eGEMs
Generating Evidence & Methods to improve patient outcomes

Analytical Methods for a Learning Health System: 2. Design of Observational Studies

Michael Stoto, PhD;[i] Michael Oakes, PhD;[ii] Elizabeth Stuart, PhD;[iii] Elisa L. Priest, DrPH;[iv] Lucy Savitz, PhD, MBA[v]

ABSTRACT

The second paper in a series on how learning health systems can use routinely collected electronic health data (EHD) to advance knowledge and support continuous learning, this review summarizes study design approaches, including choosing appropriate data sources, and methods for design and analysis of natural and quasi-experiments.

The primary strength of study design approaches described in this section is that they study the impact of a deliberate intervention in real-world settings, which is critical for external validity. These evaluation designs address estimating the counterfactual – what would have happened if the intervention had not been implemented. At the individual level, epidemiologic designs focus on identifying situations in which bias is minimized. Natural and quasi- experiments focus on situations where the change in assignment breaks the usual links that could lead to confounding, reverse causation, and so forth. And because these observational studies typically use data gathered for patient management or administrative purposes, the possibility of observation bias is minimized. The disadvantages are that one cannot necessarily attribute the effect to the intervention (as opposed to other things that might have changed), and the results do not indicate what about the intervention made a difference.

Because they cannot rely on randomization to establish causality, program evaluation methods demand a more careful consideration of the "theory" of the intervention and how it is expected to play out. A logic model describing this theory can help to design appropriate comparisons, account for all influential variables in a model, and help to ensure that evaluation studies focus on the critical intermediate and long-term outcomes as well as possible confounders.

[i]Georgetown University, [ii]University of Minnesota, [iii]Johns Hopkins University, [iv]Baylor Health Care System, [v]Kaiser Permanente Center for Health Research

Introduction

Learning health systems use routinely collected electronic health data (EHD) to advance knowledge and support continuous learning. Even without randomization, observational studies can play a central role as the nation's health care system embraces comparative effectiveness research and patient-centered outcomes research. However, neither the breadth, timeliness, volume of the available information, nor sophisticated analytics, allow analysts to confidently infer causal relationships from observational data. Rather, depending on the research question, careful study design and appropriate analytical methods can improve the utility of EHD.

This is the second paper in a series (see Box 1) on how learning health systems can use routinely collected electronic health data (EHD) to advance knowledge and support continuous learning, this review summarizes study design approaches, including choosing appropriate data sources, and methods for design and analysis of natural and quasi-experiments. The first paper[1] began by drawing a distinction between big-data style analytics of electronic health data (EHD), with its claims that randomized studies were no longer necessary, and traditionalists who believe that without randomization little can be known with certainty. Of course this is a false distinction; some questions do not involve assessing a cause and effect relationship, but when causal assessment is

Box 1. Series on Analytic Methods to Improve the Use of Electronic Health Data in a Learning Health System

This is one of four papers in a series of papers intended to (1) illustrate how existing electronic health data (EHD) data can be used to improve performance in learning health systems, (2) describe how to frame research questions to use EHD most effectively, and (3) determine the basic elements of study design and analytical methods that can help to ensure rigorous results in this setting.

- Paper 1, "Framing the Research Question,"[2] focuses on clarifying the research question, including whether assessment of a causal relationship is necessary; why the randomized clinical trial (RCT) is regarded as the gold standard for assessing causal relationships, and how these conditions can be addressed in observational studies.
- Paper 2, this paper, addresses how study design approaches, including choosing appropriate data sources, methods for design and analysis of natural and quasi-experiments, and the use of logic models, can be used to reduce threats to validity in assessing whether interventions improve outcomes of interest.
- Paper 3, "Analysis of observational studies,"[3] describe how analytical methods for individual-level electronic health data EHD, including regression approaches, interrupted time series (ITS) analyses, instrumental variables, and propensity score methods, can be used to better assess whether interventions improve outcomes of interest.
- Paper 4, "Delivery system science,"[4] addresses translation and spread of innovations, where a different set of questions comes into play: How and why does the intervention work? How can a model be amended or transported to work in new settings? In these settings, causal inference is not the main issue, so a range of quantitative, qualitative, and mixed research designs are needed.

necessary observational studies of existing EHD can be a useful complement to RCTs. In particular, when the question is whether an intervention "works" – improves outcomes of interest, causal inference is indeed critical, but appropriately designed and analyzed observational studies can yield valid results that better balance internal and external validity than RCTs.

This second paper in the series addresses study design approaches including choosing appropriate data sources, and methods for design and analysis of natural and quasi-experiments. The primary issue addressed by evaluation designs is how to estimate the counterfactual – what would have happened if the intervention had not been implemented. Without randomization to establish causality, these methods demand a more careful consideration of the "theory" of the intervention. Logic models can help to design appropriate comparisons, and to ensure that evaluation studies focus on the critical intermediate and long-term outcomes as well as possible confounders. Another paper in this series[5] discusses analytical methods for EHD, including regression approaches, interrupted time series (ITS) analyses, instrumental variables, and propensity score methods that, together with the design principles discussed in this paper, can also be used to address the question of whether the intervention "works."

This paper does not attempt to serve as a textbook or describe these approaches in detail. Rather, it presents these methods in a consistent framework rather than provide detailed information on each topic. Because the use of existing EHD is not yet well developed, some of the examples use other types of data but were chosen to illustrate the methods.

The two major potential sources of bias in non-experimental studies of health care interventions are (1) different, if not zero, probabilities of exposure to experimental conditions (2) confounding. Cochran[6] recommends developing a clear definition of treatment and comparison conditions as well as clear inclusion and exclusion criteria for the study, so we begin this paper with a brief discussion of selecting data sources, patient populations, and comparators for which these types of bias are likely to be minimal. We continue with a discussion of how specific study designs – drawing mainly from analytical epidemiology – can also address issues of selection and confounding bias. This includes the standard cohort and case-control designs, and briefly introduces specialized designs including case-cohort, case-crossover, case-time controlled, and self-controlled case series. This paper introduces a series of methods that are drawn primarily from the social sciences and the program evaluation literature. In this tradition, logic models help to clarify the expected relationship among program elements, outputs, and short- and long-term outcomes, and how they relate to both program theory and external events. Such clear specification can help determine whether the results of observational studies do in fact reflect causal effects. The central concepts in this perspective are the natural and quasi-experiment. Both approaches capitalize on deliberate changes in exposure to the treatment unrelated to other factors influencing outcomes, breaking links that could lead to selection bias, reverse causation, and so on. In the Cochran tradition, this paper concludes with a discussion of how the design of experiments can be a useful perspective in designing observational studies, even if randomization is not possible.

Selection of Data Sources to Minimize Bias

Careful selection of data sources, patient populations, including criteria for inclusion/exclusion of individuals from those populations, and comparators for the statistical analyses described in this section can help to minimize confounding and selection bias. Two considerations are particularly important:

- the key variables must be available to define an analytic cohort and identify exposures, outcomes, and confounders,
- the data resource be sufficiently granular, contain historical information to determine baseline covariates, and represent an adequate duration of equivalent follow-up.

For example, consider the study by Stuart and colleagues[7] to estimate the effect of the Medicare Part D prescription drug program on individuals with serious mental illness clearly defines the population inclusion and exclusion criteria. The cohort they selected for analysis was defined by

- continuous enrollment in 2005 and 2006, so data on both baseline covariates and outcomes were available
- eligibility for both Medicare and Medicaid on Jan. 1, 2006, when Part D came into effect
- Maryland residency, to control for variation in state-level factors and policies
- diagnosis of schizophrenia, bipolar, or depressive disorders (as defined by pre-specified diagnostic codes in Maryland's all-payer billing files), because the impact on other health conditions could be quite different
- not being enrolled in Medicare Advantage, which has different drug benefits.

