Temple University
Beasley School of Law

Center for
Public Health Law
Research

NATURAL EXPERIMENTS: RESEARCH DESIGN ELEMENTS FOR OPTIMAL CAUSAL INFERENCE WITHOUT RANDOMIZATION

Alexander C. Wagenaar, PhD

Research Professor, Emory University Rollins School of Public Health; Professor Emeritus, University of Florida College of Medicine

Kelli A. Komro, PhD, MPH

Professor, Rollins School of Public Health, Emory University

A Methods Monograph for the Center for Public Health Law Research Temple University Beasley School of Law

OCTOBER 2023

NATURAL EXPERIMENTS: RESEARCH DESIGN ELEMENTS FOR OPTIMAL CAUSAL INFERENCE WITHOUT RANDOMIZATION

Alexander C. Wagenaar

Kelli A. Komro

Summary

Most changes in laws and regulations affecting population health represent natural experiments, in which scientists do not control when and where these changes are implemented and thus cannot randomly assign the legal "treatments" to some and not to others. Many research design elements can be incorporated in evaluations of public health laws to produce accurate estimates of the size of a law's effect with high levels of confidence that an observed effect is caused by the law:

- Incorporate hundreds of repeated observations before and after a law takes effect, creating a time series.
- Measure outcomes at an appropriate time resolution to enable examination of the expected pattern of effects over time that is based on a theory of the mechanisms of legal effect.
- Include comparisons in the design, such as multiple jurisdictions with and without the law under study, constructed synthetic comparison groups, comparison groups within a jurisdiction of those exposed and not exposed to the law, and comparison outcomes expected to be affected by the law and similar outcomes not expected to be affected by the law.
- Replicate the study in additional jurisdictions implementing similar laws.
- Examine whether the "dose" of the law across jurisdictions or across time is systematically related to the size of the effect.

Combining design elements produces the strongest possible evidence on whether a law caused the hypothesized effect and magnitude of that effect. Well-designed studies of public health laws in

natural real-world settings facilitate diffusion of effective regulatory strategies, producing significant reductions in population burdens of disease, injury, and death.

Learning Objectives

- Understand advantages of time-series data, with many repeated observations before and after a change in law, for evaluating the law's effects.
- Create a nested multiple comparison group study design for evaluating health effects of a law.
- Combine several design elements in a single study to strengthen causal inference.
- Identify resources for further study of legal epidemiology methods.

Evaluating the health effects of a law or regulation, or any treatment or intervention, most fundamentally requires a comparison of the experience with the law to the experience when everything is the same but without the law. Imagine the pure counterfactual, which involves the same people at the same time in the same place experiencing a law, compared to the same people at the same time and same place not experiencing the law (Rubin, 1974). The counterfactual requires the same people at the same time and place in the two conditions – with and without a specific law – to ensure that everything is identical between the two conditions, except the specific law. If everything but the law is identical, the difference in health outcomes of interest then directly represents the effect of the law. But such a comparison is impossible since the same people at the same time cannot experience both conditions. Thus the fundamental quandary of scientific research – how do we know that the difference in outcomes observed is really caused by the law, since the difference might be due to something else and not be a true effect of the specific law under study?

Random assignment was a major advancement in creating the counterfactual (Fisher, 1935). Relying on the law of large numbers, randomly selecting sets of people from the whole population, randomly selecting times of intervention implementation, and randomly selecting from the set of all places or settings creates groups of people, times, and settings that *on average* are expected to be equivalent in every way but for the law or intervention we exposed one set to but not the other set. Thus, any single experiment might be wrong, because the treated and untreated groups might simply, by chance, differ in some unknown way and that difference might be the true cause of an observed difference in outcome. But, on average, over many replications of the randomized experiment, the two sets of people, times, or settings compared are expected to be the same, and any difference in outcome can be confidently attributed to the effects of the one planned difference between the two conditions – one is exposed to the law under study and the other is not.

Despite its appeal, randomly exposing treatment groups and control groups is rarely possible when evaluating most new laws and regulations. Most laws are implemented at particular times

and in particular settings and, obviously, passage and implementation are not under the control of researchers. They are therefore commonly called *natural experiments*. When laws are changed, they almost always of necessity apply to everyone in the given jurisdiction. Characteristically, there are few units in the study – for example, one or a few cities or states pass an innovative law, and the entire population within the unit is exposed to the new law all at once. In short, randomization is rarely available as a strategy or *design element* to improve the likelihood of correctly assessing a law or regulation's effects.

There is an unfortunate tendency by many scientists and others to dichotomize studies into strong "experimental" studies (that use random assignment to treatment and control groups) that are assumed to provide clear evidence regarding the effects of an intervention, and weak "observational" studies (not using random assignment) that are assumed to provide ambiguous and often inaccurate evidence of effects (Benson & Hartz, 2000; Concato, Shah, & Horwitz, 2000; Guyatt, DiCenso, Farewell, Willan, & Griffith, 2000). This is a false dichotomy. Random assignment is only one of a dozen or more design elements that increase confidence in a causal interpretation of an observed difference (Bärnighausen, Røttingen, Rockers, Shemilt, & Tugwell, 2017; Bärnighausen, Tugwell, Røttingen et al., 2017; Leatherdale, 2019; Shadish, Cook, & Campbell, 2002). When evaluating the effects of local, state, or national laws and regulations, under which random assignment is rarely feasible, careful attention to full use of many other design elements is warranted. Moreover, effectively combining many design elements into a single study can produce real-world legal evaluations with higher overall levels of validity and strength of causal inference than randomized trials, which are typically limited to special circumstances or artificial environments. The objective of this chapter is to review design elements of particular importance when evaluating laws and regulations that naturally occur in the field and improve the quality of empirical studies of public health law by illustrating their use. In this chapter, we assume as a prerequisite the data on the laws under study have been carefully collected and coded using reliable and valid methods, as described in chapters 11 and 12.

Design Elements for Strong Legal Evaluations

There are several design elements for strengthening causal inference of particular importance when random assignment to treatment conditions is not possible.

