Simulated Power Analyses for Observational Studies: An Application to the Affordable Care Act Medicaid Expansion

Bernard Black, Alex Hollingsworth,

Letícia Nunes, and Kosali Simon*

March 25, 2021

Abstract

Power is an important factor in assessing the likely validity of a statistical estimate. An analysis with low power is unlikely to produce convincing evidence of a treatment effect even when one exists. Of greater concern, a statistically significant estimate from a low-powered analysis is likely to overstate the magnitude of the true effect size, often finding estimates of the wrong sign or that are several times too large. Yet statistical power is rarely reported in published economics work. This is in part because modern research designs are complex enough that power cannot always be easily ascertained using simple formulae. Power can also be difficult to estimate in observational settings where researchers may not know—and have no ability to manipulate—the true treatment effect or other parameters of interest. Using an applied example—the link between gaining health insurance and mortality—we conduct a simulated power analysis to outline the importance of power and ways to estimate power in complex research settings. We find that standard difference-in-differences and triple differences analyses of Medicaid expansions using county or state mortality data would need to induce reductions in population mortality of at least 2% to be well powered. While there is no single, correct method for conducting a simulated power analysis, our manuscript outlines decisions relevant for applied researchers interested in conducting simulations appropriate to other settings.

JEL: C15, I13,

Keywords: simulated power analysis, health insurance, mortality, Medicaid expansion

*Black: Pritzker Law School and Kellogg School of Management, Northwestern University, bblack@northwestern.edu. Hollingsworth: O'Neill School of Public and Environmental Affairs, Indiana University, hollinal@indiana.edu. Nunes: Institute for Health Policy Studies (IEPS), leticia.nunes@ieps.org.br. Simon: O'Neill School of Public and Environmental Affairs, Indiana University and National Bureau of Economic Research, simonkos@indiana.edu. We thank Marcus Dillender, William Evans, Ted Joyce, Anuj Gangopadhyaya, Robert Kaestner, Amanda Kowalski, Sarah Miller, Dan Sacks, Jeanette Samyn, Ben Sommers, Coady Wing, Susie Van Doren, Engy Ziedan, and participants in workshops at the Bar-Ilan University Conference on Law and Big Data (May 2018); Georgia State University, Economics Department (August 2018); Hebrew University of Jerusalem, Economics Department (Nov. 2018); Interdisciplinary Center Herzliya (May 2018); ASSA 2018; Stanford University Health Economics Seminar (May 2018); ASHEcon (2018); APPAM (2017); Tel Aviv University, Economics Department (Nov. 2018), Texas A&M (May 2017); and UK Institute for Fiscal Studies (Nov. 2018). We also thank the Editor and two anonymous reviewers. Code related to this project can be found at https://github.com/hollina/simulated-power-analysis

In a typical research setting, when an estimate is found to be statistically indistinguishable from zero, a researcher may conclude that a studied treatment has no apparent effect or that there is a wide confidence interval. Similarly, when the estimate is large enough to be deemed statistically different from zero, a researcher may report that there is a treatment effect, providing their point estimate as a best guess of its magnitude. However, without considering statistical power, both of these seemingly reasonable conclusions are incomplete.

Statistical power is the probability that an estimate will be found to be statistically significant when the true treatment effect is non-zero. A conventional measure of whether an analysis has sufficient power is whether the study—if repeated many times across different samples—would detect an effect at least 80% of the time using a two-sided test with a 5% threshold for significance.

Power is an important consideration whenever a treatment induces small changes in an outcome relative to its underlying variation. If the underlying variation is large, a small treatment effect may not be large enough to be considered sufficiently rare to be statistically significant. Thus, analyses with low statistical power are unlikely to produce convincing evidence of a treatment effect even when one exists. Of greater concern, a statistically significant estimate from a low-powered analysis is likely to overstate the magnitude of the true effect size, and can lead to estimates of the wrong sign or with magnitudes that are several times too large (Gelman and Carlin, 2014). This occurs because tail estimates that are much larger than the true effect are more likely to overcome underlying noise to be considered significant. Thus, without considering statistical power, it is difficult to interpret both insignificant and significant estimates.[1]

Despite this importance for inference, estimates of statistical power are rarely reported in published economics work. This is in part due to the difficulty in measuring power. For the simplest statistical models, power can be calculated using formulae and canned statistical software. However, as the complexity of a statistical model grows, the corresponding power formula must also grow more complex and a closed formula is not available for many contemporary research designs (e.g., synthetic controls). For example, only recently did Burlig et al. (2020) derive a power formula for the common experimental setting of difference-in-differences with serial correlation robust standard errors.

Power analyses are most commonly conducted when designing an intervention where the researcher has control over many parameters that affect power (e.g., size of treatment effect, number of treated and control subjects, and covariates collected). These settings are conducive to the use of simple formulae to estimate power before the intervention occurs. However, much of modern applied work studies observational settings, where the researchers may not know—and usually

[1]The potential for underpowered studies to produce misleading results is exacerbated by file-drawer bias, in which insignificant results are less likely to be published, and by non-pre-specified research designs, in which researchers may inadvertently explore different sample and regression specifications and are more likely to report those producing statistically significant results (Ioannidis, 2005).

have no ability to manipulate—the true treatment effect or other parameters important for power. Observational settings are less likely than experimental settings to have clean research designs, as concern for identification often requires a complex design for which a power formula may not exist.

Observational settings present other challenges for power as well. Consider the common research objective of estimating the effect of a geography and time based policy (e.g., a state policy passed in a certain year) that affects some individuals but not others. In many cases the researcher will only have access to aggregate population-level rather than the underlying individual-level data. Population aggregates may lump together treated and untreated individuals, groups, or time periods. If individuals within each aggregated unit differ in treatment status or there is heterogeneity in treatment intensity, aggregation will introduce noise that reduces statistical power. It does so by adding variation that any treatment effect would need to "overcome" to be statistically significant. This is particularly relevant in observational settings where researchers do not themselves assign treatment, but are instead reliant on available aggregate data—perhaps at resolutions too coarse to eliminate such measurement error in treatment.

To explore these issues, we examine an applied setting—the effect of gaining health insurance on mortality. Mortality is a rare event; even large changes in the probability of death—from changing insurance status or any other policy—may not be detectable at the population level. Most sources of U.S. mortality data do not contain information on the decedent's health insurance status or income. To overcome this data limitation, researchers often exploit policy changes that affect health insurance status at an aggregate level (e.g., state-by-year Medicaid expansion under the Affordable Care Act). Since treatment is not uniform across all individuals within a state (e.g., some already have health insurance) there is measurement error in treatment status. Recent research using detailed individual data and strong research designs has provided robust evidence that health insurance decreases mortality (Miller et al., Forthcoming; Goldin et al., 2021) In this paper we ask the question, can this true effect be detected using aggregate rather than individual data, and if so how large would the effect need to be to yield sufficient statistical power in the aggregate data setting?

The Affordable Care Act (ACA) substantially expanded health insurance coverage for the low-income, non-elderly adult population through both expanding Medicaid and creating marketplaces with private insurance subsidies. While every state created a subsidized marketplace, only some states chose to expand Medicaid eligibility. This setting provides an opportunity to study the link between health insurance and mortality, as well as to examine issues of statistical power in natural experiment studies of low-frequency outcomes. We focus on near-elderly (55-64) mortality, both because this age group is more likely than younger persons to have health conditions for which healthcare is important for survival, and because this provides a comparable group in the young-

elderly (65-74), who did not experience an insurance expansion. We consider two natural research designs. First a difference-in-differences (DD) comparison that estimates how the gap between the non-elderly mortality rate in Medicaid expansion versus non-expansion states changed before and after Medicaid expansion. Second, a triple-difference design that incorporates an additional comparison between the near-elderly and young-elderly, and can control for state-specific mortality trends, not captured in the DD design.

For each research design, our empirical strategy proceeds in three stages. First, we examine the reduced form relationship between Medicaid expansions and insurance coverage, which will be used to help interpret aggregate treatment effect estimates of mortality. Second, we obtain point estimates of the relationship between Medicaid expansion and mortality. Finally, we conduct a simulated power analysis, in which we impose known treatment effects of varying sizes and measure how power varies with the imposed effect, to demonstrate the smallest imposed effect size that could be detected with sufficient power using our data and research designs.

We illustrate the statistical challenges of studying our research question using this design in a way that is intended to be broadly applicable to other questions and designs. In particular, we use a simulation-based approach to obtain estimates of statistical power and minimum detectable effect sizes that can be adapted to a variety of settings. Traditional power analyses rely on closed form mathematical solutions and are only available for the simplest statistical models. Many modern, complex research designs have no corresponding power formula. In situations where a researcher does not have access to or cannot derive the closed form solution, they must choose between not conducting a power analysis, conducting a power analysis for a different, simpler model with the hope that the results are informative for the actual, more complex model, or performing a simulation. The steps to perform a simulated power analysis are not identical across every setting and the most appropriate simulation will be idiosyncratic for each project. A goal of this paper is to serve as a guide for other applied researchers seeking to conduct their own simulated power analysis by outlining how we determined and conducted a suitable simulation for our research design.