Table 1, drawn from AHRQ'S User's Guide for Developing a Protocol for Observational Comparative Effectiveness Research,[8] gives more specifics on this approach.

Study Designs for Individual-level EHD

Stürmer and Brookhart[9] write that the choice of study design often has profound consequences for the causal interpretation of study results and provide an overview of various study design options for non-experimental comparative effectiveness research (CER), with their relative advantages and limitations in controlling the potential for bias. See Velentgas[10] and Rothman[11] for more detail on the study designs described in this section.

Cohort Studies

Cohorts are defined by their exposure (including the receipt of a treatment), or lack thereof, over a specified time. Subjects are then typically followed for the occurrence of the outcome. If designed properly, the main advantage of the cohort is that it has a clear timeline separating potential confounders from the exposure, and the exposure from the outcome. This means cohorts allow the estimation of incidence (risk or rate) in all exposure groups and thus the estimation of risk or rate differences. Cohort studies also allow investigators to assess continuous outcomes as well as multiple outcomes from given treatments/exposures. The cohort design is also easy to conceptualize and is readily compared to an RCT, a design with which most clinical researchers are very familiar. The principal disadvantages of a cohort design are the lack of randomization and resulting potential for confounding bias as well as loss to follow-up, which may result in selection bias.[12]

A variant of the traditional cohort design is the case-cohort design. Here cohorts are defined as usual, but additional information required for analysis (e.g., blood levels, biologic materials for genetic analyses) is collected for all cases (those with the outcome of interest) and a random sample of others in the cohort. This design has the same advantages and disadvantages as a cohort study, but is more efficient in terms of data collection costs.[13]

Case-control Studies

Unlike a cohort design, a case-control design identifies all incident cases in a database that developed the outcome of interest and compares their exposure history with the exposure history of controls selected from the imagined cohort

Table 1. Questions to Consider When Choosing Data for an Observational Study

QUESTION TO ASK	EXAMPLE
Do the data contain a sufficiently long duration of followup after exposures?	Are there data on weight for at least three years after bariatric surgery?
Are there sufficient historical data to determine baseline covariates?	Is there information of hospitalizations int he year prior to cardiac resynchronization therapy for an observational study of outcomes from the device?
Is there a complete dataset from all appropriate settings of care to comprehensively identify exposures and outcomes?	Is there a record of emergency department visits in addition to a record of outpatient and hospitalized care in a study of children with asthma?
Are data available on other exposures outside of the healthcare setting?	Are there data on aspirin exposure when purchased over the counter in a study of outcomes after myocardial infarction?
Are there a sufficient number of observations in the dataset if restricting the patient population is necessary for internal validity (e.g., restriction to new users)?	Are there a sufficient number of new users (based on a "washout period" of at least 6 months) of each selective and non-selective nonsteroidal anti-inflammatory drug (NSAID) to study outcomes in users of each of these medications?
What is the difference between the study and target population demographics and distributions of comorbid illnesses? Will these differences affect the interpretation and generalizability of the results?	Is the age range of the data source appropriate to address the study question? Can any differences in demographics between data source and target population be addressed through appropriate design or analysis approaches?
Are the key variables available to define an analytic cohort (the study inclusion and exclusion criteria)?	Do the data contain height and weight or BMI to define a cohort of overweight or obese subjects?
Are the key variables available for identifying important subpopulations for the study?	Do the data contain a variable describing race for a study of racial differences in outcomes of coronary stenting?
Are the key variables available for identifying the relevant exposures, outcomes, and important covariates and confounders?	Do the data contain information on disease severity to assess the comparative effectiveness of conservative versus intensive management of prostate cancer? (Disease severity is a likely confounder.)
Are the data sufficiently granular for the purpose of the study?	Is it adequate to know whether the individual has hypertension or not, or is it important to know that the individual has Stage I or Stage III hypertension?
Are there a sufficient number of exposed individuals in the dataset?	Are there enough individuals who filled prescriptions for exenatide to study the outcomes from this medication?

Source: Stürmer and colleagues.[8]

of interest. Control selection is outside of the scope of this paper, but more information can be found in Rothman.[14] Given rare outcomes and proper sampling of controls from the risk set, the estimation of the odds ratio in a case-control study is a more efficient way to estimate the otherwise identical incidence rate ratio in the underlying cohort, especially for rare outcomes such as adverse reactions to medications. This efficiency is especially important if additional data (e.g., blood levels, biologic materials, validation data) need to be collected. The case-control design also allows for assessment of multiple exposures, but breaks down if outcome incidence exceeds 10 percent or the controls do not have the same risk of developing the outcome of interest.

For example, Glanz and colleagues[15] use a case-control design to examine the association between under vaccination (receipt of less than the recommended number of doses) and pertussis in children 3 to 36 months of age who were members of the eight managed care organizations participating in the Vaccine Safety Datalink study between 2004 and 2010. Each laboratory-confirmed case of pertussis (72 patients) was matched to 4 randomly selected controls (for a total of 288 controls). Cases were matched to controls by managed care organization site, sex, and age at the index date, defined as the date of pertussis diagnosis for the case patients. Using a conditional logistic regression analysis, the study found that under-vaccination with DTaP vaccine increases the risk of pertussis. The results, shown in Table 2, demonstrate that the risk of pertussis increases with the number of doses missed, and that the risk of pertussis is significantly higher for children who miss any number of doses (OR = 4.36, 95 percent C.I. = 2.23, -8.55, P<0.001). Because the number of adverse events is too small for differences to be detected in the RCT's that were done to establish the vaccines' efficacy, and because it would be both unethical and impractical to design a large-scale RCT to establish the vaccines' safety, Glanz and colleagues[16] observational study using existing EHD fills in an important knowledge gap.

Although they can be efficient, case-control studies have some important limitations. Because they begin with the outcome, case-control studies can be difficult to understand and explain. The major limitation of case-control studies is the potential for selection-bias if the controls are not representative of the imagined cohort.[17] Unless additional information from the underlying cohort is available, risk or rate differences cannot be estimated from case-control

Table 2. Estimates of the Risk of Laboratory-confirmed Pertussis for Those Undervaccinated vs. Age-appropriately Vaccinated

NUMBER OF DOSES UNDERVACCINATED BY	ODDS RATIO (OR) AND 95% CONFIDENCE INTERVAL	*P* VALUE
1 vs. 0	2.25 (0.97 – 5.24)	0.06
2 vs. 0	3.41 (0.89 – 13.05)	0.07
3 vs. 0	18.56 (4.92 – 69.95)	<0.001
4 vs. 0	28.38 (3.19 – 252.63)	0.002
1, 2, 3 or 4 vs. 0	4.36 (2.23 – 8.55)	<0.001

Source: Glanz and colleagues.[15]

studies. Because the timing between potential confounders and the treatments is often not taken into account, case-control designs typically assess confounders at the index date rather than prior to treatment initiation, so the results may be biased if they inadvertently control for covariates that may be affected by prior treatment. If information on treatments needs to be obtained retrospectively, such as from an interview with study participants identified as cases and controls, there is the potential that treatments will be assessed differently for cases and controls, which will lead to information bias, including when either cases or controls are more likely to recall exposure than the other group.

Case-crossover Design

For this study design, only patients with the outcome (cases) who have varying exposures during the study period contribute information, and the cases serve as their own controls. The self-control removes the confounding effect of any characteristic of subjects that is stable over time, such as genetics, so measures of stable confounding factors are not needed. Because this design depends so much on timing, it is appropriate only for acute effects of transient exposures. It is not appropriate for exposures that may have long-lasting effects. It can be useful in situations in which patients switch between two similar treatments without stopping treatment as long as the causes of switching are unrelated to health events (e.g., due to changes in health plan drug coverage) and as long as there is not carryover of the effects of the first treatment into the second time period. If the switching were triggered by health events, however, within-person confounding would bias the parameter estimates. For instance, if some patients stop taking a medication because of side effects, comparisons or outcomes between those on treatment and not might yield a biased estimate of the treatment effect.

