other factors). In theory, the comparison group provides an estimate of how much the outcome changed in the policy states due to co-occurring factors; however, the assumption is that the comparison group experienced the *same exact* co-occurring factors that had the *same exact* impact on the outcome as did the policy states. Depending on the context, this may or may not be a reasonable assumption.

In particular, it is important to remember that policies are not enacted in a vacuum. In each state, the policy of interest commonly occurs alongside other policies that may potentially affect the outcome being studied. It is essential that evaluation studies differentiate between the effect of the policy of interest and that of co-occurring policies.

Consider a study evaluating the effect of naloxone laws on opioid overdose deaths. In addition to naloxone laws, some states may also have policies in place to expand access to treatment and/or recovery support groups, which may also affect opioid overdose rates. A rigorous policy study will analytically control for co-occurring policies, such as treatment access laws, so as to estimate the impact that *only* naloxone laws have on opioid overdose deaths.

Disentangling the effects of a policy from co-occurring state policies is very challenging when states enact multiple policies in a very short period, as seen in response to economic or public health crises (such as the COVID-19 pandemic).

Overall, rigorous studies will include a discussion of other co-occurring factors and policies that may have impacted the outcome and will clearly describe how they analytically disentangled the effects of the policy of interest from these other factors.

Does the study account for other co-occurring factors and policies that may have impacted the outcome?

Key Questions:

Does the study account for other factors (such as economic events or sociodemographic shifts) that could also affect the outcomes?

Does the study account for other co-occurring policies that could also affect the outcomes?

Red flags:

The study overlooks important factors that occurred during the study period and are likely to impact the outcome.

The study is evaluating a policy that was implemented simultaneously with other policies.

If the study has a comparison group, does it make an "apples to apples" comparison between states with and without the policy?

Many policy evaluation studies use outcomes observed in a *comparison group* of states that were not exposed to the policy in order to estimate *what would have happened* if states that did enact the policy had not done so. Not all evaluation studies use a comparison group; in the absence of a comparison group, studies generally assume that the outcome trend observed among the policy states would have continued in the same manner had the policy not been enacted. Studies without a comparison group generally provide weaker evidence regarding the extent to which the outcome would have changed due to other factors in the absence of the policy. This is particularly true in the case of disruptive events that alter historical trends—for example, the flood of fentanyl into the opioid market.

Consider a study examining the impact of a 2015 policy regulating opioid prescribing on opioid overdose rates. From 2015 onward, opioid mortality rates have dramatically increased due to the rise of fentanyl. A study without a comparison group would simply project forward the trend of opioid overdose rates prior to 2015, which were growing at a much slower rate. As such, this study design would not provide an accurate estimate of what opioid mortality rates would have looked like in the policy states in the absence of the policy. Alternatively, a study design with a contemporaneous comparison group would provide a much more accurate estimate of what opioid overdose rates would be in the absence of the policy, yet in the era of rising fentanyl.

For studies with a comparison group, the nature of the comparison group affects a study's quality of evidence. Notably, the process by which certain states choose to implement policies, and others do not, is not random. For example, certain states may implement opioid-related policies because they are experiencing elevated overdose deaths, which is spurring them to policy action, and/or because they generally have more financial resources or take a more proactive approach to public health. This *self-selection process* presents a central challenge because it means that policy and comparison states may differ substantially, setting up an "apples to oranges" rather than "apples to apples" comparison.

Fundamentally, studies that use a carefully selected comparison group and appropriate statistical methods to improve the similarity of the policy and comparison groups provide stronger evidence of policy effectiveness. Statistical techniques can be used to make the policy and comparison groups similar, minimizing concerns about self-selection. For example, studies may do this through their choice of study design (some of which use weighting or matching methods to make the policy and comparison groups more similar) or by measuring and controlling for the factors that policy and comparison groups differ on (e.g., as variables in a regression analysis). Sometimes studies simply choose neighboring states as the comparison group; however, geographic proximity alone does not ensure comparability.

Two key types of differences that studies may adjust for are: (1) different outcome trends in the policy and comparison groups during the pre-policy period and (2) different state-level characteristics. The ability of studies to rigorously adjust for these differences depends in part on the quality of the data. The first type of difference requires multiple repeated measures of the outcome during the pre-policy period; the second type of difference requires data on time-invariant or pre-policy state characteristics (including other relevant state policies) that may also impact the outcome.

If the study has a comparison group, does it make an "apples to apples" comparison between states with and without the policy?

Key Questions:

Does the study use a comparison group (not exposed to the policy) to help distinguish the true policy effect from changes due to other factors (e.g., economic factors, social trends, other policy or regulatory efforts, etc.)?