We conduct our simulated power analysis by artificially introducing treatment effects of different sizes into untreated, but real data, drawn from the pre-Medicaid expansion period. We do this many times, randomly assigning treatment at the state-level in each iteration. After imposing each pseudo-treatment effect, we run both the DD and triple difference specifications, recording the coefficient and p-value for the artificial treatment effect estimate. For each imposed effect size, we use this information to compute the percent of times a treatment effect is found to be statistically different from zero (i.e., power) and the minimum effect size that has 80% power at the 5% significance level. Conditional on an estimate being significant, we record the rate of severe magnitude errors (i.e., estimate more than twice the imposed effect size) and sign errors (i.e., estimate with

the wrong-sign). We also examine how power and the minimum detectable effect size vary as we perform the same analysis on subgroups of interest that may be more affected by the intervention. Finally, we demonstrate how particular simulation choices (e.g., the use of real vs. simulated data for this exercise, the geographic level of clustering, how treatment is assigned, and the length of the pre-treatment period) affect our power estimates.

There is no uniform method for conducting a power simulation. However, a key goal of this manuscript is to highlight issues that will appear in many efforts to use simulation to conduct a power analysis and to guide researchers in constructing analyses suitable for their research settings. Consider just a few decisions that a project may face: 1) the choice between using real, but untreated data, or entirely artificial data; 2) how to select untreated units to be pseudo-treated—should they be drawn from the pre-treatment period or from never-treated units? and 3) how to impose the treatment effect—should all units receive an identical treatment or should it follow some realistic structure? It general, it is important to understand how such decisions affect power estimates. When possible we present power estimates that show the effect of varying each of these choices.

In our research context, we find that Medicaid expansions resulted in average coverage gains of around 2% for those aged 55-64, with some subpopulations (e.g., Hispanics) seeing larger gains.[2] Using county-level specifications, we do not find a statistically significant relationship between insurance expansion and decreased mortality. However, with state-level data and a triple difference specification, we find evidence that healthcare amenable mortality decreased for the newly insured by around 1.53%.[3] Moreover, we find that standard difference-in-differences and triple differences analyses of Medicaid expansions using county or state mortality data would need to induce reductions in population mortality of at least 2% to be well powered. For example, power simulations for the state-level triple difference specification show that when the true, imposed treatment effect is 1.53%, statistical power at the 5% significance level is around 51.2%. We also show that when using our two research designs, focusing on sub-populations that were more affected by the treatment does not improve statistical power. This is because the increased variance in mortality for these groups offsets any power gains from greater treatment or a larger share of the population gaining insurance. Finally, we show that if individual-level data reported both mortality and insurance status, much smaller treatment effects (below 3% treatment on the treated) could be reliably detected with sufficient power.

[2]This estimate of expanded insurance coverage could be attenuated due to documented under-reporting of Medicaid enrollment in survey data and the availability of conditional/retroactive coverage (Davern et al., 2009; Marton and Yelowitz, 2015).

[3]The county-level and state-level specifications produce different estimates due to different regression weights. We use inverse propensity score weights so that weighted control observations will be similar to treated observations on observable covariates. Weighting for balance on state—rather than county-level characteristics—leads to different point estimates.

We note several caveats related to our work. First, our analysis is primarily focused on the subject of power analysis and does not examine whether insurance affects mortality; several recent studies answer this question (Goldin et al., 2021; Miller et al., Forthcoming). Second, we caution against the interpretation that a low-powered study contains no important information, instead we suggest that significant results from a low-powered analysis should be interpreted with more caution than those from an otherwise identical study with more power and that insignificant results from a low-powered analysis should not be interpreted as clear evidence of no effect. A collection of low or moderately-powered studies with different data and research designs may provide better collective evidence of an effect than a single well-powered analysis. Third, we also caution against using discrete thresholds to indicate whether or not a study has sufficient power. Estimates of statistical power are an important piece of information that can be used to evaluate the strength of an empirical result, but are only one metric. Results from a low-powered, but well-identified study may be preferred over a well-powered but unidentified result. Fourth, we do not recommend conducting post-hoc power analyses that use estimated treatment effects from regressions of the actual treatment; instead we recommend estimating power by perturbing real, but untreated data by imposing a known treatment effect.

Our paper proceeds as follows. Section 1 provides a background on statistical power, power in the economics literature, and an introduction to simulated power analyses. Section 2 contains all information unique to our idiosyncratic application: summarizing prior work examining the link between health insurance and mortality, the data used in our study, and our estimates linking Medicaid expansion to insurance enrollment and mortality. Section 3 outlines details of our simulated power analysis, results, and sensitivities to idiosyncratic choices. Section 4 concludes.

1 Background on statistical power

Power calculations can be useful in many settings. Discussions of power in economics typically relate to one of two broad goals. The first is to determine statistical power for literatures or specific questions of interest through the meta-analysis of already completed studies. The second—and the one most related to this manuscript—is to assess the statistical power of a specific research design for a given project. By definition, power for meta-analyses must occur after treatment has occurred. However, power for a specific study can be assessed before or after treatment occurs. The ability to conduct an *ex ante* power analysis can be valuable in planning a study design and these analyses are often used in the design of randomized controlled trials (RCTs).

In this section, we first document the rarity of power in published economics work. We then discuss attempts to estimate statistical power for studies within economics and outline differences between *ex ante* and *ex post* estimates of statistical power. Next, we explore how observational

data introduces challenges to using closed form formula for estimating power. Finally we discuss how simulated power analyses can alleviate some of these challenges.

1.1 Statistical power in the economics literature

Presence of power in published work: The simplest attempts to document concern for power involve simple keyword searches of published work. Examining a sample of empirical papers from the *American Economic Review* in the early 1980s, McCloskey (1985) finds that none of the papers mention the word power; examining all empirical papers published in top economic journals in the 1980s, McCloskey and Ziliak (1996) find that only 4% mention power and 1.1% examine the power function.[4] In a similar vein, we searched for the phrase "power analysis" in published and working economics papers and in pre-registrations. We found very few instances of the phrase, indicating that power analyses are not typically reported in published economics research.[5] This does not mean that economists do not conduct power analyses, as many grant applications require them, but merely that they are not reported in searchable text in the listed databases.

Concern for power in *collections* of economic studies: A lack of reported power has not prevented a number of efforts to estimate statistical power in economics. A growing literature across a number of fields, including economics, documents the frequent appearance of underpowered studies, as well as the potential causes and implications (Button et al., 2013; Maxwell, 2004; Ioannidis, 2005; Ioannidis et al., 2017).

Using data collected from top economic publications in the 1980s and 1990s, De Long and Lang (1992) argue that many null hypotheses are not rejected due to insufficient power. More recently, Ioannidis et al. (2017) use data extracted from meta-analyses to examine newer economics publications and estimate average power in the profession, both on the whole and in certain subfields. The authors report that the median statistical power in economics research is 18%, far from the 80% often used as a threshold for sufficient power. The authors determine power for each set of studies by comparing a weighted effect size from meta-analysis to a weighted standard error, with adequate power being when the effect is at least 2.8 times the magnitude of the standard error. While this is an appealing and intuitive approach to determining the power of a set of studies, it relies on the assumption that the weighted average effect size is close to the true effect size. This approach, as the authors point out, is not sufficient to determine if a single study is well-powered;

[4] Ziliak and McCloskey (2004) analyze the same journals for the 1990s and find that the percent of papers mentioning power rose to 8%.

[5] For papers released before January 2019, only 45 published and peer-reviewed papers and 39 working papers (in EconLit and NBER Working papers respectively) contain the phrase. Similarly, in the AEA RCT registry, 58 of the 4274 trials registered contained the search phrase "power analysis."

it can only be applied to sets of similar studies.[6]

There have also been thoughtful explorations of power within more narrowly defined questions. One example is Banerjee et al. (2015), who demonstrate that the literature evaluating the impact of microcredit suffers from low statistical power due to a limited take up rate, which is in part due to study design. For instance, Zhang and Ortmann (2013) identify median power across all papers using the classic experimental economics dictator game to be 25%. In addition, Gallet and Doucouliagos (2017) show that 59% of studies examining the impact of healthcare spending on life expectancy have adequate power. Both of these studies use methods similar to Ioannidis et al. (2017) to ascertain power for a set of similar studies.