Case-time Controlled Design

This study design adjusts for calendar time trends in the prevalence of treatments that can introduce bias in the case-crossover design. To do so, it uses controls as in a case-control design but estimates a case-crossover odds ratio (i.e., within individuals) in these controls. The case-crossover odds ratio (in cases) is then divided by the case-crossover odds ratio in controls. This design has the same advantages as the case-crossover design, but is not dependent on the assumption of no temporal changes in the prevalence of the treatment. The control for the time trend can introduce confounding, however, although the magnitude of this problem for various settings has not been quantified.

Self-controlled Case Series

As with the case-crossover design, the self-controlled case-series design estimates the immediate effect of treatment in those treated at least once. It is similarly dependent on cases that have changes in treatment during a defined period of observation time. The observation time is divided into treated person-time, a washout period of person-time, and untreated person-time. The immediate effect of treatment is estimated using conditional Poisson regression to estimate the incidence rate ratio within individuals. This design was originally proposed for rare adverse events in vaccine safety studies for which it seems especially well suited. As with the case-crossover design, the self-control removes the confounding effect of any characteristic of subjects that is stable over time.

Example: Intussusception Risk After Rotavirus Vaccination

Post-licensure studies have identified an increased risk of intussusception after vaccination with the second-generation rotavirus vaccines RotaTeq

(RV5) and Rotarix (RV1). Yih and colleagues[18] use a self-controlled risk interval (SCRI) design to assess this risk in infants. The study included data from infants 5.0 to 36.9 weeks of age who were enrolled in three U.S. health plans that participate in the FDA Mini-Sentinel program. Cases of intussusception and vaccine exposures from 2004 through mid-2011 were identified through procedural and diagnostic codes. Medical records were reviewed to confirm occurrence of intussusception and rotavirus vaccination.

The primary analysis used a self-controlled risk-interval (SCRI) design that included only vaccinated children. Two alternative risk intervals were used: 1-7 days and 1-21 days post vaccination. Assuming that any effect would occur in the first 21 days, the control interval was 22-42 days post vaccination. Logistic regression was used to adjust for age in risk and control intervals based on intussesception risk in external hospital data. The advantages of this design are that it inherently controls for all fixed potential confounders such as sex, race or ethnic group, and chronic predisposing conditions, and uses data only from exposed children, minimizing potential misclassification bias due to incomplete data on vaccine exposure.

A secondary analysis used a cohort design that included exposed (1-21 days post vaccination) and unexposed (5.0-36.9 weeks of age except 0-21 days post vaccination) person-time. Poisson regression was used to adjust for age (using a quadratic risk function), sex, and data partner. Calendar time and interactions were not included. Although the primary design was thought to be better for controlling selection bias, the latter was expected to have higher power.

As can be seen in more detail in Table 3, Yih and colleagues[19] found that for the first dose of RV5, the estimated relative risk for intussusception was 9.1 (95 percent CI, 2.2 to 38.6), representing an attributable risk of 1.5 (95 percent CI, 0.2 to 3.2) per 100,000 doses. The secondary analysis of RV1 suggested a potential risk, although the study of RV1 was underpowered. Thus, more than the simple results of one analysis, the patterns of positive findings where

Table 3. Case Counts and Risk Estimates for Confirmed Intussusception After First Dose of RV5 and RV1

DESIGN	DAYS AFTER VACCINATION IN RISK WINDOW	NUMBER OF CASES IN RISK WINDOW	NUMBER OF CASES IN CONTROL WINDOW	RELATIVE RISK (RR)	95% CONFIDENCE INTERVAL
RV5					
SCRI	1 to 7	5	3	9.1	(0.3 – 2.7)
SCRI	1 to 21	8	3	4.2	(1.1 – 16.0)
Cohort	1 to 21	8	97	2.6	(1.2 – 3.2)
RV1					
SCRI	1 to 7	1	0	–	–
SCRI	1 to 21	1	0	–	–
Cohort	1 to 21	1	97	3.2	(0.4 – 22.9)

SCRI = self-controlled risk interval design. Source: Adapted from Yih and colleagues.[18]

study design considerations would most expect them helps to strengthen the evidence for a causal effect.

Design and Analysis of Natural and Quasi-Experiments

Natural and quasi-experiment are central concepts in the social sciences oriented program evaluation literature. In a quasi-experiment, allocation to treatment and comparison groups is not random, as occurs when a law or policy changes in one jurisdiction but not others. In a natural experiment, subjects are randomly assigned to treatment and control conditions by a random process not controlled by the researcher. A good example of the latter is the "Oregon experiment" in which funding limitations permitted expansion of Oregon's Medicaid program to only 30,000 residents, chosen by lottery from 90,000 eligible persons.[20]

Although researchers differ in how they distinguish between natural and quasi-experiments, both approaches capitalize on deliberate changes in exposure to the treatment unrelated to other factors influencing outcomes, breaking the usual links that could lead to selection bias, reverse causation, and so on. Thus they address the first and third of Riegelman's criteria for a contributory cause: the cause precedes the effect and altering the cause results in a change in the effect. As a result, associations between the intervention and the outcome are more easily regarded as causal. And because these studies typically use data gathered for patient management or administrative purposes, the possibility of observation bias is minimized.

The primary focus of the study design of natural and quasi-experiments is in making an unbiased estimate of the counterfactual in order to assess Riegelman's second criterion, that the cause be associated with the effect. If assignment is random as in the Oregon experiment, this is not an issue as long as outcomes in the eligible non-participants are assessed. Otherwise, causal inference can be strengthened by (1) using a logic model to position data in context and consider likely confounders (observed and/or unobserved), (2) quasi-experimental design, especially finding settings where exogenous changes are unlikely and appropriate control subjects are available (3) multiple pre- and post-intervention measurements and the use of Interrupted Time Series (ITS) methods (see below) when appropriate, (4) individual-level statistical analysis of the type discussed in the third paper in this series[21] to control for relevant confounders, and (5) the use of pre-existing data to control recall bias (e.g. medical records rather than patient recall about use of a drug thought to cause adverse effects).

A logic model is a graphical representation of the logical relationships between the resources, activities, outputs and outcomes of a program. Logic models clarify the "theory of change" for an intervention and help to ensure that evaluation studies focus on the critical intermediate and long-term outcomes as well as possible confounders. They also can help to identify appropriate control groups. The generic logic model in Figure 1 illustrates how logic models represent the expected relationship among program elements, outputs, and short- and long-term outcomes and how they relate to both program theory and external events. Specifying these expectations in advance can help determine what intermediate and mediating variables should be measured, and help to determine whether the results of observational studies do in fact reflect causal effects.

Wagenaar & Komro[22] illustrate how quasi-experimental design elements can produce strong evidence of both whether a law or policy change caused an effect as well as the magnitude of effect. These can be summarized as follows:

1. Incorporate dozens or hundreds of repeated observations before and after an intervention creating a time series (i.e., use an interrupted time series analysis – see Stoto[23]).
2. Measure outcomes at an appropriate time resolution to enable examination of the expected pattern of effects over time based on a theory of the mechanisms of intervention's effect.
3. Include comparisons in the design, including multiple jurisdictions with and without the intervention under study, comparison groups within a jurisdiction of those exposed and not exposed to the intervention, and comparison outcomes expected to be affected by the intervention and similar outcomes not expected to be affected by the intervention under study.
4. Replicate the study in additional jurisdictions implementing similar interventions.
5. Examine whether the "dose" of the intervention across jurisdictions or across time is systematically related to the size of the effect.