MANY REPEATED MEASURES

A fundamental criterion for inferring whether a given law or regulation caused a change in outcomes is that the cause preceded the effect. For this reason, we measure the outcome before the law is implemented and again after. But having just one observation before and one observation after produces weak inference, because any difference observed might simply reflect the natural variation in the outcome over time. Figure 14.1 illustrates a situation in which a simple before-and-after design shows a major effect of the law, but that effect is no longer considered real when seen

in the context of more observations both backward and forward over time further away from the effective date of the law.Law

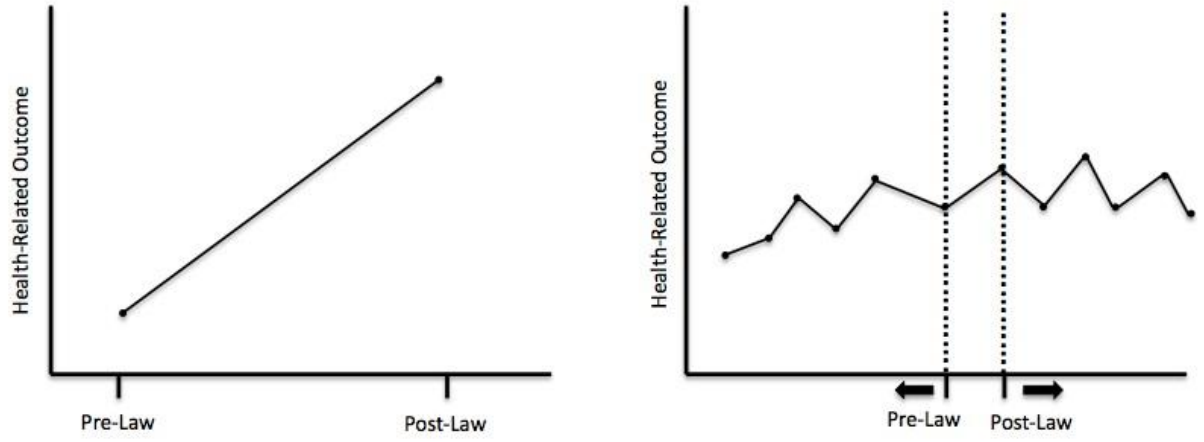

Figure 14.1. Observed Effect: Simple Pre-Post Design Versus Time-Series Design.

Collecting dozens or hundreds of observations in a *time series* before and after a new law takes effect makes it easier to see whether changes in the outcome of interest right around the time of the new law are larger than typical variation over time, and enhances confidence an observed difference occurring just at the time a new policy legally takes effect is due to that law. Any time series of observations can be viewed as a single sample (one window) from a time series that runs infinitely back in time and infinitely forward in time. The larger the time window observed around the time of a change in law, the easier it is to reliably assess that law's effects.

Beyond collecting many repeated measures, one must choose an appropriate *time resolution* for the observations. Are the observations a measure every minute, day, week, month, or year? Selecting the optimal time resolution is a complex tradeoff of multiple considerations. First is the speed by which a new law is expected to show effects. If the effects are expected to show up within weeks of the law's effective date, using weekly or monthly observations will make that effect easier to discern than using annual observations (Figure 14.2).

A second consideration when selecting the best time resolution to measure is the variation in the outcome over time at each time resolution. If there is little to no variation week by week in an outcome a new state law is meant to improve – say, math ability of teens – then monthly or even annual measures might be more appropriate. Consideration of the variation in the outcome over time interacts with a third important dimension, whether the underlying phenomenon being measured is *continuous*, or a *count*. For example, math ability, air pollution levels, water quality – like the temperature – all are continuous. The outcome is always there, we just choose intervals when we check the level. For continuous outcomes, the most important basis on which to choose

the time resolution of the measures is theory regarding the mechanism of a law's effect – when is the law first expected to show a difference in the outcome, and when (that is, at what interval) are further improvements expected?

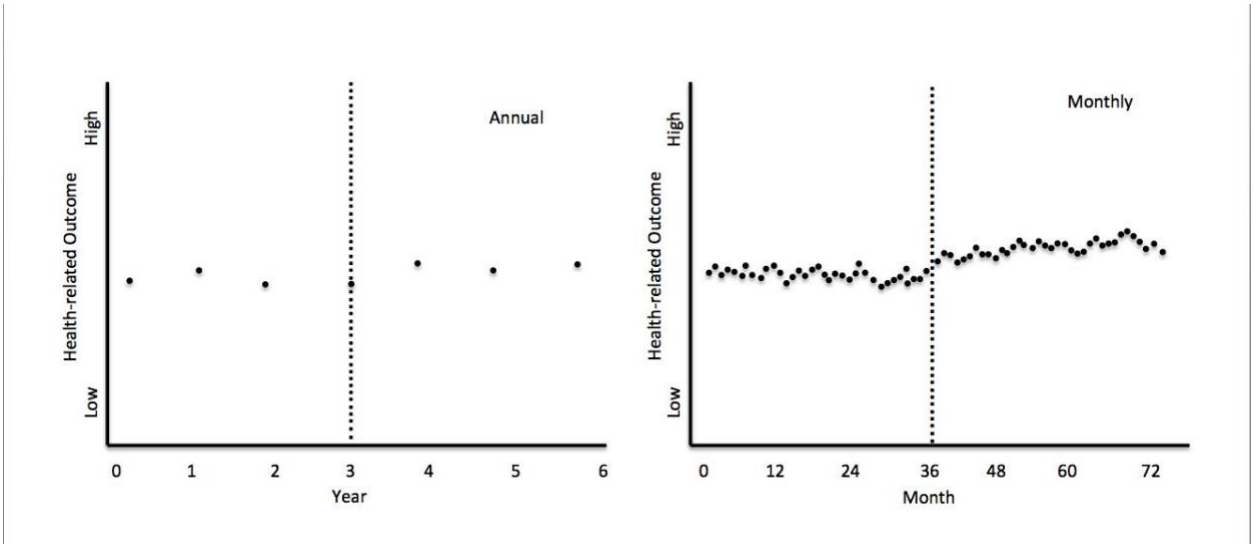

Figure 14.2. Observed Effect: Annual Versus Monthly Measures.

Many public health outcomes are not continuous, but are counts or frequencies of new infections or disease cases, counts of injuries, or counts of deaths. For count outcomes, the time resolution must roughly match the frequency of the event. If there is only, on average, one or two infections, injuries, or deaths per month in the geographic unit under study, choosing a daily or weekly time resolution is not appropriate since it will not help discern a law's effect on that outcome. Conversely, for example, if there are 50 or a 100 car crash deaths per month, lumping those data up to the yearly level for evaluating a new law's effects impairs the ability to accurately measure the law's effects. At the extreme, the problem of low counts expresses itself as numerous observations that are all zeros. Anything more than a very small fraction of zero-count observations complicates statistical analyses and makes discerning policy effects difficult or impossible. Thus, when the study design is being finalized, one must be aware of the expected outcome frequencies, and if numerous zero counts are expected at the preferred time resolution, the typical practical solution is moving to the next lowest resolution (for example, moving from monthly to quarterly counts).