1.2 *Ex ante*, *Ex post*, and *Post hoc* power

***Ex ante* power calculations:** Most power analyses are done *ex ante*, before conducting a study. This has become standard practice for many grant applications. For example, the National Institutes of Health (NIH) requires reviewers to evaluate statistical power and advises potential grant applicants to aim for studies with at least 80% power (NIH, 2016; Gerin et al., 2017).

For simple *ex ante* power calculations, it is common to used canned statistical software or closed form mathematical solutions that involve assumptions on a variety of parameters. In a simple comparison of means from two groups, a treatment effect needs to be 2.8 standard errors from zero in order to be detected with 80% power at the 5% significance level. For two samples of equal size and equal variance the minimum detectable effect size is $\frac{5.6\times \text{(s.d.)}}{\sqrt{N}}$ (Gelman and Hill, 2006). Thus, after making assumptions about the mean and distribution of an estimated treatment effect, a researcher designing an RCT could use a standard formula to estimate the minimum number of subjects needed to detect the effect at a 5% significance level with 80% power. This approach is helpful in ruling out poor study designs that are underpowered given realistic assumptions. Moreover, it allows the researcher to maximize power subject to realistic constraints (e.g., a budget) by manipulating the research design before treatment occurs. These manipulations could include altering the treatment effect size, reducing the number of participants to the minimum number needed for adequate power, or altering the length of treatment or control.

As the complexity of a statistical model grows, closed form power calculations also grow more complex and may not be derived for specific research designs (e.g., a synthetic control analysis). Consider Bertrand et al. (2004), who show that failure to account for serial correlation in a fixed-

[6] Another appealing feature of this approach is that conditional on distributional assumptions, 2.8 times the weighted standard error should be the minimum effect size that is detectable 80% of the time at the 5% significance level in the set of analyzed papers. This approach will not be sufficient for determining the minimum detectable effect size for any particular specification or for any set of papers that deviate from the assumptions of normality on coefficient estimates. For example, for non-parametric or structural estimates, this short cut for calculating the minimum detectable effect size will not work.

effects panel data setting can dramatically affect power and lead to an increased probability of a Type-I error. The authors recommend clustering standard errors as a solution, but do not provide a closed form solution for power in such a setting. Only recently did Burlig et al. (2020) derive a power formula for this common experimental setting of difference-in-differences with serial correlation robust standard errors.[7] Prior to this, a researcher wishing to use this experimental design would have been unable to use a power formula that correctly accounted for fully arbitraty serial correlation. In such situations, when a closed form power calculation does not exist or is not easily derivable, simulated power analyses are a viable alternative (Arnold et al., 2011; Burlig et al., 2020).

***Ex post* power calculations:** In addition to helping design studies before they are conducted, power analyses can also be calculated *ex post*—or after the treatment has occurred. *Ex post* power analyses are rarer in social science and in economics despite the fact that much of social science research is conducted after a treatment has already occurred and in situations where the researcher may have little opportunity to manipulate features such as sample size or treatment effect. The majority of studies that conduct power analyses after the fact estimate power for collections of papers. Importantly these studies all rely on comparing the standard errors of each study estimate to a proxy for the true value, not comparing the estimated treatment effect to the standard error– which would be equivalent to the t-statistic or p-value of the study. In principle the same methods used to determine *ex ante* power can also be used to estimate power *ex post*, however *ex post* power analyses are made more complex because the researcher does not necessarily know (or have control over) the true treatment effect.

Lewis and Rao (2015) provide an instructive example evaluating the power of an experiment *ex post*. They study the return on investment of twenty-five large scale online advertising campaigns using detailed micro data. They show that despite having millions of observations, most of the campaigns are significantly underpowered to detect plausible effect sizes.

They assume that sales variance and the number of observations are equal across treated and untreated groups. Using these assumptions they derive an R-squared and t-statistic formula as a function of the ratio of the effect size to the standard deviation (Cohen's D) and the sample size. The authors consider the "ideal scenario" of adding observations to their dataset. They then ask the simple question, how many of these additional observations need to be added to reach a t-statistic of 3? The critical value of 3 is chosen to have 90% power with a one sided p-value at the 5%

[7] Burlig et al. (2020) provide an excellent overview of ex ante and ex post power calculations. They derive an analytic formula for difference-in-difference settings that allows for arbitrary serial correlation. They demonstrate that failure to account for serial correlation can lead to a miscalculation of power and that this miscalculation is ambiguous in sign. Sometimes serial correlation can improve power, while often in longer panels it dampens power relative to the scenario of no serial correlation.

significance level. Using their assumptions, they show that essentially all of the experiments in question are underpowered and would require much larger sample sizes to detect plausible effect sizes.

Other existing experimental work uses a similar approach and reaches similar conclusions; despite sample sizes in the millions, many large scale advertising campaigns have low statistical power (Lewis and Reiley, 2014; Johnson et al., 2017). Our work differs from these studies since we do not make assumptions regarding the underlying variance and we opt to perturb existing data through simulation in a manner that will work for more complex, non-experimental research designs.

Concerns related to *post hoc* power analyses: We note that *ex ante* and *ex post* are relative to the time of treatment, not the time of initial empirical analysis. Conducting a power analysis *post hoc* or after the analysis has been conducted is a different matter.[8] Some have argued that *post hoc* power analysis should not be pursued (Hoenig and Heisey, 2001; Gouveia and Fletcher, 2000; Senn, 2002), citing concerns that power analyses would be used selectively, with researchers omitting the power analysis or arguing that a power analysis is unecessary after finding a significant effect, and only conducting a power analysis as justification for an insignificant finding. *Post hoc* analyses are problematic because conditional on finding a statistically significant effect, tests with low power have a higher likelihood of the estimated treatment effect being overstated in magnitude or having an incorrect sign relative to the true effect (Gelman and Carlin, 2014; Button et al., 2013). Thus, finding a treatment effect estimate at least 2.8 standard errors away from zero does not ensure sufficient power because *ex ante* power calculations require that the *true*, not the estimated, treatment effect is 2.8 standard errors away from zero.

Hoenig and Heisey (2001) similarly demonstrate that in a *post hoc* analysis there is a direct relationship between the estimated p-value of a test and the *post hoc* power estimate. Since p-values are estimates that are in part based upon the treatment effect estimate, insignificant findings will tend to appear as if they have lower power regardless of the true underlying power. Thus an insignificant finding cannot be used as evidence of low power when the true treatment effect is unknown.

We agree with these concerns; in both of these cases of *post hoc* analysis, a fundamental error occurs because the estimated treatment effect is being used to estimate power instead of the true treatment effect. This is not an issue when conducting *ex ante* power calculations since the true treatment effect is either assumed or known (e.g., an RCT where the researcher controls the treat-

[8] We recognize that the literal difference between the terms *ex post* and *post hoc* is minimal at best. We use the terms differently for two reasons. First, for clarity. *Ex* refers to pre/post treatment while *pre/post hoc* refers to pre/post analysis. Second, the use of the phrase *post hoc* has previously been used in the literature to discuss power analyses that use an estimated effect in the analysis (Gilbert and Prion, 2016).

ment). *Ex post* power analyses should be careful to avoid this conflation. That is, the estimated effect from a single analysis is not enough to determine if a study is well powered. In our empirical analysis, we minimize any possibility that our treatment effect estimate influences our power estimate by creating measures of power that do not use data from treated units following treatment and by imposing treatment effects of known size.

Power in observational settings Observational researchers often have no control over the sample size, treatment effect, or other features that may be malleable when conducting a randomized controlled trial or other intervention. Thus a researcher using a design for which there is no associated power formula is left with a dilemma. How to—without basing the power analysis on the estimated effect size—check the power of her study design to ensure 1) that significant effects are not of the wrong magnitude/sign and 2) that insignificant results are not a function of low power.

One solution is to evaluate power at a variety of effect sizes that the researcher believes are plausibly true (Arnold et al., 2011; Gelman and Carlin, 2014).[9] This range can be taken from the literature through a meta-analysis or asserted logically, but it should not be based upon the estimated value of the coefficient in the original regression. For our setting, a closed form power calculation does not exist, leaving simulation as the only viable option. Our approach evaluates power using observed, but pre-treatment data, for a variety of known and imposed effect sizes, which serve as a proxy for the true treatment effect. Before explaining this approach in more detail, we first outline two broad methods of estimating power through simulation.

1.3 Measuring power using simulation:

Simulation can involve either modifications of actual data, to which a treatment effect is added, or entirely artificial data (from an assumed data generating process). A study of bird nest visitation by Hannon et al. (1993), the earliest simulated power analysis we have found, is similar in spirit to our approach in that the authors apply a simulated treatment effect to actual data. The authors modify their outcome variable (nest visitation) using draws from the binomial distribution, gradually increasing (or decreasing) the probability of visitation. For each modified sample, they draw 50 bootstrapped samples, re-estimate their statistical model, and report power for each imposed effect size as the percentage of times the imposed effect is statistically significant among the bootstrapped samples.