Figure 2 demonstrates how high time-resolution data have another important advantage furthering the quality of a policy evaluation. Based on theory regarding the mechanisms of an intervention's effects, one has an implicit or (even better) explicit hypothesis on the expected pattern of the effect over time. Using several jurisdictions in comparison with the one implementing the intervention rather than just one often enhances causal inference. And comparisons of different kinds nested in a hierarchical fashion substantially strengthen the design. Figure 3 illustrates this approach, using a change in the legal drinking age as the example.[24]

Example of Natural and Quasi-Experiments: Massachusetts's Health Care Reform

In 2006 Massachusetts passed comprehensive health care reform with the goal of near-universal coverage. The law offered subsidized private insurance, expanded Medicaid, and created an individual mandate, serving as a model for the

Figure 1. Generic Logic Model

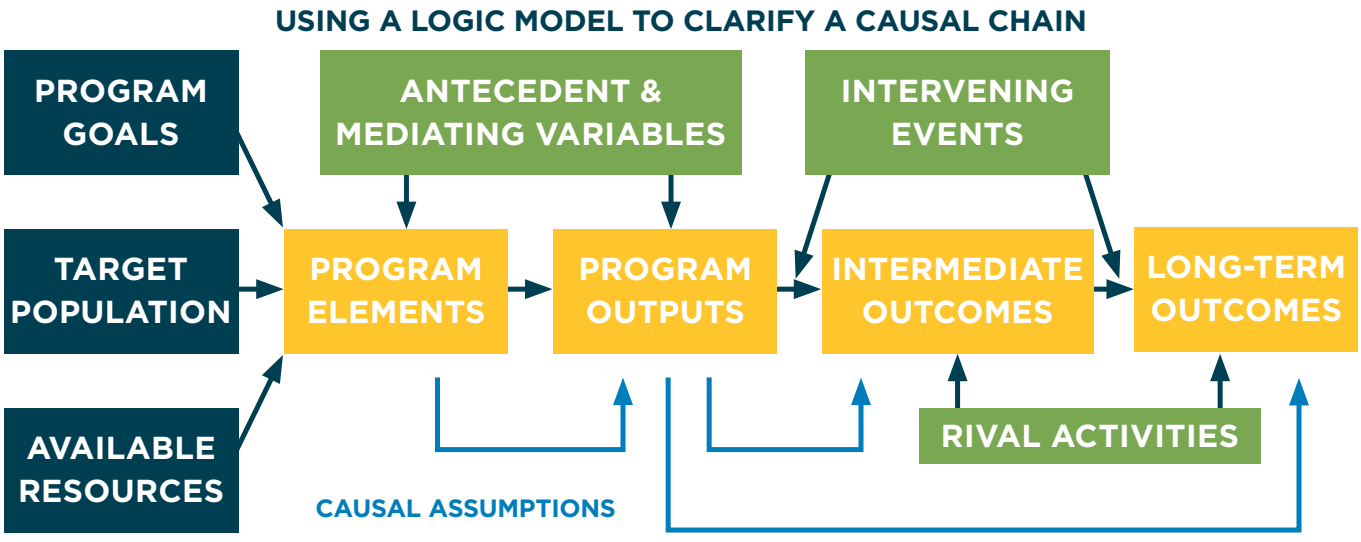
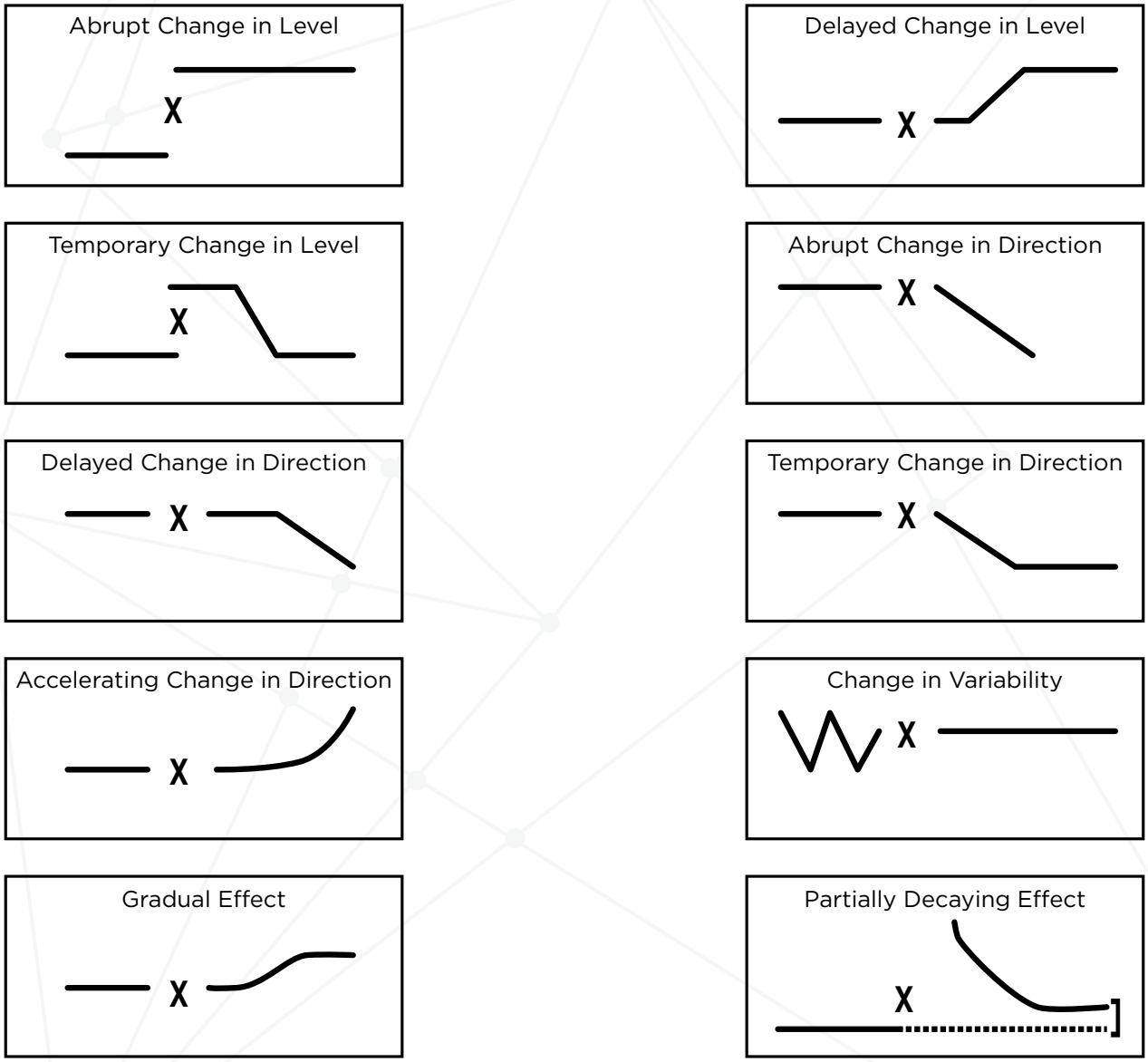
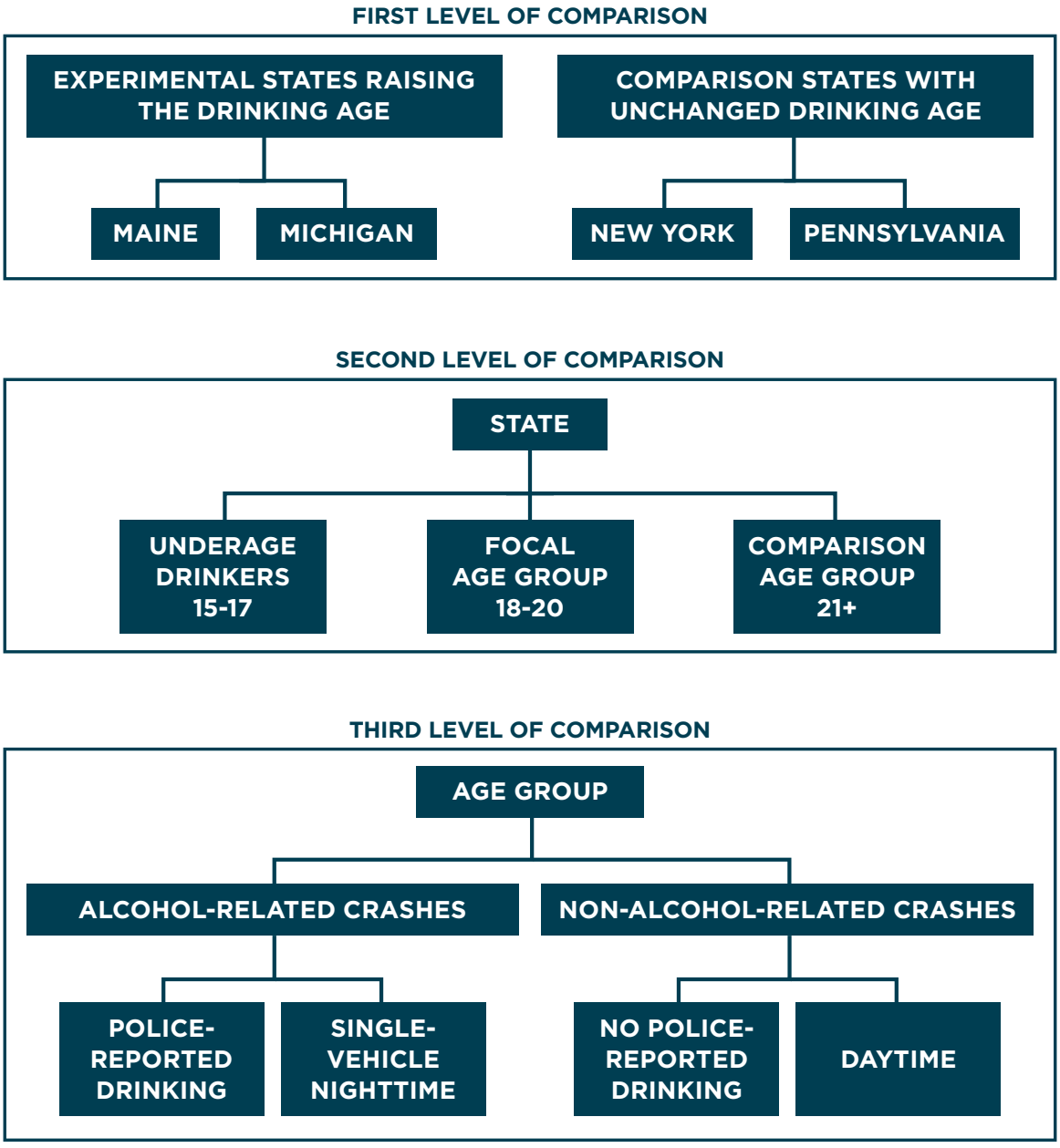