Selecting the best time resolution for count data presents a tension between the desire for high-time-resolution observations and the resulting time series being "well-behaved," that is, exhibiting smooth regularities, cycles, or trends and not dominated by random unpredictability. In any study, minimizing the random, unpredictable variation from one observation to the next is important for maximizing the ability to detect the underlying "signal" of the law's effects. This is also known as maximizing statistical power (Cohen, 1988).

A fourth factor affecting the best time resolution to measure is exactly when the law took effect – a January 1 effective date works well with annual data, but typical effective dates of public health laws are distributed throughout the year. Using annual data with laws that take effect mid-year requires assumptions that the effect is going to be, say, half the size of effect in the subsequent full-year implementation (if the effective date is July 1), but those annual data will not permit the investigator to evaluate the validity of that assumption. Sometimes anticipatory effects of a new law are seen starting a couple of months before it takes effect, or there are lagged effects that do not start until a few months after it legally takes effect. Perhaps the short-term effects are much larger than long-term effects, a situation common with laws that require public attention and active enforcement. Or the longer-term effects might be larger than the short-term effects, a situation common with laws that require construction of or refinements in an implementation structure before the full effects are seen. All these situations are obscured by selecting outcome data at too coarse a time resolution (for example, annual rather than monthly).

Keep in mind that a date may seem like a straightforward data element, but actually requires careful thought. Several dates are important to consider in evaluating a law's effects. There is the date the law is *introduced for debate;* the date it is *enacted,* passed by a legislative body or signed into law by the executive; the date specified by law that it *legally takes effect*; the date *actual implementation* of the law begins. The specific dates of most interest in the evaluation are based on hypothesized mechanisms of action drawn from theory.

Finally, when designing a study with lower time-resolution measures of continuous outcomes, it is critically important to take the measure at exactly the same time each year. This is because most physical, behavioral, and social phenomenon are characterized by seasonality – a nonrandom cycle within the time unit of observation. Pollution levels, dietary vegetable intake, infection rates, injuries, and most other health-relevant outcomes exhibit cyclic or other systematic differences across hours of the day, days of the week, weeks of the month, or months of the year (Figure 14.3).

So, if one is surveying individuals once per year, or inspecting restaurants or schools once per year to collect an outcome for evaluating a public health law, it is important to do the data collection the exact same month of the year. This applies at all time resolutions of measurement – if one is collecting data once per month, measure the same day each month (for example, first Wednesday of the month). If one is collecting data weekly, measure on the same day and same time of day each time. The further the data collection procedures diverge from measurement at the exact same time within the time unit, the less confident one can be in interpreting observed differences from before to after a new law is implemented as representing the effect of the law – it might be just because the measures were taken at a different point in the cycle.

In summary, a strong public health law evaluation has as many observations as possible before and after the law takes effect – a lengthy time series – and uses the highest time resolution possible,

constrained by the nature of the hypothesized effect, the frequency of underlying outcome counts, and feasibility limits due to resources or data available.

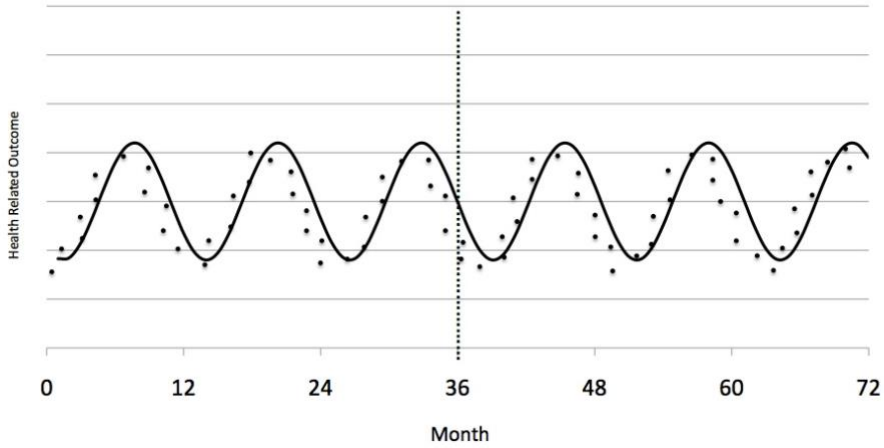

Figure 14.3. Time Series Illustrating Seasonality.

FUNCTIONAL FORM OF EFFECTS

High-time-resolution data have another important advantage furthering the quality of a policy evaluation. On the basis of theory regarding mechanisms of a law's effects, one has an implicit or (even better) explicit hypothesis on the expected pattern of the effect over time (Figure 14.4).

Imagine one's theory of legal action is based on deterrence. In that case, one may expect a lag before effects are seen due to enforcement taking time to ramp up and news about enforcement actions taking time to spread in the relevant population. Alternatively, if one's theory focuses more on normative compliance, initial timing of expected effects is based on when the relevant population first hears about the new law, suggesting that effects might be observed even before it legally takes effect, due to attention gained by hearings on the proposed law or to publicity surrounding a governor's signing the law. Therefore, one might expect effects to emerge at the enactment date, rather than the more typical expectation of little or no effect until the new law legally takes effect. The amount of time between enactment and the date the law takes effect might affect the magnitude and timing of expected effects. A longer lead time may enhance effects if it allows for better design and ramp up of implementation structures and practices. Some statutes include specific provisions for implementation, but most do not; such provisions might affect the hypothesized timing and size of effects. Implementation is not limited to the public sector. Private organizations and individuals also might require time to put in place what is needed for the law to fully take effect (e.g. train staff, purchase compliance equipment). Furthermore, implementation might vary significantly across sub-units of the jurisdiction (e.g. counties might differentially implement a state law). Consideration of likely implementation features and their timelines, both by public sector officials and relevant private organizations and individuals, will influence facets of the research design.