In contrast, Hsiang et al. (2009) provide an example of estimating power using synthetic data.

[9] Burlig et al. (2020) also suggest estimating power using simulation for complex research designs/data generating processes. They provide a suggested plan and Stata package on how to approach this simulation in their Appendix Section D.2. Their suggested procedure has many appealing features, but does not apply exactly to as broad a range of research designs or use perturbed real, but untreated data as we do in our approach.

They generate the dependent variable (likelihood of conflict) using a normal distribution with a fixed mean and standard deviation; they impose a treatment effect by varying the mean to indicate a "treatment effect." For each imposed effect size, they analyze the synthetic data using their preferred specification and report power as the percent of times they find a statistically significant result at the 95% confidence level.[10]

An advantage of entirely synthetic data in a panel setting is that there will be no pre-treatment trends or treatment effect unless they are imposed. However, fully artificial data involves large sacrifices, similar to those noted for closed form power analyses by Burlig et al. (2020); one must implicitly impose structure on the variance-covariance matrix, for which the true structure may not be known. In panel data settings values could be correlated across time, pre-treatment trends could be non-parallel in complex ways, and unobserved covariates could predict both treatment and outcome. All of these complexities can affect statistical power. As Stigler (1977) points out, real data are rarely drawn from a "perfect distribution."

Our approach, applying a simulated treatment effect by modifying existing data during the pre-treatment period, does not guarantee a distribution centered around the null when we impose a zero treatment effect (the data can exhibit an "accidental" effect). But it preserves both the obvious and more subtle relationships present in the actual data that can affect power, and lets us take accidental effects into account when estimating power.

Another example of a simulated power analysis is Bertrand et al. (2004), who demonstrate that DD estimators in panel settings often suffer from autocorrelation. Similar to our setup, they estimate power with a single imposed effect size of 2%, finding that due to serial-correlation the null is incorrectly rejected two-thirds of the time when using simple OLS analyses.[11] These specifications in Bertrand et al. (2004) compare the relative gains in power from using different econometric specifications, with a consistent treatment effect. Our work differs from this since it is an inversion of the issue; we examine how power changes while changing the treatment effect, holding constant the econometric model. That is, we wish to know the smallest treatment effect size for which a certain power threshold is achieved given a certain research design.

Outline of our simulation approach Our contribution focuses on the value of conducting and reporting a power analysis in an observational study with a complex research design. We conduct a power analysis by artificially introducing treatment effects of different sizes into the data from the pre-treatment period, and then assessing how often our preferred analyses can detect these effects at the 90%, 95%, 99%, and 99.9% confidence levels (using two-tailed tests). Essentially we

[10] As another example of simulated power using entirely artificial data, Croke et al. (2016) examine a meta-analysis done by Taylor-Robinson et al. (2015) on the impact of administration of deworming drugs on childhood health. Croke et al. (2016) demonstrate that the meta-analysis is under-powered by using a simulation similar to Hsiang et al. (2009).

[11] Bertrand et al. (2004) examine the effect of placebo laws using state-level data on female wages.

perform a simulation in which we attempt to identify an imposed treatment effect while varying the treated units and the treatment effect size.

By modifying actual data, we preserve many of the complex relationships between and within variables that would be difficult to model credibly using artificial data. In our preferred analyses, we use pre-treatment data to avoid any possible influence of treatment in our setting. We use the results of our simulation to inform us about the research design and the power to detect true effects. This simulation based approach has the advantage that it can be applied to a wide-variety of research settings, including both structural and non-parametric work. One simply needs to run the same procedure on many versions of modified data, with known and varying treatment effects, storing the results each time. Following this Monte Carlo procedure, the stored results from each run are collectively analyzed to determine power at each significance level across varying sizes of the imposed treatment effect.

We also outline how we approach ad hoc choices that will likely be faced in most power simulations so that our study can serve as a guide to other applied researchers conducting simulated power analyses. While no uniform solution exists, we recommend outlining why certain choices were made and—when possible—demonstrating that other reasonable decisions do not materially affect power estimates. For example, we use real rather than simulated data in our study, but in another context, obtaining additional data may be impossible and thus generating artificial data may be the only viable path forward. We explore how changes in which, and how much untreated data are used (both across space and time), and whether the use of artificial data affects our power estimates. As another example, we strive to impose pseudo-treatments that are as similar as possible to how an actual treatment effect would be implemented. In our context this means that we do not simply shift the death rate of every pseudo-treated unit by an identical amount, but we randomly remove deaths probabilistically.

2 Applied example: Health insurance and mortality

Our illustrative example of the importance of statistical power, the effect of health insurance on mortality, fits into a large literature that examines the connection between health insurance and health status. This literature spans experimental and quasi-experimental settings, and examines morbidity and mortality, physical and mental health, elderly and non-elderly adults, pregnant women, children, infants, short- and long-run effects, and specific diseases and demographic subpopulations. Recent work using individual level data has provided strong evidence that health insurance lowers mortality (Miller et al., Forthcoming; Goldin et al., 2021). Our question is simply whether this effect can be detected using aggregate rather than individual data. Specifically we examine whether state-level Medicaid expansion is associated with lower healthcare amenable

mortality for the young elderly (i.e., those aged 55 to 64) using data aggregated at the county-year and state-year levels. In the next section we address the question of the smallest effect size that could be detected with sufficient statistical power.

2.1 Prior work on effect of health insurance on health and mortality

Historically, the first rigorous evidence on how health insurance affects health and mortality comes from the RAND Health Insurance Experiment (HIE) (Brook et al., 1983; Keeler et al., 1985; Newhouse, 1993), which provided experimental exposure to varying degrees of insurance generosity; none of the study subjects was fully uninsured. Brook et al. (1983) found no significant overall effect on mortality for the full sample—point estimate -0.02; 95% CI [-0.05, +0.02] for persons aged 14 to 61, followed for 3-5 years—but found 10% lower mortality for high-risk individuals who received generous insurance. The RAND HIE also found some improvements in blood pressure for low-income populations receiving generous insurance, but otherwise found limited evidence that generous insurance led to improved health.

Finkelstein and McKnight (2008) study Medicare's introduction in 1965, which remains the largest health insurance policy change in US history. They find a large increase in the insurance rate of around 75%, because private insurance for the elderly was uncommon before Medicare (Finkelstein, 2007). They find a 40% drop in out-of-pocket medical expenditures, but no discernible mortality effects over a 10-year period (point estimate after 5 years = -0.15%; 95% CI [-3.9%, +3.6%]). They observe that these results may be driven by the fact that prior to Medicare, those with life-threatening but treatable conditions likely sought care even if they were uninsured.

Card et al. (2004) exploit the age-65 discontinuity in Medicare coverage using more recent data from 1989-1998; they find no significant effect of turning 65 on population mortality (point estimate +0.5%, 95% CI [-3.3%, +4.3%]). They find an increase in the rate of insurance coverage of approximately 8% for the full sample and 14% for a low-education subsample. In a related study that speaks to possible mechanisms, Card et al. (2009) find a drop in mortality at age 65 among those admitted to hospital through the emergency department—for severe, non-deferrable reasons for which most individuals would seek emergency department care whether insured or not. They find that having insurance through Medicare increases treatment intensity by around 3% and results in a 1% absolute (20% relative) reduction in 7-day mortality and a 3% relative reduction in 1-year mortality.

Doyle (2005) studies a subpopulation with strong need for emergency medical care—victims of auto accidents who are alive when they reach the hospital—and finds higher adult mortality rates for uninsured persons in Wisconsin during 1992-1997. He finds that being uninsured increases in-hospital mortality by 39%, relative to other auto accident victims—1.5 more deaths per 100 (95%

CI [0.3, 2.7]), relative to a mean of 3.8 deaths per 100. He attributes this finding to differences in treatment intensity, rather than pre-accident differences in health.

Levy and Meltzer (2004, 2008) review the literature and conclude that, consistent with Finkelstein and McKnight (2008) and Card et al. (2004), there is at most modest evidence of some health benefits from general adult health insurance expansions. They note potential exceptions for specific vulnerable populations, but conclude that "for most of the population at risk of being uninsured (adults of ages 19 to 50), we have limited reliable evidence on how health insurance affects health" (Levy and Meltzer, 2008, p.404).