Figure 2. Possible Patterns of Policy Effects Over Time



Note: X = policy change. Source: Adapted from Wagenaar & Komro.[22]

Figure 3. Hierarchical Multi-level Time-series Design: Legal Drinking Age Example



Source: Adapted from Wagenaar & Komro.[22]

Affordable Care Act. Thus, understanding the effects of the Massachusetts law has important policy implications. Here we review two separate evaluations of the law's impact that illustrate the natural experiment design approach. Figure 4 is a simplified logic model for this reform. The primary goal was to make health insurance available to most residents of the Commonwealth. Insurance coverage, first of all, provides financial security. The link to health outcomes begins with an increase in the utilization of primary care services, which in turn would lead to an immediate decrease in hospital and emergency room (ER) utilization. Improved primary care access was also expected to eventually lead to a reduction in health care amenable mortality and improved health status. Health care costs and affordability was explicitly not a goal of the 2006 law, so there is no link to the affordability box in Figure 4.

Starting before the law was implemented, Long and colleagues[25] have been conducting an annual telephone survey of approximately 3,000 Massachusetts adults ages 19 to 64, the group most likely to be affected by the insurance expansion. Low-income areas and uninsured adults were oversampled. Based on this analysis, the researchers found that uninsurance rates remained low and access to health care, particularly primary care, was strong.[26] Figure 5 shows that the proportion of adults with any insurance rose immediately after the law went into effect and remained high. Primary care utilization, measured by the proportion of adults with a usual source of care showed a similar pattern. Emergency department visits and hospital inpatient stays (not shown) both declined starting in 2010, suggesting improvements in effectiveness of primary care in the year since the law was implemented. However, the affordability of health care (measured by the proportion who had problems paying bills), which was not a focus of the 2006 reform, remains an issue.

Noting these results, Sommers and colleagues[27] address the impact on population health, specifically to determine whether the Massachusetts reform was associated with changes in all-cause mortality and mortality from causes amenable to health care. Treating the reform as a quasi-experiment, the authors compared county-level mortality rates