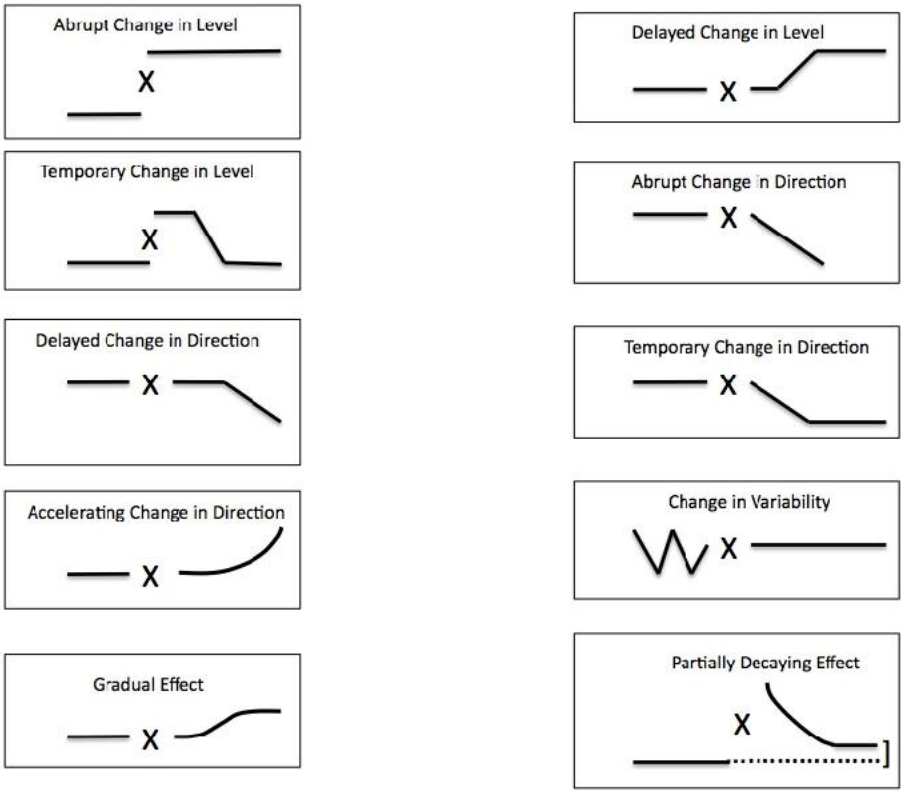

Figure 14.4. Possible Patterns of Policy Effects Over Time.
*Note: **X** equals policy change.*
Source: Adapted from Glass, Wilson, and Gottman, 1975.

Hypothesizing particular functional forms for legal effects leads to the following types of questions that shape the design of the study and the nature of the data to be collected. Is the effect expected to show up immediately when the law takes effect? Or is a delay of weeks or months expected, as enforcement or other implementation systems are developed and ramped up? Might there be an anticipatory effect before the legal effective date, due to publicity and attention to the issue surrounding debate on the new legislation, or widespread media reports at the time the law is passed? Is the effect expected to emerge gradually, as various implementation systems change or norms and behaviors gradually change? Or is the effect hypothesized to be temporary, dissipating over time as organizations and individuals adapt to the new law in ways to maintain previous conditions or behaviors?

Most public health laws are designed to affect the *level* of relevant outcomes, but there may be rare situations in which the expected effect is on another dimension, such as the *variance*. For example, laws and regulations might affect the amount of health care utilization by individual citizens, when the optimal public health objective might be to reduce both over-utilization and under-utilization – reducing the variance – while not affecting the overall level of services provided. Another example might be policies designed to reduce variance in caloric intake among children

eating school lunches – some children overeating and others undereating both represent health and school performance risks. Thus, the objective of regulations may be to reduce variance in calories consumed at school with no effect on overall level of calories or amount of food consumed at the school.

The bottom two panels in Figure 14.4 illustrate common patterns of effect of public health laws. The first illustrates the conventional "S-curve," when change starts slowly until reaching some "tipping point" at which change accelerates, followed by a leveling off at the (new) long-term level (Granovetter, 1978). The last panel of Figure 14.4 illustrates a sizable, fairly immediate effect that then partially dissipates over time (perhaps due to reduced attention to the issue), resulting in a much smaller, but often still important, long-term effect. One can see this in effects of strengthened driving-while-intoxicated laws, which often receive considerable media attention around the time they are passed or implemented, sometimes magnified by advocacy groups such as Mothers Against Drunk Driving, substantially raising the perceived probability of being detected and punished for driving impaired. As the short-term publicity declines, the magnitude of effect on driving behaviors also declines. But as the strengthened laws are integrated into ongoing enforcement efforts, the real and perceived probabilities of detection and punishment remain higher than baseline before the law, with a more modest but still important long-term effect.

In short, decisions on time resolution of outcome data to collect and their analyses should be informed by expected patterns of effect over time. It is important to note that if the observed pattern of effect closely matches the hypothesized pattern that is based on a particular theory regarding the operating legal mechanism, the level of confidence in causally attributing the observed effect to the change in law or regulation is substantially strengthened.

COMPARISON JURISDICTIONS

With many repeated observations correctly measured and analyzed, it is possible to determine with a high degree of accuracy whether a change in the outcome coincides with the time of implementation of a new law or regulation – a change that is larger than expected from normal variation over time, and a change that matches the theoretically expected pattern. However, we still have the problem of the counterfactual – what if the same change in outcome would have occurred regardless of whether the new law was implemented or not? The observed change might have been caused by something else happening at the same time. A fundamental way to further improve causal inference – to assess whether the law caused the change in outcome or whether something else caused it – is to use comparison jurisdictions that did not implement the law under study. One collects the same outcome data for another city or state that did not change their law, and examines whether the observed change in the "experimental" jurisdiction is also seen in the comparison jurisdiction. If no similar change is seen in the comparison, one is more confident that the observed change at the time of the law is in fact due to the law, and not to some other factor occurring in

common across jurisdictions. On the other hand, if a similar change is seen in the comparison, the observed change in outcome in the experimental site cannot be attributed to the change in law.

A key design consideration is selecting an appropriate comparison site. This is most commonly described as a site that is similar to the experimental site in terms of observable factors correlated with the outcome if not the outcome itself (in terms of level, trend line or variance). Typically, evaluators select a site with broadly similar sociodemographic profiles of the population, or similar counts or rates on the key outcome variables. There are many dimensions on which one might assess degree of similarity, so it is important to consider the underlying reason why one seeks similar jurisdictions. Choosing a site with similar counts or rates on the outcome is a helpful but relatively minor consideration – it makes it easier to determine whether the comparison site experienced a change in outcome that is similar to the change observed in the experimental site. In other words, it helps ensure approximately equal statistical power to estimate change in the outcome in both the experimental and comparison sites.

The fundamental criterion for comparison site selection has much deeper significance, since it is directly connected to achieving the best possible counterfactual. The fundamental criterion for selection of a comparison site is that all the *causes* of the outcome variable are similar across the two sites. Thus, the conventional approach to choose sites of similar demographics might be appropriate *if* demographics are a key influence on the outcome under study. But in many cases, other factors are more important in any particular study. For example, if car crashes are the outcome, similar urbanization and climate are likely more important than demographics, with the exception perhaps of proportion of young drivers, since they are at such elevated risk.