In addition to the RAND Experiment, three other randomized experiments deserve attention. Weathers and Stegman (2012) find no significant mortality effect for adults receiving Social Security Disability Insurance who receive Medicare immediately rather than after the usual 2-year waiting period (point estimate for odds ratio 1.28, 95% CI [0.71,1.85]. They do find that those receiving insurance have better self-reported health. The second is the Oregon Experiment, involving Medicaid expansion for adults, administered through a lottery among those who applied. Finkelstein et al. (2012) and Baicker et al. (2013) find limited changes in mortality or measures of physical health after 2 years. They find increased healthcare use, increased diabetes detection and care (but not lower blood sugar levels), reduced financial strain, and less depression. Their estimated increase in health insurance coverage is large, indicating around a 25% relative rise in coverage for those in the treatment group, but shrinks rapidly and is only half as large after 16 months. Their point estimate for mortality reduction is large, at -13%, but with a wide 95% CI [-26%, +13%]. These first two experiments find statistically insignificant effects for relatively vulnerable populations (the disabled for Weathers and Stegman (2012), and poor adults who signed up for the Medicaid lottery and enrolled if eligible for the Oregon Experiment).

In contrast, the third experiment in the literature finds large and significant effects of insurance on mortality. Goldin et al. (2021) randomly notified 3.9 million households—which owed an Affordable Care Act induced penalty for failure to have health insurance—of their tax burden and instructions on how to sign up for health insurance. This nudge increased health insurance enrollment by 1.9% relative to a control group. This increase in health insurance reduced mortality of 45-64 year olds by 0.06 percentage points (95% CI[-0.112 to -0.014]) relative to a mean of 1 percent.

In a similar vein as Goldin et al. (2021), several recent—but, non-experimental—papers on insurance expansions for nonelderly adults find large effects of health insurance on mortality rates. Sommers et al. (2012) considers Medicaid expansion for non-elderly adults in three states (Arizona, Maine, and New York) that expanded Medicaid in the early 2000s compared to neighboring non-expansion states; Sommers et al. (2014) and Powell (2018) consider the Massachusetts insurance expansion in 2006. McClellan (2017) considers the ACA mandate that requires employers

to cover young adults under their parents' employment-based insurance policies until age 26, and Dunn and Shapiro (2019) consider the effect of Medicare Part D prescription drug coverage for elderly adults. Using a weighted match and trimmed sample, Borgschulte and Vogler (2020) find that Medicaid expansions are associated with a 3.6 percent decrease in mortality for those aged 20 to 64.

Although not an RCT, recent work by Miller et al. (Forthcoming) uses large samples of individual-level data to study the impact of ACA Medicaid expansions on insurance rates. They find a 9.4 percent reduction in mean elderly mortality associated with Medicaid expansion. Along with Goldin et al. (2021), this work provides the strongest evidence that health insurance can affect mortality, with both studies leveraging large-scale individual level data, known treatment, and strong research designs to demonstrate a clear link between health insurance and mortality.

2.2 Conceptual Concerns

Several concepts inform our power analysis and the interpretation of our mortality results. One is the existence of prior policies that provide vulnerable populations with health insurance, or with healthcare regardless of health insurance status. These include existing Medicare or Medicaid avenues to health insurance and healthcare for the disabled persons in the 50+ age range we study; emergency care through the Emergency Medical Treatment and Active Labor Act (EMTALA); coverage for persons with specific high-cost health conditions (e.g., AIDS through the Ryan White Act and end-stage renal disease under Medicare); coverage for those who suffer workplace or automobile injuries; and healthcare for those with access to public hospitals, publicly supported clinics, or the charity care provided by nonprofit hospitals). Thus, health insurance expansions will affect principally populations and medical conditions outside these groups. This issue is exacerbated in our setting since we do not have detailed data on subgroups or individuals that experienced meaningful increases in health care access due to expanded insurance. Our analysis examines treatment effects at the county age-group level, despite treatment being at the individual level.

A second concept that informs our analysis is selection into coverage for a new program, such as the ACA Medicaid expansion, including selection effects for both take-up of new coverage and crowd-out of other coverage. The less policymakers are practically or politically able to target groups likely to be uninsured and promote a high take-up rate, the less likely it is that studies like ours will have sufficient statistical power to find detectable effects on health or mortality. For example, the ACA changes eligibility but does not directly provide insurance. As in any "intent-to-treat" (encouragement) design, we can estimate a treatment effect only for the "compliers" with the encouragement. Multiple selection effects are possible, including that those who sign up: (i) may be more health-conscious in other ways; (ii) may have greater healthcare needs (e.g., Ken-

ney et al., 2012); (iii) may be more likely to use additional healthcare once insured; (iv) may be more compliant with medical advice than the "never-takers" who do not sign up; and (v) some Medicaid-eligible persons may wait to enroll until care is needed for a catastrophic event creating adverse selection (Marton and Yelowitz, 2015). Thus, estimates for compliers may differ from those for never takers or always takers (the already insured). For example, Kowalski (2020) reconciles differences in the effects of the Oregon experiment and the Massachusetts health insurance expansion on emergency department visits on the basis of better initial health for the Massachusetts compliers.

Third, there could be substantial treatment heterogeneity even among the compliers, with health insurance improving health for some, but being neutral for others ("flat of the marginal benefit curve" medicine) or even detrimental due to overtreatment (e.g., opioid addiction as an unintended possible effect of pain treatment). Yet the available data limits our ability to study specific subpopulations. A fourth concern is heterogeneous health insurance quality. In many states, Medicaid insurance is considered to be of lower quality than commercial insurance (Polsky et al., 2015). Fifth, health insurance is only one factor potentially affecting trends in health and mortality. Other factors can vary by age and ethnic group (e.g., Case and Deaton (2015) find rising mortality in middle-age for less-educated whites, but not other groups), and by state. Differing trends can complicate efforts to define a suitable control group.

These concerns collectively highlight the complex relationship between health insurance and health outcomes, and anticipate the limitations of the available data and policy shocks.

2.3 ACA Insurance Expansions and Identifying Variation

In 2014, the two main insurance expansions under the ACA took place, with Medicaid expansions occurring in 27 states (including the District of Columbia) on or soon after January 1, 2014; expansions took place in three more states in late 2014 or soon after January 1, 2015, and then in two more in late 2015 or the beginning of 2016. Standard expansion included coverage for all non-elderly adults with family income less than 138% of the federal poverty level (FPL). Of these 32 expansion states, 10 had conducted significant expansions prior to 2014 and are not included in our main specifications (following other studies on Medicaid expansion, e.g.,Wherry and Miller (2016)).[12] The treated states for our principal analyses are the remaining 22 "Full Expansion States;" the control group consists of the 19 "Non-Expansion States." Table A1 lists the states in each expansion group, as well as the change in percent uninsured in each state from 2013 to 2016 for persons between the ages of 18 and 64.

[12]Louisiana expanded Medicaid in mid-2016. We consider the first expansion year for Louisiana to be in 2017, which occurs after the last year of data used in our analyses.

The second major way in which the ACA expanded coverage was by creating marketplaces with private insurance subsidies for those with low income. Although our study design exploits variation in Medicaid expansion, Table A1 shows that the uninsured population fell in both Expansion and Non-Expansion States due to these other aspects of the ACA. Subsidized marketplace insurance was available for persons with incomes from 100-400% of FPL in non-expansion States, and thus to some extent substituted for Medicaid expansion; this reduces the relative effect of Medicaid expansion on insurance rates between expansion and non-expansion states, and thus the first-stage for our study.

2.4 Data

We measure mortality using the confidential version of the Compressed Mortality File (CMF), which contains individual death records for approximately 2.5 million deaths a year. This dataset is compiled by the National Center for Health Statistics. The data in the mortality files include (1) race, ethnicity, and gender; (2) year of death; (3) age at death (which we collapse into 10 year-age groups, e.g., 55-64); and 4) primary cause of death. In our preferred analysis we use ten years of pre-treatment data (from 2004 to 2013) and three years of post-treatment data (from 2014 to 2016). We conduct county and state-level analyses, using population (from the National Cancer Institute's SEER) and inverse propensity score weights, which more heavily weight untreated units that closely match treated units and that are representative of the respective populations.

To examine the impact of ACA expansion on health insurance estimates—our first-stage—we use information on uninsurance rates from both the Census Bureau's Small Area Health Insurance Estimates (SAHIE) and from the American Community Survey (ACS). These data help place our mortality estimates into better context. The periods used in SAHIE and ACS were 2006-2016 and 2008-2016, respectively. We analyze ACS data at the state level, which allows us to examine results by gender, race, ethnicity, education and income. SAHIE estimates, on the other hand, are available both at the county and state level, but with only gender and income subgroups. We estimate the first stage using two data sources since each source offers its own idiosyncratic advantages. The ACS data have the advantages of allowing for a direct comparison with our population of interest by age (i.e., those aged 55-64 versus those aged 65-74 as a control group), which enables us to perform the same DD and triple difference analyses for the first stage as we conduct using the mortality data. The SAHIE data have the advantage of providing county level insurance estimates as well as a longer time horizon than the ACS—with usable data beginning in 2006 rather than 2008.