Figure 4. Simplified Logic Model for Massachusetts Health Care Reform

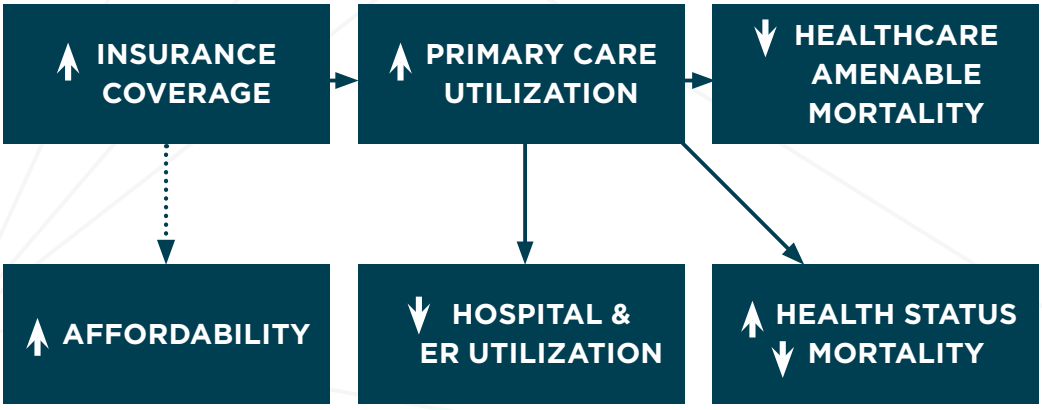
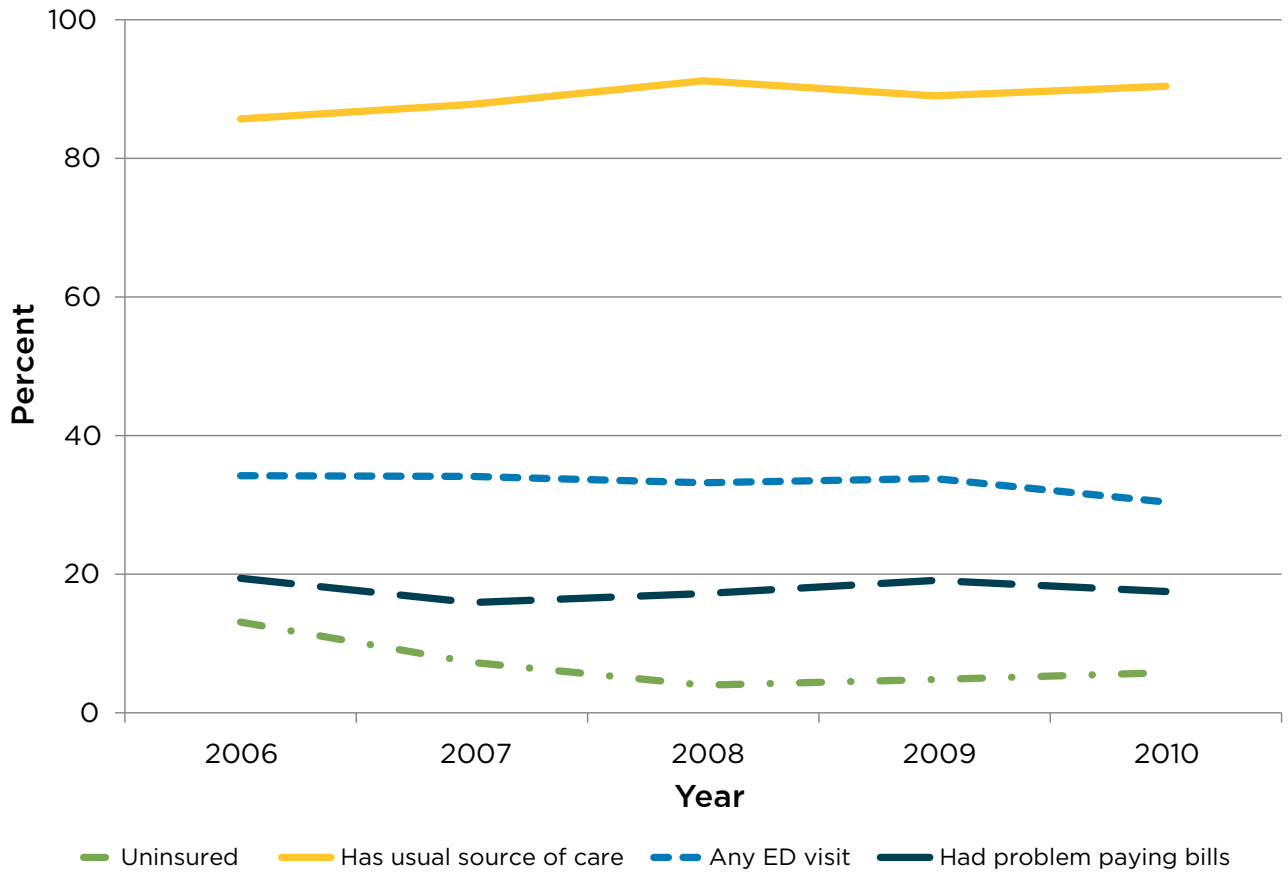


Figure 5. Impact of Massachusetts Health Care Reform on (a) Insurance, (b) Access to Primary Care (Usual Source of Care), (c) Emergency Department Use, and (d) Affordability (Problems Paying Bills)



Source: Adapted from Long and colleagues.[26]

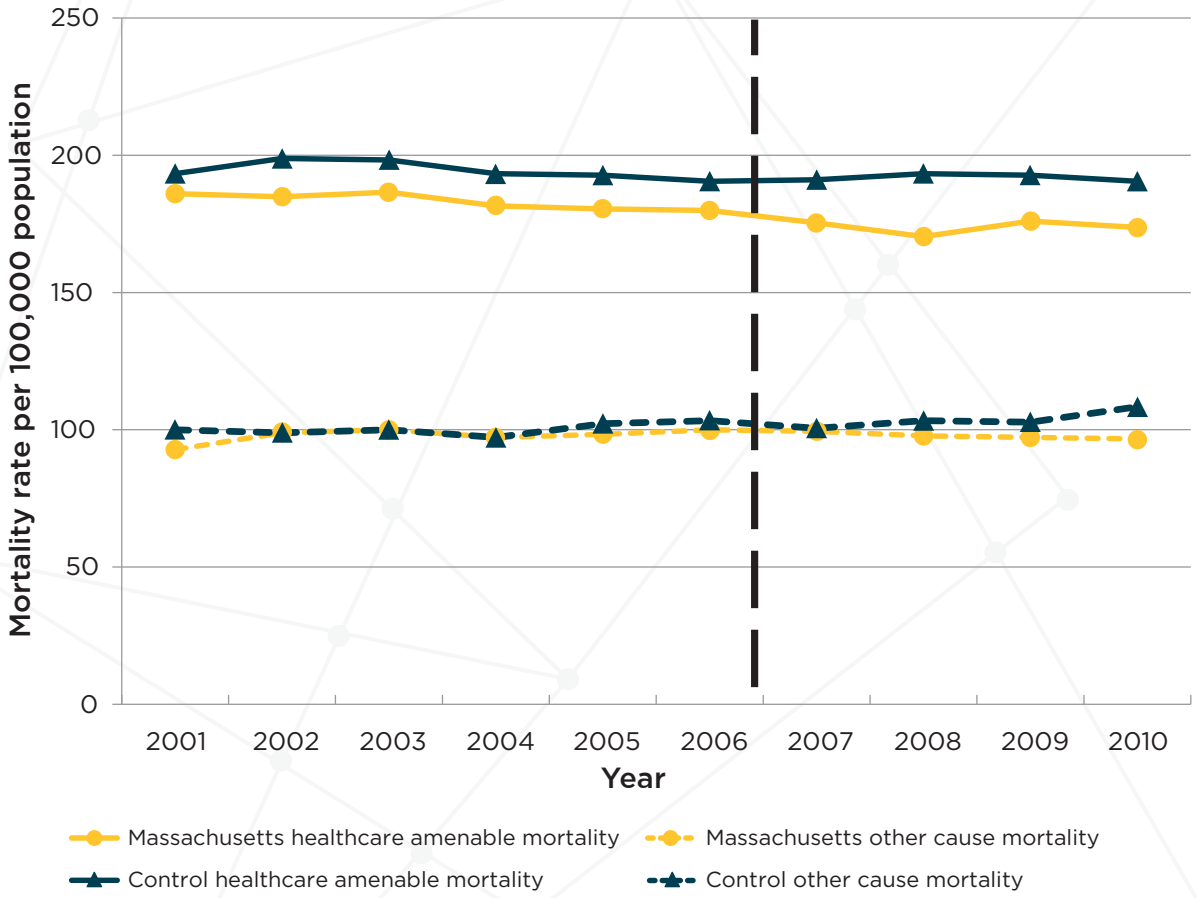
before and after reform in Massachusetts versus a propensity score matched control group of counties in other states using a difference-in-differences analysis. So for outcome variable X,

$$\text{Effect} = (X_{\text{Mass, after 2006}} - X_{\text{Mass, before 2006}}) - (X_{\text{control, after 2006}} - X_{\text{control, before 2006}})$$

The primary outcome was age-, sex-, and race-specific all-cause mortality, and deaths from causes amenable to health care was a secondary outcome.

Based on this analysis, Sommers and colleagues[28] found that reform in Massachusetts was associated with a significant decrease in all-cause mortality compared with the control group (-2.9 percent; P=0.003, or an absolute decrease of 8.2 deaths per 100 000 adults; see Figure 6 and Table 4). Deaths from causes amenable to health care also significantly decreased (-4.5 percent; P=0.001). In addition, the changes were larger in counties with lower household incomes and higher pre-reform

Figure 6. Unadjusted Mortality Rates for Adults Aged 20 to 64 Years in Massachusetts Versus Control Group (2001–2010)



Note: The vertical line designates the beginning of the Massachusetts state health care reform that was implemented starting in July 2006.
Source: Adapted from Sommers and colleagues.[27]

uninsured rates. They also found that reductions in mortality were largest in Massachusetts counties with lower incomes and lower insurance coverage before reform. Considered in conjunction with Long and colleagues'[29] demonstration of significant gains in coverage, access to care, and self-reported health, Sommers and colleagues[30] concluded that health reform in Massachusetts was associated with significant reductions in all-cause mortality and deaths from causes amenable to health care.