Stratification before selection of comparison sites optimally is based on multiple characteristics. For example, in policy research focused on promoting healthy food environments, it may be important to find comparison sites based on urbanity, sociodemographic factors, and the overall food environment, all of which are generally associated with outcomes of interest. The goal is to achieve two groups as similar as possible in an attempt to mimic the counterfactual – what a particular outcome would look like with or without a particular policy among the same group of people at the same time in history. Selecting an optimal comparison group is an attempt to rule out competing alternative explanations for the outcomes observed post-intervention. The goal is to be able to attribute any difference between the jurisdictions to the legal intervention of interest, and rule out any other plausible explanations as best as possible. For example, if the goal was to evaluate effects of a new food policy, it would be critical to select comparison sites with similar socioeconomic and food environments prior to the new policy to help rule out alternative explanations for change in outcomes. If data for a longer baseline period with many observations are available (as is recommended), a useful tactic is to examine the correlation of the outcome variables between the experimental site and candidate comparison sites during the baseline period only; then select comparison sites with the highest correlations.

Of course, a perfect comparison jurisdiction is unachievable, because no two jurisdictions are identical in every way but for the law under study. For this reason, it helps improve inference by including multiple comparison jurisdictions. If a clear change in outcome is observed in the one with the law change, but no such change is seen in several other similar jurisdictions that did not change their law, inference that the law caused the change in the first site is enhanced.

Synthetic Comparisons

The discussion of comparisons thus far has focused on finding the best available comparison jurisdiction(s), selected from the set of all available comparison jurisdictions. But none of those jurisdictions is likely to be the perfect comparison – exactly the same as the focal jurisdiction in every way but for the specific law being evaluated. This has led to the development of synthetic control methods (Abadie, 2021) that build on ideas from propensity score methods (Holmes, 2014). Synthetic controls are a creative and important advance, but the idea is simple: instead of selecting one (or a few) comparison jurisdictions, a single weighted average of the units in the pool of comparisons is constructed and used as the comparison. This weighted average is constructed in such a way as to maximize the correlation between the experimental jurisdiction (the one that changed a law) and the synthetic control over the baseline time period. Thus, a long baseline with many repeated measures is essential for valid construction of a good synthetic control.

Depending on the structure of available data, there are additional ways to improve the construction of synthetic controls. Many applications of synthetic controls remain at one level of aggregation. For example, imagine a US state changes a law, perhaps making mask use compulsory, and the study design is using the pool of other states that have never implemented compulsory mask use as candidate comparisons to assess effects on state-level COVID-19 test positivity rates. When constructing the synthetic control, we have, at most, measures on n=49 other states in the comparison pool to use as the "raw material" to construct the optimum synthetic control. However, there are also many times when we have data available measured at a sub-unit of the research design unit – data on individual state residents, for example, while the research design is evaluating a state-wide law by comparing to other states. In such a case, we have outcome measures on thousands or millions of people, and individuals can be differentially weighted to create the optimum synthetic comparison group.

Synthetic controls are a great design element to minimize the risk of some types of selection bias – differences between the treatment and comparison groups (other than the law being evaluated) that may confound a causal interpretation of observed effects. However, they do not eliminate the risk of history confounds. To take our example further, imagine many of the states that did not make masks compulsory made vaccines mandatory at about the same time the focal states were making masks mandatory. Despite the synthetic control closely matching the treatment states in outcome rates during the baseline period, it nevertheless is not a good comparison because of the contemporary history confound. It is always best to layer in multiple design elements to strengthen causal inference.

COMPARISON GROUPS

The notion of incorporating comparisons not expected to be affected by the law under study can be fruitfully extended in other directions beyond comparison jurisdictions. If a law or regulation is targeted to particular groups of people or organizations within a jurisdiction, effects on that focal group targeted should be compared with other similar groups within the same jurisdiction that are not likely to be affected. The benefit of within-jurisdiction comparisons is the equal exposure of treatment and comparison groups to the totality of all other laws and conditions present in the state, except for the specific law under evaluation. For example, consider a new state regulation intended to reduce worker injuries in auto repair shops. The injury rate can be tracked before and after the new regulation, and an observed reduction in injuries among auto-repair shop personnel is suggestive of an effect of the law. But inference of a causal effect would be strengthened by tracking similar measures of injuries among workers in the state that work in types of workplaces other than auto-repair shops. If similar declines in injuries were observed, then the observed auto-repair injury reductions are likely due to some other broader factor, and are not an effect of the new regulation specific to auto-repair shops. On the other hand, an observed reduction in injuries only for the specific group covered by the new law, with no reduction for workers in other similar settings not covered by the law, substantially strengthens the inference that the new law caused the reductions in auto-repair injuries. Most laws and regulations are inherently targeted in some way, opening important opportunities for enhancing causal inference regarding the law's effects by incorporating relevant within-jurisdiction comparison groups. For example, zoning rules that prohibit elementary schools from being sited adjacent to major highways (as a means to reduce air pollution exposure and asthma) can be evaluated by incorporating comparisons consisting of preschool, or middle- and high-school students not covered by the law.

COMPARISON OUTCOMES

Additional options for comparisons are provided by outcome variables. Appropriate comparison outcomes are related to the primary outcome, but, of importance, are not affected by the law or policy under study. For example, to evaluate the effect of New York City's regulation to post calorie information in chain restaurants, one might compare sale receipts for food purchased at chain restaurants to receipts from non-chain restaurants. To evaluate effects of motorcycle helmet laws, comparisons of car to motorcycle fatality and injury rates have been conducted (Sosin & Sacks, 1992). To evaluate effects of a graduated driver's license for teen drivers that forbids night-time driving, comparisons have been made between daytime and nighttime teen driver fatalities (Morrisey, Grabowski, Dee, & Campbell, 2006). The difference between labeling a comparison a "group" or an "outcome" is sometimes just a matter of convention. The central importance of the notion of comparison outcomes is to expand one's thinking and highlight the many comparisons, even within the jurisdiction enacting a new law or regulation, that can be effectively used to create strong research designs for evaluating the law's effects.

REPLICATIONS

A fundamental way to strengthen causal inference regarding a law's effects is to replicate the evaluation across jurisdictions. Naturally, the larger the number of sites in the treatment and control groups, the more (statistical) power one has to measure potential effects. And, if similar effects are observed in each place a similar law is implemented, confidence in causal inference is clearly enhanced. If multiple jurisdictions implementing the policy are included in an initial study, observed effect sizes can be directly compared across sites. If statistical power remains too low to reliably compare site-by-site effects, incremental removal of one site at a time from the analysis model can help determine whether the observed effect is driven by a subset of sites, rather than consistently across sites. But often, replications happen later, and are separately studied, perhaps by other investigators. If the initial observed effect is not seen in subsequent replications, suspicion increases that some other idiosyncratic or uncontrolled factor accounts for the observed effect in the first jurisdiction, and the law under study may have had no effect.