2.5 Empirical Approach

To investigate the effect of Medicaid expansion on mortality, we use several DD specifications: (i) a "simple DD" specification, which assumes a one-time change in mortality rates; (ii) an event-study or "leads-and-lags" model, which allows for a separate treatment effect by years since expansion, and lets us assess whether pre-treatment trends are parallel; and (iii) a "triple difference" model, in which the third difference is persons aged 55-64 versus persons in the same county or state aged 65-74. Treatment is recorded in event time, relative to the year in which each expansion state expanded Medicaid. Our preferred specifications use county-year level data, county and year fixed effects (FE), inverse propensity score based population weights, standard errors clustered at the state level, and data from 2004 through 2016, for the sample of 22 full-expansion and 19 non-expansion states.

The simple county-level DD model is:

$$Y_{it} = \alpha + \beta D_{st} + \delta X_{jt} + \tau_t + \vartheta_j + \varepsilon_{jt} \tag{1}$$

Here, j indexes county; s indexes state; t indexes time in years, the dependent variable; Y_{jt} is ln((deaths/100,000 persons)+1); we add 1 to the mortality rate to avoid dropping county-years with zero deaths. We limit the sample to Full- and Non-Expansion States to form a stronger comparison. The predictor variable of interest is $D = 1$ for Full Expansion States in post-expansion years (2014-2016 for the 17 states that fully expanded Medicaid in 2014; 2015 and 2016 for the 3 states that expanded in 2015; and 2016 for the 2 states that expanded in 2016). The covariate vector X_{jt} includes the following county-level variables: managed care penetration (Medicare Advantage beneficiaries as % of all Medicare beneficiaries); % disabled (% of Medicare beneficiaries receiving SSDI benefits); % in poverty; unemployment rate; median household income; mean per-capita income; % with diabetes; % obese; % physically inactive; % smokers; and active practicing non-federal physicians/1,000 persons. We convert all amounts to 2010 dollars. In some specifications, we use a narrower set of covariates or no covariates, partly to assess whether our results are sensitive to including observable, time-varying, county-level factors, and also because expansion could affect some covariates. We include county and year fixed-effects in all models to control for potential unobserved covariates that vary across counties but are fixed over time, and for determinants of mortality that are constant across counties but vary over time. Standard errors are clustered at the state level.

To address potential differences in control and treatment states, we implement an inverse propensity score weighting approach in which we compute average treatment on the treated (ATT) weights that also reflect differential population. To generate the ATT weights, we first average covariates in each unit over the pre-treatment period (2004 to 2013 in our preferred analysis).[13]

[13]We use the following list of covariates: % uninsured under 138% federal poverty line, % uninsured 50-64, total

We then run a logistic regression, predicting whether a county has full-expansion or non-expansion status. We generate the fitted propensities p for each untreated unit and calculate ATT weights as $(p/(1-p))$; ATT weights for treated units are set to one. The inverse propensity score weights are windsorized at the 95^{th} percentile to avoid assigning extremely high weights to a small number of control counties.[14] We interact these inverse propensity score weights by population so that our results better reflect the treatment effect for the average person rather than the average county. This procedure more heavily weights non-expansion counties that look like expansion counties.

We principally study mortality due to healthcare-amenable causes (Nolte and McKee, 2003), but we also provide some estimates for non-amenable and total mortality. The concept of amenable mortality seeks to capture deaths from conditions that are potentially preventable with timely care; including mortality related to HIV, cardiac, diabetes, and respiratory causes. Studying non-amenable mortality has value as a placebo check—any effect of health insurance expansion should be weaker, or absent entirely, for non-amenable mortality.

To study the time pattern of any apparent treatment effect, and to assess whether pre-treatment trends differ between Full- and Non-Expansion States, we use a leads-and-lags model in event time, with the first expansion year set to zero, following Equation (2):

$$Y_{jt} = \alpha + \sum_{k=-10}^{2} (\beta_k D_{jt}^k) + \delta X_{jt} + \tau_t + \vartheta_j + \varepsilon_{jt} \tag{2}$$

Here, k indexes "event time" in years relative to Medicaid expansion. $D_{jk} = 0$ for Non-Expansion States for all j and k. For Full-Expansion States, D_{jk} = 1 for the k^{th} year relative to the adoption year, and 0 otherwise. Thus, β_1 provides the estimated population average treatment effect for the first expansion year, while β_{-1} is the estimated effect one year before adoption, and so on. All coefficient estimates are reported relative to the difference between expansion and non-expansion counties in the year before adoption ($k = -1$). The true identifying assumption is that there would be parallel trends in the post-treatment period in the absence of treatment. While this assumption is not directly testable, the evidence of parallel trends in the pre-treatment period is suggestive that this identifying assumption would hold so long as we use the appropriate weights in our analysis.

In addition, we use a further source of within-state variation: mortality trends among those who are 65 or older (and thus always insured) can potentially control for the otherwise unobserved state-specific factors that generate non-parallel trends. We thus also use a triple-difference specification,

% uninsured, per capital income, median household income, % in poverty, % Medicare Advantage penetration, % Medicare beneficiaries receiving disability, % disabled, % obese, % daily-smokers, % occasional smokers, % former smoker, active MDs per 1000, population that is male, white non-Hispanic, Black non-Hispanic, aged 55-64, and aged 65-74.

[14] Our main results are not sensitive to any reasonable amount of windsorization. We display sensitivity across a range of choices in Figure A2.

where the third difference is mortality among persons between the ages of 65 and 74, who are eligible for Medicare and should not be affected by Medicaid expansion; we limit the sample to persons between the ages of 55 and 74, thus comparing mortality trends for the 55-64 and 65-74 age groups. This specification implicitly controls for all county-year unobservables that equally affect death rates for both age groups. The triple-difference specification, analogous to the simple DD, is:

$$Y_{jt} = \alpha + \beta D_{st} \times Under65_{jt} + \gamma D_{st} + \theta Under65_{jt} + \delta X_{jt} + \tau_t + \vartheta_j + \varepsilon_{jt} \quad (3)$$

We also estimate separate models for subsamples stratified on covariates that may predict uninsurance rates or response to health insurance, for which we also have mortality data: cause of death, gender, and race/ethnicity.

2.6 Results

We present full-sample results in this section for adults aged 55-64. We first present univariate results, and then results from DD and triple difference models. In the Appendix, we assess whether we could obtain a better match between treated and control states, and thus tighter confidence bounds. We conclude that we cannot rely on these methods for inference due to poor pre-treatment fit.