Does the study clearly describe the analytical method used to ensure that the policy and comparison groups are similar?

Red flags:

Evaluation studies that do not have a comparison group provide weaker evidence of policy effectiveness than studies using a comparison group, which helps to isolate the policy effect from the impacts of other factors.

How consistent is the evidence of a policy effect?

Several fundamental factors can strengthen the consistency of study findings.

1. *Evaluating policy effects based on a greater number of states that have implemented the policy can provide stronger, more consistent evidence.* Depending on the timing of the study, the policy of interest may have been implemented in a very small number of states or widely implemented. When only one state (or very few states) has implemented a novel policy, evaluation studies can only assess outcomes in a single state (or very few states). In this case, it is hard to know if the policy would also be effective in other states or if the observed effect was dependent on factors unique to state(s) that were early adopters of the policy. Studies that assess the policy effectiveness across multiple states that have implemented the policy can provide stronger evidence. In this case, findings of a policy effect suggests that policy effectiveness is indeed generalizable across states, providing a more solid foundation for policy decisionmaking.

2. *Evaluating policy effects on a greater number of outcomes can provide stronger, more consistent evidence*. If a study examines only a single outcome, questions may remain as to whether the observed effect was truly due to the policy or if the outcome simply exhibited a coincidental change. Examining a single outcome also does not provide evidence about how the policy may have affected other outcomes, including possible unintended consequences. A study that evaluates multiple, well-reasoned outcomes and finds evidence that the policy affected more than one outcome provides stronger evidence of a true policy effect. For example, a study finding that regulating "pill mills" led to reductions in the total number of opioid prescriptions, the duration of opioid prescriptions, and opioid overdoses provides stronger evidence about the effectiveness of pill mill regulations than a study examining only the total number of opioid prescriptions.

3. *Conclusions from a well-designed study are further strengthened if sensitivity analyses demonstrate that changing certain aspects of the study does not fundamentally alter the findings.* In some cases, multiple study designs and analytic approaches would be appropriate. Sensitivity analyses are secondary analyses that investigate whether study findings are unique to the specific analytic approach used in the study or whether similar findings are observed when the data are analyzed somewhat differently. For example, sensitivity analyses may look at the impact of using a different kind of statistical analysis, a different data source to measure the outcome, or a shorter or longer time frame. If the main analyses presented in a study are appropriate and rigorous, sensitivity analyses

indicating that findings hold across different, appropriate analytic approaches should bolster confidence in the study's main analyses.

- *Results across a greater number of evaluation studies can provide stronger, more consistent evidence, particularly if studies evaluate independent datasets.* Results from a single study often do not provide sufficient evidence to inform policy decisions. Multiple studies of a given policy showing consistent results regarding the same or related outcomes provide stronger evidence than a single study. Independent replication of findings across different studies conducted by different research teams is a hallmark of good science.

- *Sometimes multiple studies of the same policy may have different findings regarding similar outcomes.* Assuming that these studies are well-designed and directly comparable, mixed findings across studies could indicate that the overall evidence of the policy's effects is still ambiguous and does not provide a solid foundation for policymaking.

How consistent is the evidence of a policy effect?

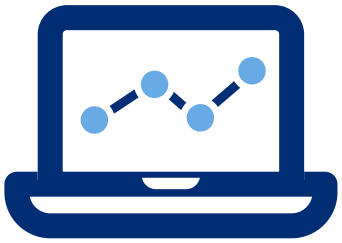
Key Questions:

Is there evidence of a policy effect across multiple states, rather than simply in one state?

Is there evidence of a policy effect across multiple outcomes, rather than a single outcome?

Does the study conduct a sensitivity analysis using an alternative (yet appropriate) approach to determine whether changing some aspect of the study could change the overall findings?

If study findings are very different from prior studies, do the authors explain what factors might explain their unique findings?



Conclusion

Overall, policy evaluation studies vary in terms of their study design, data sources, and rigor of analytic approach. Findings from rigorous, high-quality evaluation studies are more credible and provide a sounder basis for decisionmaking. This Perspective highlights key questions that can help assess how much to trust study findings as well as some red flags indicating that a study's findings should be viewed with caution.

Acknowledgements

This work was supported by the National Institute on Drug Abuse (OPTIC; P50DA046351). We are grateful to our colleagues at the National Governors Association Center for Best Practices and the National Conference of State Legislatures for their contributions towards the development of the pragmatic guide for policymakers.

https://americanhealth.jhu.edu/news/evidence4policy