It is often better to evaluate each instance of a law, rather than the all-too-common practice of lumping all similar laws together into a single group and estimating the average effect across all specific instances. Consider the situation in which such a pooled analysis hints that a law might have small effects, but the effect is too small to be reliably measured (that is, is not statistically significant), leading to the conclusion that the regulatory approach is ineffective. Now imagine that in that pooled analysis lurk five states with large clear beneficial effects but 10 other states with no effects. The pooled analysis might prematurely discredit the regulatory approach and miss the opportunity for more in-depth analyses of the individual states to better understand why the law works in some cases and not in others, leading to improvements in implementation and further replication of effective approaches.

Replications occur not only across sites but also over time. As jurisdictions change the law on a particular subject in different years or decades, evaluation designs incorporating those replications ensure that observed effects are not due to other factors specific to a given era, again increasing confidence the observed effects are caused by the law under study. A whole area of research design in general involves manipulating the timing of a treatment or intervention. As expected, random assignment of a treatment to a particular time of implementation is a great strategy, just like randomly assigning a treatment to groups or jurisdictions, but it is rarely feasible. However, even without random assignment, naturally occurring (that is, induced by legislatures, courts, or administrators) variation over time in law in a single jurisdiction can be used effectively to dramatically strengthen the evaluation.

Psychologists call these "ABAB" designs, in which a treatment is applied, then removed, then later reapplied, and they can support strong causal inference (Kratochwill et al., 2010). Thus, we know with little doubt the causal effects of compulsory motorcycle helmet laws, since some states implemented such laws, later rescinded them, then still later reinstated them, creating an ABAB design (an "A" period without compulsory helmet law, then a "B" period with, then an "A" period without, followed by another "B" period with the law, all within one jurisdiction). The match

between the legal changes and morbidity or mortality outcomes in both directions supports strong causal inference (deaths decline abruptly when helmets become compulsory and abruptly return to the higher levels again when the law is rescinded [Mertz & Weiss, 2008; Ulmer & Preusser, 2003]).

DOSE RESPONSE

The notion of replications, when a similar law is implemented in multiple jurisdictions, and reversals, when a law is implemented and then removed, can be straightforwardly extended to replications in which the dose of a particular regulatory approach varies by jurisdiction or within jurisdiction over time. Dose can represent many different dimensions, tied to theory on the mechanism of the law's effects. All good legal evaluation studies should be based on a clear understanding of the underlying theory regarding legal mechanisms. This is especially true for designing a good dose-response study, because what constitutes different "doses" of the law is inherently tied to how one thinks the particular law works. It could be the size and speed of application of a penalty in a deterrence-based statute, for example, or many other dimensions of breadth, strength, or reach of a law. After effects of the law are assessed within each jurisdiction, jurisdictions are arrayed in order of low to high "dose" of the law. If the magnitude of observed effect tracks the dose – low-dose jurisdictions have small effects and high-dose jurisdictions have large effects – the causal attribution of the observed effects to the laws is substantially strengthened.

Dose-response studies substantially strengthen causal inference, but can have complications. Because the dosages are not randomly assigned to different jurisdictions and different times, it is possible the dose applied in a particular situation is correlated with some other characteristic of that situation or that time period. For example, if all high-dose locations are highly urbanized areas, and all low-dose locations are rural, perhaps dose does not truly affect the magnitude of legal effect and the observed dose-response relationship is really due to urbanism. The risk of such misattribution of effect is lowered by examining the pool of jurisdictions with differing doses for other differences that plausibly might explain the pattern of effects observed.

MULTIPLE DESIGN ELEMENTS

Evaluating effects of laws on public health outcomes should be guided by optimum use of multiple design elements for constructing experiments and quasi-experiments. For most cases, when randomization is not feasible, the use of matched comparisons (jurisdictions, groups, and outcomes) in combination with many repeated measures is recommended. Keep in mind that comparisons need not be matched one for one. One jurisdiction implementing a new law is typically compared with a similar jurisdiction that has not. Causal inference is often enhanced by using several jurisdictions in comparison with the one implementing a new law rather than just one. And comparisons of different kinds nested in a hierarchical fashion substantially strengthen the design. Finally, when multiple sites pass new laws, replications can be built directly into the design.

An illustration of such a combination of design elements that produced strong causal inferences about a law's effects can be seen in studies of the legal drinking age (Figure 14.5).

Two states that changed the legal age for possession and consumption of alcoholic beverages (Maine and Michigan) were compared to two states with, at that time, an unchanged drinking age (New York, with a consistent legal age of 18 since Prohibition ended, and Pennsylvania, with a consistent legal age of 21). Experimental states versus comparison states constituted the first level of comparison. Second, nested within each state, the focal age group affected by the change in law (18- to 20-year-olds) was compared to younger and older age groups. Third, nested within each age group, frequencies and rates of alcohol-related car crashes were compared to frequencies and rates of non-alcohol-related crashes. Fourth, to avoid the possibility that the law changed reporting of alcohol involvement perhaps more than the actual incidence of such crashes, two measures of alcohol-related crashes were observed – one based on normal crash reports by police officers regarding drivers' drinking, and an alternative that did not rely on officer reports of drinking (single-vehicle nighttime crashes, which are well-known from other research to have a high probability of involving a drinking driver). These two measures were compared with crashes with no police report of drinking and crashes occurring during the day – providing two measures of non-alcohol-related crashes.