2.6.1 Univariate Graphical Evidence

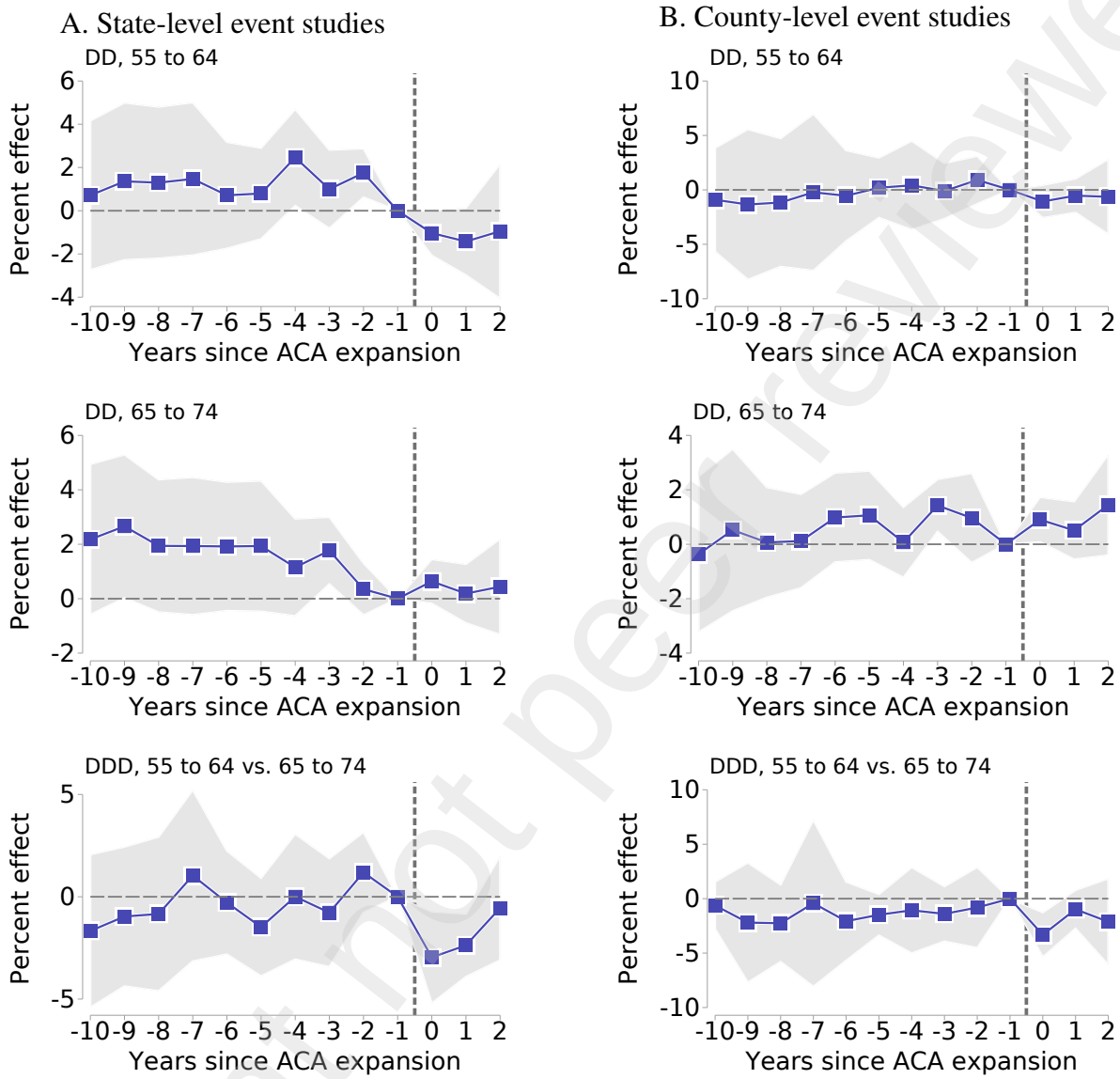
In Figure A3, we display trends in the amenable mortality rate for the four state groups, for the full time period with available data (2004-2016) in event time. For those states that did not expand Medicaid, time is measured relative to 2014 (i.e., at $t = 0$ the data come from 2014). Several features of Figure A3 are important. First, there are substantial level differences in mortality rates across the state groups, although these are smaller between our principal comparison groups—the Full-Expansion vs. Non-Expansion States. Second, the trends of mortality across full and non-expansion states are relatively stable across time, even following ACA adoption.

2.6.2 Event Studies

To investigate these differences more formally, we next turn to an event study analysis, using Equation 2 for our DD specifications and an analogous specification for the triple difference analysis. Figure 1 provides, in its first-row, event study estimates and 95% confidence intervals using data from 2004 to 2016 at the state (panel A) and county (panel B) levels, for amenable mortality among persons aged 55-64. For the county-level specifications—consistent with the trends observed in the raw data—there is no evidence of a change in relative mortality in the first three expansion years.

The same conclusion can be reached for the 65-74 age group in the second-row of Figure 1. The DD event-studies at the state level show declines in mortality for the young elderly and elderly that begin in the year before expansion. For the state-level data and the triple difference specification, we see both flat pre-trends and a decline in mortality in the first two years after the ACA expansion.

Figure 1: Event study estimates showing the effect of Medicaid expansion on amenable mortality



Note: Number along the x-axis indicates years since most Medicaid expansion. For non-expansion states, $t = 0$ corresponds to 2014. Years included in this figure are 2004 to 2016. Since each dependent variable is a natural log transform of the healthcare amendable death rate per 100,000, we transform each regression coefficient by 100(exp(coef.)-1) and the standard errors using the delta method. Thus each reported coefficient can be interpreted as the percent effect of Medicaid expansion t years ago on the healthcare amenable mortality rate for a given age group (i.e., A coefficient of -2 would mean that the mortality rate would decrease by 2%). Point estimates are depicted by blue squares and come from regressions analogous to Equation 2. 95% confidence intervals are displayed by gray area and are calculated using robust standard errors clustered at the state level.

Table 1: The effect of Medicaid expansion on mortality by age group and specification

	State-level						County-level					
	DD 55-64 years		DD 65-74 years		DDD		DD 55-64 years		DD 65-74 years		DDD	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Healthcare amenable mortality												
Full expansion dummy	-3.23**	-2.32**	-1.50	-0.98	-1.63	-0.88	-0.29	-0.65	0.85	0.41	0.61	0.25
	(1.27)	(1.01)	(1.22)	(0.89)	(1.18)	(0.90)	(2.00)	(1.74)	(1.21)	(0.80)	(1.19)	(0.88)
Full expansion dummy x Age 55-64 dummy					-1.55**	-1.53*					-0.83	-0.85
					(0.76)	(0.77)					(0.94)	(0.94)
Non-amenable mortality												
Full expansion dummy	0.93	1.45	2.95**	3.32***	2.76**	3.23***	3.03	2.60	1.63	0.90	1.45	0.90
	(1.36)	(1.35)	(1.24)	(0.76)	(1.21)	(0.94)	(2.59)	(2.36)	(1.32)	(1.12)	(1.31)	(1.20)
Full expansion dummy x age 55-64 dummy					-1.73	-1.73					1.56	1.58
					(1.11)	(1.12)					(1.79)	(1.79)
All Mortality												
Full expansion dummy	-2.27*	-1.27	-0.75	-0.18	-0.88	-0.06	0.60	0.28	1.03	0.57	0.81	0.46
	(1.18)	(0.97)	(1.12)	(0.77)	(1.08)	(0.81)	(1.89)	(1.70)	(1.06)	(0.68)	(1.03)	(0.78)
Full expansion dummy x Age 55-64 dummy					-1.35*	-1.32*					-0.17	-0.18
					(0.72)	(0.72)					(1.00)	(1.01)
Weights	ATTxPop	ATTxPop	ATTxPop	ATTxPop	ATTxPop	ATTxPop	ATTxPop	ATTxPop	ATTxPop	ATTxPop	ATTxPop	ATTxPop
Unit fixed-effects (state or county)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed-effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	533	533	533	533	1066	1066	36478	36478	36478	36478	72956	72956

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors clustered at the state level in parentheses. Each DD regression corresponds to Equation 1 while each triple-difference regression corresponds to Equation 3, with any deviations noted. Years included in this table are 2004 to 2016. Each dependent variable is a natural log transform of the death rate per 100,000 population for the respective group. We transform each regression coefficient by 100(exp(coef.)-1) and the standard errors using the delta method. Thus each reported coefficient can be interpreted as the percent effect of Medicaid expansion on the specific mortality rate for a given age group (i.e., a coefficient of -2 would mean that the mortality rate would decrease by 2%). Covariates are used in even-numbered regressions. Third difference (regressions (5)-(6)) is ages 55-64 versus aged 65-74. Covariates include % Medicare Advantage Penetration, % Medicare enrollees receiving social security disability benefits, Per Capita Income, Median Household Income, % Persons in Poverty, Unemployment Rate, 16+, Number of individuals with diabetes as percent of the population, Age-adjusted percent of the population that is physically inactive, Percent of population that is obese, Smoking Prevalence (all genders), Non-Federal Active MDs per 1000 people.

Table 2: The effect of Medicaid expansion on the percentage uninsured by age group and demographics

	ACS (County-level)			SAHIE	
	DD 55-64	DD 65-74	DDD	State DD 50-64	County DD 50-64
	(1)	(2)	(3)	(4)	(5)
All	–1.75***	–0.11*	–1.60***	–1.97***	–1.99***
	(0.30)	(0.06)	(0.41)	(0.34)	(0.49)
Female	–1.81***	–0.15**	–1.59***	–1.99***	–2.11***
	(0.39)	(0.06)	(0.48)	(0.41)	(0.50)
Male	–1.68***	–0.08	–1.63***	–1.93***	–1.89***
	(0.25)	(0.07)	(0.36)	(0.29)	(0.48)
White (non-Hispanic)	–1.57***	–0.03	–1.38***	–1.79***	
	(0.32)	(0.04)	(0.41)	(0.33)	
Black (non-Hispanic)	–2.55***	–0.27	–2.64**	–2.66***	
	(0.74)	(0.23)	(0.99)	(0.75)	
Other (non-Hispanic)	–0.99	–1.06	–1.93		
	(0.94)	(0.74)	(1.17)		
Hispanic	–3.22*	0.34	–4.00**	–3.59***	
	(1.74)	(0.54)	(1.92)	(1.00)	
Elementary school	–4.40***	–0.96***	–3.55***		
	(1.05)	(0.35)	(0.98)		
High school incomplete	–4.58***	–0.55***	–4.12***		
	(0.69)	(0.20)	(0.82)		
High school complete	–2.03***	0.03	–2.02***		
	(0.42)	(0.08)	(0.54)		
Some college	–1.32***	–0.05	–1.19***		
	(0.28)	(0.05)	(0.35)		
Below 138% FPL	–7.11***	–0.43*	–6.68***	–7.03***	–7.27***
	(0.83)	(0.23)	(1.10)	(0.81)	(1.33)
138%-400% FPL	–1.83***	–0.10	–1.68***	–2.13***	–1.81**
	(0.44)	(0.07)	(0.55)	(0.47)	(0.73)
Observations	369	369	738	451	19628
Unit fixed-effects (state or county)	Yes	Yes	Yes	Yes	Yes
Year fixed-effects	Yes	Yes	Yes	Yes	Yes
Covariates	Yes	Yes	Yes	Yes	Yes
Weights used	ATTxPop	ATTxPop	ATTxPop	ATTxPop	ATTxPop

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors clustered at the state-level are reported in parentheses. Each estimate comes from a separate regression and the point estimate can be interpreted as the effect of Medicaid expansion on the percent uninsured for a given group (i.e., a coefficient of -2 would mean that those without insurance in that group decreased by 2 percentage points). ACS data refers to the American Community Survey and are at the county-level covering 2008 to 2016. SAHIE data refers to the Census Bureau's Small Area Health Insurance Estimates and are reported at both state and county-levels covering 2006 to 2016. See Table A4 in Appendix to see the same Table, but with population weights rather than ATT$\times$Pop weights.