The authors note the limitations of the nonrandomized design and the possibility of unmeasured confounders, as well as the chance that the post-reform reduction in mortality in Massachusetts was due to other factors that differentially affected Massachusetts, such as the recession. As a result, they appropriately stop short of claiming definitive evidence of a causal relationship between the reform and the decline in mortality. Although the evidence is not definitive,

Table 4. Drop in Mortality After Massachusetts Health Care Reform Among Adults Aged 20 to 64 Years (2001-2010)

OUTCOME	UNADJUSTED MORTALITY PER 100,000 ADULTS		ADJUSTED RELATIVE CHANGE (POSTREFORM - PREREFORM)		
	PRE-REFORM	POST-REFORM	DIFFERENCE	95% CI	*P* VALUE
ALL-CAUSE MORTALITY					
Massachusetts	283	274	2.9	(4.8 – 1.0)	0.003
Control group	297	299			
HEALTH CARE-AMENABLE MORTALITY					
Massachusetts	185	175	4.3	(6.2 – 2.7)	<0.001
Control group	197	195			

Source: Adapted from Sommers and colleagues.[27]

they cite a number of factors deriving from the non-experimental design principles that strengthen the case:

- The analysis was based not just on single pre- and post-intervention measures but annual data going back as far as 2001 for the mortality data.
- The relevant outcomes were measured annually for four years after the law's implementation, permitting the authors to see first an increase in insurance coverage, second, improvement in primary care access, and finally, the impact on hospitalizations and health outcomes, as the theory of action would predict
- The analysis included multiple jurisdictions and comparison groups within Massachusetts and with other states. The analysis controlled for several distinct time- and county-specific economic measures by comparing each Massachusetts county to propensity score matched counties in other states. Within Massachusetts, the authors found that reductions in mortality were largest in counties with lower incomes and lower insurance coverage before reform, the areas likely to have had the greatest increase in access to care under reform.
- Within Massachusetts a comparison was made between outcomes expected to be affected by the intervention and similar outcomes not expected to be affected by the intervention under study. In particular, the larger proportional decrease in health care amenable mortality is consistent with the theory of action. The authors also found no evidence of a similar decline in mortality among elderly adults in Massachusetts, which would suggest a secular trend.
- Finally, the authors point out that it is challenging to identify factors other than health care reform that might have produced this pattern of results: a declining mortality rate since 2007 not present in similar counties elsewhere in the country, primarily for health care–amenable causes of death in adults aged 20 to 64 years (but not elderly adults), concentrated among poor and uninsured areas, and not explained by changes in poverty or unemployment rates.

Thus, while non-randomized studies can never definitively prove a cause and effect relationship, careful application of non-experimental design principles in this case make a very strong case. Such evidence may be "good enough" to inform decisions to replicate and/or spread health care interventions.

Planned Variation

When a researcher designs an experiment and allocates some subjects to treatment and other to control conditions, one result is a balance in the number of subjects in each exposure group. Making assignments by randomization helps to ensure that both observed and unobserved confounding variables are also balanced. But principles of experimental design can also help to achieve balance even without randomization. For instance, delayed-start and stepped-wedge designs,[31] in which some observational units are started on a treatment after others in a carefully planned way can be useful even if the assignment is not make at random. If attention is paid to unit characteristics thought to be associated with the outcome, researchers can at least help to control the effect of known confounders.

This "planned variation" approach can also be extended to more complex experimental designs. For instance, Zurovac and colleagues have proposed the use of multifactorial experiments paired with EHD to enable scientifically-rigorous testing of multiple facets of care provision at the same time in real-world settings where change is ongoing. This approach has the potential to help providers conduct rapid-cycle comparative effectiveness research and examine the impact of alternative ways of implementing care. Table 5 illustrates this approach in terms of assignments of alternatives that test ways of operationalizing care management. Note that although a full-factorial experiment exploring all combinations of the 4 care management factors would require 16 experimental units, in this example only 4 units are needed to provide estimates of each factor's impact as well as the first order interactions. The benefits of this design approach include statistical efficiency (minimizing the variance of parameter estimates) as well as the ability to test many facets of care provision simultaneously in real-world settings. Thus this approach combines the rigor of experimental design with the ability to produce results on the effectiveness of alternate approaches to multicomponent interventions in a single experiment. Although Zurovac and colleagues'[33] proposal assumes random assignment of experimental units (health care practices participating in the study) to specific combinations of intervention components, the principles of multifactorial design can be valuable even without randomization. These designs can also treat characteristics of units as factors to estimate their effect on outcomes as well as interactions with intervention components.

Conclusions

The primary strength of study design approaches described in this section is that they study the impact of a deliberate intervention in real-world settings, which is critical for external validity. The primary question addressed by evaluation designs is how to estimate the counterfactual – what would have happened if the intervention had not been implemented. At the individual level, epidemiologic designs focus on identifying situations in which bias is minimized. Natural and quasi-experiments focus on situations where the change in assignment breaks the usual links that could lead to confounding, reverse causation, and so forth. And because these observational studies typically use data gathered for patient management or administrative purposes, the possibility of observation bias is minimized. As a result, associations between the intervention and the outcome are more easily regarded as causal.

Table 5. Assignments of Alternatives That Test Ways of Operationalizing Care Management

CARE MANAGER	FREQUENCY OF ROUTINE CONTACT BETWEEN CARE MANAGER AND MEMBER	INVOLVEMENT OF A MEDICAL NURSE IN MANAGEMENT OF COMPLEX MEDICAL CASES	FOLLOW-UP DURING HOSPITAL ADMISSION AND AFTER DISCHARGE	BROWN BAG REVIEW OF MEDICATION*
1	a) Contact frequency based on member risk	b) Medical nurse is always involved	a) Current practice: care manager contacts member during the admission, conducts an in-person follow-up at discharge, and monitors as needed	a) No brown bag review of medication
2	b) More frequent contact (also based on member risk)	a) A medical nurse is involved as needed	b) Current practice, plus additional follow-up within a week of discharge, plus monitoring	b) Care manager performs a brown bag review for members with 4+ prescriptions
3	a) Contact frequency based on member risk	b) Medical nurse is always involved	b) Current practice, plus additional follow-up within a week of discharge, plus monitoring	a) No brown bag review of medication
4	b) More frequent contact (also based on member risk)	a) A medical nurse is involved as needed	a) Current practice: care manager contacts member during the admission, conducts an in-person follow-up at discharge, and monitors as needed	b) Care manager performs a brown bag review for members with 4+ prescriptions

Source: Adapted from Zurovac and colleagues.[32]

The disadvantages are that one cannot necessarily attribute the effect to the intervention (as opposed to other things that might have changed), and the results do not indicate what about the intervention made a difference.

Because they cannot rely on randomization to establish causality, program evaluation methods demand a more careful consideration of the "theory" of the intervention and how it is expected to play out. A logic model describing this theory can help to design appropriate comparisons, account for all influential variables in a model, and help to ensure that evaluation studies focus on the critical intermediate and long-term outcomes as well as possible confounders.