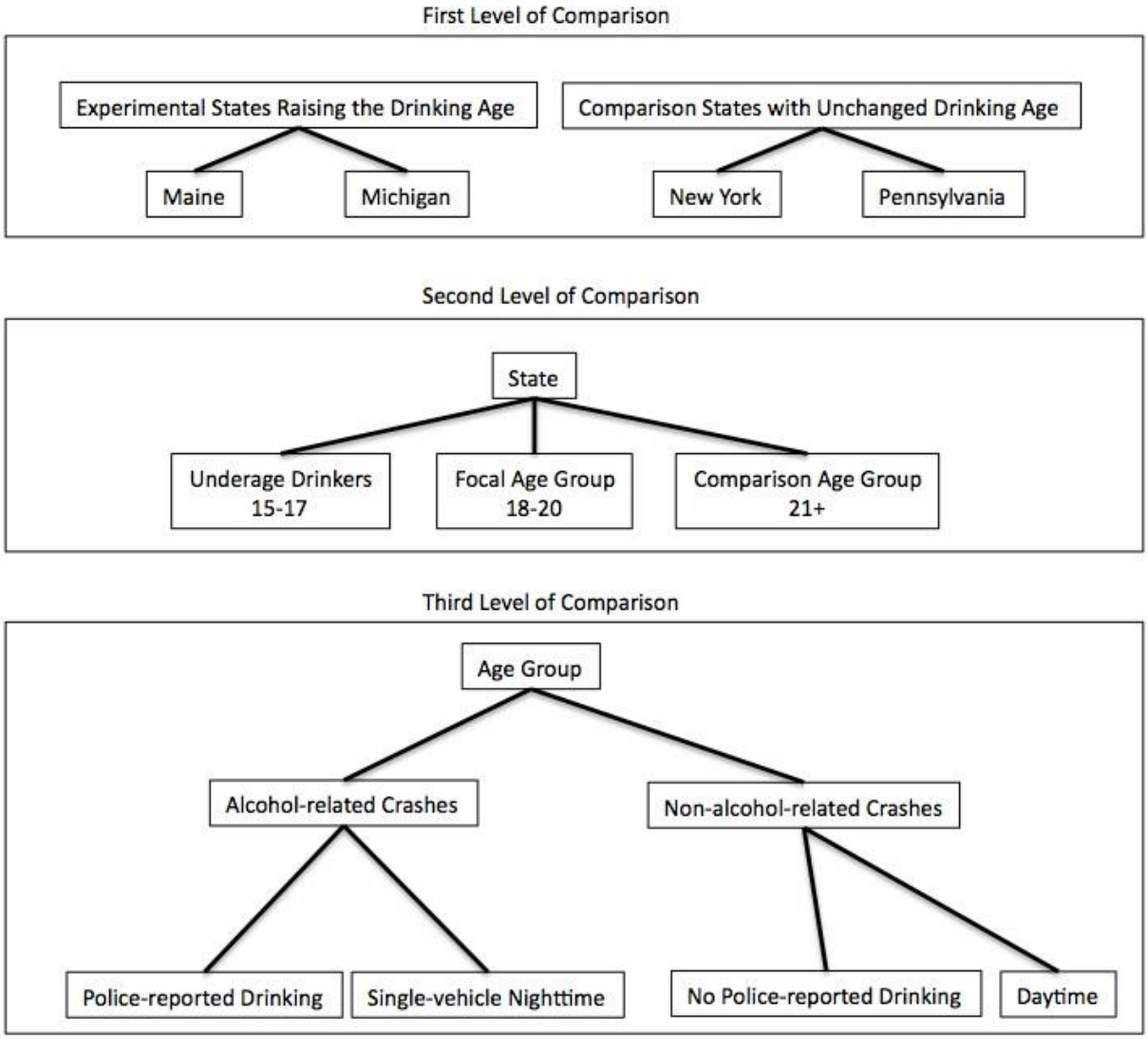

Figure 14.5. Hierarchical Multilevel Time-Series Design: Legal Drinking Age Example.
Source: Wagenaar, 1983a.

For each cell in this hierarchical design, outcomes were measured monthly for many years before and after the legal changes. The pattern of observed effects – reductions in crashes beginning the first month after the new law, only in the "experimental" states that raised their legal drinking age (and not in the comparison states), only among teenagers (not among drivers 21 and over who were not affected by the change in legal age from 18 to 21), only among alcohol-related crashes (and not among non-alcohol-related crashes, and confirmed with two alternative measures of alcohol-related crashes) – together produced an inference with high levels of confidence that this particular law caused a change in car crashes. Replications in other states that raised the legal age confirmed this pattern of effects. Moreover, a look back to reports and studies from a decade earlier in the 1970s, when 29 states lowered their legal age for drinking, produced an implicit ABAB

or intervention reversal design. After many states lowered the legal age for drinking in the early 1970s, teen car crashes increased; when a decade later the legal age was returned to 21, crashes decreased, reversing the earlier increase.

Despite periodic renewed attention to the legal age issue, with various individuals and organizations occasionally arguing in favor of returning again to a lower legal drinking age, the fundamental findings from the decades-earlier research have not been seriously challenged by scientists or most evidence-based review panels. In fact, the US National Highway Traffic Safety Administration (2010) estimates the age-21 law continues to prevent about 900 teen crash fatalities per year, saving more than 25,000 lives since the 1970s (Fell, Fisher, Voas, Blackman, & Tippetts, 2009; Voas, Tippetts, & Fell, 2003). Empirical legal evaluations that creatively took advantage of numerous design elements for strong causal inference produced important empirical results that have continuing policy relevance decades later.

REGRESSION-DISCONTINUITY DESIGN

All of the discussion thus far has assumed a time-series design, where the treatment assignment cut-point (i.e. discontinuity) is a particular point in time – no treatment before the law takes effect vs. treatment implementation after a law takes effect. The regression-discontinuity design is conceptually similar, except the cut-point being analyzed for a possible treatment effect is along some other variable, not time (Cattaneo & Escanciano, 2017). For example, rather than applying universally, some laws might apply only to persons in a jurisdiction below or above some specified criterion. Common examples are income limits – a law applies to anyone below a specified income level and to nobody above that income level. In that case, the income cut-off can be used in a regression-discontinuity design to evaluate the law's effects. The people just a couple dollars above the limit are, on average as a group, plausibly identical to those just a couple dollars under the limit. Another example: consider a state law that makes masks mandatory in any county when the COVID-19 ICU occupancy rate exceeds a specified level. The ICU rate cut-point is then used to evaluate a potential discontinuity in the outcome, such as rate of new infections.

Conclusion

Given the number of design elements available to strengthen empirical evaluations of public health laws and regulations, opportunities for continued improvement in the science on public health laws is clear. Awareness and understanding of available research designs for use in the real world outside the laboratory, where random assignment to both treatment and control conditions is often difficult, is important not only for scientists and legal scholars but also for policy makers, public health professionals, and advocates as well. Advancing the effectiveness of heath policy requires differential weighting of the evidence coming from various studies based on the quality of the research design – how well a given study incorporates multiple design elements and thus produces high-confidence causal conclusions. A simple before-and-after design should get little weight in policy deliberations compared to a high-time-resolution, time-series study that includes a hundred

or more repeated observations. A single-state study with no comparisons should get less weight than one that incorporates multiple comparison states and multiple comparison outcomes within each state.