References

1. Stoto MA, Oakes M, Stuart EA, Stuart L, Priest E, Zurovac J. Analytical methods for a learning health system: 1. Framing the research question. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):28.
2. Stoto MA, Oakes M, Stuart EA, Stuart L, Priest E, Zurovac J. Analytical methods for a learning health system: 1. Framing the research question. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):28.
3. Stoto MA, Oakes M, Stuart EA, Brown R, Zurovac J, Priest E. Analytical methods for a learning health system: 3. Analysis of observational studies. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):30.
4. Stoto MA, Parry G, Savitz L. Analytical methods for a learning health system: 4. Delivery system science. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):31.
5. Stoto MA, Oakes M, Stuart EA, Brown R, Zurovac J, Priest E. Analytical methods for a learning health system: 3. Analysis of observational studies. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):30.
6. Cochran WG, Chambers SP. The planning of observational studies of human populations. J. R. Statist. Soc. A. 1965; 128(2): 234-266.
7. Stuart EA, DuGoff E, Abrams M, Salkever D, Steinwachs D. Estimating causal effects in observational studies using Electronic Health Data: challenges and (some) solutions. EGEMS (Wash DC). 2013; 1(3).
8. Stürmer T, Brookhart MA, Study Design Considerations. Chapter 2 in Velentgas P, et al., editors. Developing a protocol for observational comparative effectiveness research: a user's guide. AHRQ Publication No. 12(13)-EHC099 [Internet]. Rockville (MD): Agency for Healthcare Research and Quality; 2013 Jan [cited 2014 Dec 17]. Available from: http://www.effectivehealthcare.ahrq.gov/ehc/products/440/1166/User-Guide-to-Observational-CER-1-10-13.pdf
9. Stürmer T, Brookhart MA. Study design considerations [Internet]. Rockville (MD): Agency for Healthcare Research and Quality; 2013 Jan [cited 2014 Dec 17]. Available from: http://www.effectivehealthcare.ahrq.gov/ehc/assets/File/Ch_2-User-Guide-to-OCER_130129.pdf
10. Stürmer T, Brookhart MA, Study Design Considerations. Chapter 2 in Velentgas P, et al., editors. Developing a protocol for observational comparative effectiveness research: a user's guide. AHRQ Publication No. 12(13)-EHC099 [Internet]. Rockville (MD): Agency for Healthcare Research and Quality; 2013 Jan [cited 2014 Dec 17]. Available from: http://www.effectivehealthcare.ahrq.gov/ehc/products/440/1166/User-Guide-to-Observational-CER-1-10-13.pdf
11. Rothman KJ. Epidemiology: an introduction. 1st ed. New York: Oxford University Press; c2002. Chapter 4, Types of Epidemiologic Study; p. 57-93.
12. Rothman KJ. Epidemiology: an introduction. 1st ed. New York: Oxford University Press; c2002.
13. Stürmer T, Brookhart MA, Study Design Considerations. Chapter 2 in Velentgas P, et al., editors. Developing a protocol for observational comparative effectiveness research: a user's guide. AHRQ Publication No. 12(13)-EHC099 [Internet]. Rockville (MD): Agency for Healthcare Research and Quality; 2013 Jan [cited 2014 Dec 17]. Available from: http://www.effectivehealthcare.ahrq.gov/ehc/products/440/1166/User-Guide-to-Observational-CER-1-10-13.pdf
14. Rothman KJ. Epidemiology: an introduction. 1st ed. New York: Oxford University Press; c2002.
15. Glanz JM, Narwaney KJ, Newcomer SR, Daley MF, Hambidge SJ, Rowhani-Rahbar A, Lee GM, Nelson JC, Naleway AL, Nordin JD, Lugg MM, Weintraub ES. Association between undervaccination with diphtheria, tetanus toxoids, and acellular pertussis (DTaP) vaccine and risk of pertussis infection in children 3 to 36 months of age. JAMA Pediatr. 2013 Nov; 167(11): 1060-1064.
16. Glanz JM, Narwaney KJ, Newcomer SR, Daley MF, Hambidge SJ, Rowhani-Rahbar A, Lee GM, Nelson JC, Naleway AL, Nordin JD, Lugg MM, Weintraub ES. Association between undervaccination with diphtheria, tetanus toxoids, and acellular pertussis (DTaP) vaccine and risk of pertussis infection in children 3 to 36 months of age. JAMA Pediatr. 2013 Nov; 167(11): 1060-1064.
17. Rothman KJ. Epidemiology: an introduction. 1st ed. New York: Oxford University Press; c2002.
18. Yih WK, Lieu TA, Kulldorff MD, McMahill-Walraven CN, Platt R, Selvam N, Selvan M, Lee GM, Nguyen M. Intussusception risk after rotavirus vaccination in U.S. infants. N Engl J Med. 2014 Feb 6; 370(6): 503-512.
19. Yih WK, Lieu TA, Kulldorff MD, McMahill-Walraven CN, Platt R, Selvam N, Selvan M, Lee GM, Nguyen M. Intussusception risk after rotavirus vaccination in U.S. infants. N Engl J Med. 2014 Feb 6; 370(6): 503-512.
20. Baicker K, Taubman SL, Allen HL, Bernstein M, Gruber JH, Newhouse JP, Schneider EC, Wright BJ, Zaslavsky AM, Finkelstein AN, Oregon Health Study Group, Carlson M, Edlund T, Gallia C, Smith J. The Oregon experiment—effects of Medicaid on clinical outcomes. N Engl J Med. 2013 May 2; 368(18): 1713-1722.
21. Stoto MA, Oakes M, Stuart EA, Brown R, Zurovac J, Priest E. Analytical methods for a learning health system: 3. Analysis of observational studies. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):30.
22. Wagenaar AC, Komro KA. Natural experiments: design elements for optimal causal inference [Internet]. Philadelphia (PA): Public Health Law Research; 2011 Sept 1 [cited 2014 Dec 17]. Available from: http://publichealthlawresearch.org/resource/natural-experiments-design-elements-optimal-causal-inference
23. Stoto MA, Oakes M, Stuart EA, Brown R, Zurovac J, Priest E. Analytical methods for a learning health system: 3. Analysis of observational studies. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):30.

24. Wagenaar AC, Komro KA. Natural experiments: design elements for optimal causal inference [Internet]. Philadelphia (PA): Public Health Law Research; 2011 Sept 1 [cited 2014 Dec 17]. Available from: http://publichealthlawresearch.org/resource/natural-experiments-design-elements-optimal-causal-inference
25. Long SK, Stockley K, Dahlen H. Massachusetts health reforms: uninsurance remains low, self-reported health status improves as state prepares to tackle costs. Health Aff (Millwood). 2012 Feb; 31(2): 444-451.
26. Long SK, Stockley K, Dahlen H. Massachusetts health reforms: uninsurance remains low, self-reported health status improves as state prepares to tackle costs. Health Aff (Millwood). 2012 Feb; 31(2): 444-451.
27. Sommers BD, Long SK, Baicker K. Changes in mortality after Massachusetts health care reform: a quasi-experimental study. Ann Intern Med. 2014 May 6; 160(9): 585-593.
28. Sommers BD, Long SK, Baicker K. Changes in mortality after Massachusetts health care reform: a quasi-experimental study. Ann Intern Med. 2014 May 6; 160(9): 585-593.
29. Long SK, Stockley K, Dahlen H. Massachusetts health reforms: uninsurance remains low, self-reported health status improves as state prepares to tackle costs. Health Aff (Millwood). 2012 Feb; 31(2): 444-451.
30. Sommers BD, Long SK, Baicker K. Changes in mortality after Massachusetts health care reform: a quasi-experimental study. Ann Intern Med. 2014 May 6; 160(9): 585-593.
31. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol. 2006 Nov 8; 6: 54.
32. Zurovac J, Moreno L, Crosson J, Brown R, Schmitz R. Using multifactorial experiments for comparative effectiveness research in physician practices with electronic health records. EGEMS (Wash DC). 2013; 1(3).
33. Zurovac J, Moreno L, Crosson J, Brown R, Schmitz R. Using multifactorial experiments for comparative effectiveness research in physician practices with electronic health records. EGEMS (Wash DC). 2013; 1(3).