High-quality and consistently implemented monitoring systems of relevant population-level health outcomes is critical for increasing the number of well-designed time-series evaluations. These ongoing data-collection efforts are the "management information systems" for population health, facilitating the monitoring of health status, evaluation of changes in laws, regulations and implementation procedures, and achievement of expected standards of health and safety for the population as a whole. Continuing outcome-monitoring systems are necessary for "continuous quality improvement" in the health and well-being of the population. A great example of the role of such information systems is the Fatality Analysis Reporting System, which collects hundreds of detailed data elements on every fatal car crash in the entire United States. The system was carefully designed and tested by a large community of scientists and engineers inside and outside the federal government in the 1960s and early 1970s. Then, full implementation began in 1975, and has continued ever since. The complete data in analysis-ready formats are publicly and easily available. This data system resulted in an explosion of research on the causes and prevention of car crash deaths, and each year as additional longitudinal data are added, more high-quality time-series evaluations are possible. Because of the knowledge gained from thousands of studies using these data over the past few decades, we have saved hundreds of thousands of lives and millions of injuries. This is a truly phenomenal public health achievement (Hemenway, 2009). For decades, each time a state innovates with laws and regulations designed to further reduce crash injuries, investigators can simply access the data system and build well-designed multistate time-series studies evaluating the effect of the change.

There are many other examples of emerging data systems that will facilitate the use of strong time-series research designs to evaluate the effects of laws and regulations. The dissemination of electronic medical records (Hillestad, Bigelow, Bower, et al., 2005), including records on health risk behaviors recorded routinely in primary care practices (Hung, Rundall, Tallia, et al., 2007), will provide population-level daily, weekly, or monthly indicators of health-relevant behaviors and outcomes. New technologies implemented at scale, such as the Apple watch, which monitors and records numerous health indicators continuously for millions of users, will open up new opportunities for evaluating laws using high frequency and high-density data in long time series. All these continuing improvements in the breadth, quality, consistency, and availability of continuous monitoring data systems will facilitate further well-designed evaluations of the effects of laws and regulations.

Combining many design elements in a hierarchical multiple time-series research design represents the best approach for evaluating the effects of public health laws and regulations, in many ways providing better knowledge of effect than that gained from randomized controlled trials (RCTs). Randomization to treatment condition is a useful design strategy in many fields (for

example, testing specific treatments such as new pharmaceuticals), but has more limited utility in legal epidemiology studies. RCTs can be used productively to study the effects of specific "micro" mechanisms found in many theories of legal effect, and those results help design better laws and regulations. But RCTs, of necessity, are almost always conducted in small, localized, and unnatural laboratory-type settings, with small samples of people. Natural experiments with public health relevant laws, in contrast, are implemented in real-world settings, use the actual legal tools and implementation processes widely available in society, and apply to very broad or universal populations. And results from actual field implementations of laws and regulations are more persuasive to policy makers, public health practitioners, and citizens, facilitating diffusion of successful approaches to other jurisdictions, resulting in major improvements in population health.

Further Reading

Abadie A and MD Cattaneo (2018) "Econometric Methods for Program Evaluation" *Annual Review of Economics* 10: 465-503.

Bernal J.L., Cummins, S. & Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. International Journal of Epidemiology, 46(1), 348-355.

Box, G. E. P, Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. (2016). *Time Series Analysis: Forecasting and Control.* Hoboken, NJ: John Wiley & Sons.

Cunningham, Scott. *Causal Inference: The Mixtape*, New Haven: Yale University Press, 2021.

McCleary, R., McDowell, D. & Bartos, B. J. (2017). *Design and Analysis of Time Series Experiments*. New York: Oxford University Press.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods, 15*, 3–17.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Shadish, W. R., & Sullivan K. (2012). Theories of causation in psychological science. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 1, pp. 23–52). Washington, DC: American Psychological Association.

References

Abadie, A. (2021). Using synthetic controls: feasibility, data requirements, and methodological aspects. *Journal of Economic Literature, 59*(2), 391-425.

Bärnighausen, T., Røttingen, J.-A., Rockers, P., Shemilt, I., & Tugwell, P. (2017). Quasi-experimental study designs series -- paper 1: introduction: two historical lineages. *Journal of Clinical Epidemiology, 89*, 4-11.

Bärnighausen, T., Tugwell, P., Røttingen, J. A., Shemilt, I., Rockers, P., Geldsetzer, P., . . . Atun, R. (2017). Quasi-experimental study designs series--paper 4: uses and value. *J Clin Epidemiol, 89*, 21-29.

Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine, 342*(25), 1878–1886.

Cattaneo, M. D., & Escanciano, J. C. (2017). *Regression discontinuity designs : theory and applications* (First edition. ed.). Bingley, United Kingdom: Emerald Publishing.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hills-dale, NJ: Lawrence Erlbaum Associates.

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine, 342*(25), 1887–1892.

Fell, J. C., Fisher, D. A., Voas, R. B., Blackman, K., & Tippetts, A. S. (2009). Changes in alcohol-involved fatal crashes associated with tougher state alcohol legislation. *Alcoholism—Clinical and Experimental Research, 33*(7), 1208–1219.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, London: Oliver and Boyde.

Glass, G. V., Wilson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments*. Boulder: Colorado Associated University Press.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology, 83*(6), 1420–1443.

Guyatt, G. H., DiCenso, A., Farewell, V., Willan, A., & Griffith, L. (2000). Randomized trials versus observational studies in adolescent pregnancy prevention. *Journal of Clinical Epidemiology, 53*(2), 167–174.

Leatherdale, S. T. (2019). Natural experiment methodology for research: a review of how different methods can support real-world research. *International Journal of Social Research Methodology, 22*(1), 19-35.

Mertz, K. J., & Weiss, H. B. (2008). Changes in motorcycle-related head injury deaths, hospitalizations, and hospital charges following repeal of Pennsylvania's mandatory motorcycle helmet law. *American Journal of Public Health, 98*(8), 1464–1467.

Morrisey, M. A., Grabowski, D. C., Dee, T. S., & Campbell, C. (2006). The strength of graduated drivers license programs and fatalities among teen drivers and passengers. *Accident Analysis and Prevention, 38*(1), 135–141.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology, 66*(5), 688–701.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton–Mifflin.

Sosin, D. M., & Sacks, J. J. (1992). Motorcycle helmet-use laws and head injury prevention. *Journal of the American Medical Association, 267*(12), 1649–1651.

Ulmer, R. G., & Preusser, D. F. (2003). *Evaluation of the repeal of motorcycle helmet laws in Kentucky and Louisiana* (HS-809 530). Washington, DC: U.S. Department of Transportation.

Voas, R., Tippetts, A. S., & Fell, J. C. (2003). Assessing the effectiveness of minimum legal drinking age and zero tolerance laws in the United States. *Accident Analysis and Prevention, 35*(4), 579–587.

Wagenaar, A. C. (1983a). *Alcohol, young drivers, and traffic accidents: Effects of minimum age laws*. Lexington, MA: Lexington Books.