2.6.3 DD and Triple-Difference Regression Results

We next turn to regression analysis in Table 1, showing results from DD regressions at the state and county-level, following Equation 1. These specifications include year and unit fixed effects (at either the state or county levels) and use ATT$\times$population weights. We present results separately for our principal treatment group (ages 55-64) and the placebo group (ages 65-74). The table also shows triple-difference results, following Equation 3. We show separate results for healthcare-amenable mortality, non-amenable mortality, and total mortality. Even-numbered columns include covariates. Since each dependent variable is a natural log transform of the death rate per 100,000, we transform each regression coefficient by 100(exp(coef.)-1) and the standard errors using the delta method. Thus each reported coefficient can be interpreted as the percent effect on the mortality rate (i.e., A coefficient of -2 would mean that the mortality rate would decrease by 2%).

These regression results are consistent with those presented in the event-study plots. At the county-level we find no clear evidence that Medicaid expansions affected mortality of the young elderly; in the DD model, we find a statistically insignificant 0.65% post-expansion fall in amenable mortality and in the triple-difference design we find an insignificant decline of 0.85%. At the state-level we see evidence of declining healthcare amenable and all cause mortality for the young elderly. While we find evidence of this using both the DD and triple difference research designs, only the triple difference specification yields parallel pre-trends in the event-study results. Table 1 suggests that the young elderly health care amenable mortality rate decreased by 1.53% and the all-cause mortality rate decreased by 1.32% following expansion.

The state and county-level estimates are discrepant due to the difference in propensity score weights across the two specification types. The county-level specifications find non-expansion counties that look the most like expansion counties, while the state-level specifications do likewise for states. The use of these weights achieves balance on covariates across treated and control units in similar ways (see Table A2), but it is evident that the use of state or county weighting scheme affects our point estimates. Geographic aggregation has been shown to affect the magnitude of health effects in other settings as well. Lindo (2015) compares the relationship between economic conditions and health and—much like we do here—finds that county-level specifications result in smaller estimated effects than results from state-level specifications. As in the case of Lindo (2015), it may be that measurement error in treatment is partially responsible for this discrepancy. Since Medicaid expansion is so tightly tied to income, it may be the case that assigning expansion status to a state results in less measurement error than assigning it at the county level. That is, there may be some counties in an expansion state where very few residents of that county qualified for expanded Medicaid. Similarly, there could be counties in non-expansion states, where a large number of people gained coverage through the marketplace.

Geographic aggregation may affect power as well. We explore this issue in the next section by

conducting power simulations at both the state and county levels. However, before moving to the simulated power analysis, we explore how the insurance expansion affected the rate of insurance and possibility of treatment effect heterogeneity by subgroup.

2.6.4 The effect of Medicaid expansion on insurance coverage

In this section we estimate how Medicaid expansion affects reported health insurance status using survey data. Medicaid expansion plausibly affects mortality by first increasing access to health care by reducing the cost of care and then by patients seeking or receiving care at an increased rate due to this reduced cost. As such, we refer to the effect of Medicaid expansion on health insurance status as a "first-stage," as we anticipate that the mechanism through which the policy affects mortality is changes in insurance.[15] Our empirical analyses are analogous to the DD and triple difference specifications outlined by Equations 1 and 3, but with the percent uninsured as our outcome of interest.

Another key difference between the mortality and health insurance specifications is the geographic level of the analysis. Due to data availability, we present some specifications at the state-year level and others at the county-year level. We first use state-year level data from the American Community Survey (ACS), which corresponds directly to our DD and triple-difference specifications where those aged 55 to 64 are considered treated. We then use data from the Census Bureau's Small Area Health Insurance Estimates (SAHIE) as a second data source. The SAHIE data have the advantage of being at the county-level, but the disadvantage of lacking the same age-groups used in our ACS analysis. Thus, specifications using the SAHIE data consider those aged 50 to 64 to be treated and only employ our DD design comparing those of similar age in expansion counties to non-expansion counties.

Both the ACS and SAHIE provide additional information on race, ethnicity, income, and education. When possible—not all additional information is available for each data source and geography combination—we provide a first stage estimate for relevant subgroups as well. We are of course the most interested in the first stage for low-income individuals as Medicaid eligibility is tied to income.

Table 2 demonstrates our results. Using the ACS data and our DD specifications, we can see that there was a decrease in the uninsurance rate of about 1.75% among those aged 55-64. When using our triple difference specification, the point estimate is quite similar showing a 1.6% drop. Estimates using the SAHIE data show similar first stage estimates with uninsurance rate reductions of 2.0% using state or county-level data. We do not read too much into the difference across these two datasets since the confidence intervals overlap. However, ACA-derived insurance gains could

[15] We do not mean to conflate this term with the first-stage terminology used in two-stage-least-squares regressions, simply that this is the a necessary first effect to be present if Medicaid expansions are to affect health.

be somewhat smaller among the near elderly (on whom we focus) than among younger adults, perhaps because the near-elderly have greater healthcare needs and greater income, which may have led many to obtain insurance before the ACA induced Medicaid expansions. In any case, we use the larger of the estimates, 2.0%, throughout our paper when referring to the first stage for the near elderly.

Demographic characteristics are associated with heterogeneous first stage estimates. Certain groups—non-Hispanic Blacks, Hispanics, those without a high school degree, and low income individuals—have the largest first-stage estimates. As expected, due to the mechanical nature between income and Medicaid eligibility, low-income individuals have the largest first stage estimates—with point estimates indicating an increase in insurance coverage of between 6.7 and 7.3%.[16] Unfortunately, our mortality data do not include income of the decedent. We do however know the race, ethnicity, and education status of the decedent, which are correlated with income and result in a larger first stage estimates near 5%. Thus, around 5% is likely as large a first stage as one can likely achieve without linked individual data on some combination of income, family status (children at home), pre-expansion insurance, and mortality.

Of note, this table uses weights that interact inverse propensity scores with population, so the results are comparable with our mortality regressions. An analogous table using population weights is presented as Table A4 in the Appendix. When weighting by population the first stage estimates tend to be slightly smaller in magnitude than those results using inverse propensity score weights. However, results follow a similar pattern and are quite close to one another, with the respective confidence interval across each point estimate overlapping in most cases.

Our first-stage estimates suggesting that insurance coverage increased by 1.6 to 2% across the entire near elderly population and by around 7% for those under 138% FPL are similar to estimates from recent work. Wehby and Lyu (2018) use ACS microdata from 2011 to 2015 to show that Medicaid expansion caused a 1.6% decline in the uninsurance rate for those aged 56 to 64.[17] Courtemanche et al. (2017) use the ACS microdata from 2011 to 2014 to show that Medicaid expansion is associated with a gain in insurance for those aged 50 to 64 of 2.8%. This estimate includes all states and is not limited to full vs. non-expansion states. Borgschulte and Vogler (2020) use the SAHIE data and a slightly different specification than ours to estimate a 2.7% increase in those with insurance aged 50 to 64. Finally, Miller et al. (Forthcoming) demonstrate that for those below 138% FPL or without a high-school degree, Medicaid expansion decreased the uninsurance rate by more than 4%. As displayed in the bottom panel of Figure A4 and in Table 2, we find similarly large estimates for this population. However when we consider the entire population of

[16] This larger increase in insurance coverage for the low income sub-population is particularly salient when displayed in event time as in Appendix Figure A4.

[17] See Table 3 from Wehby and Lyu (2018), which limits the sample of states to full and non-expansion states, the same set of states used in our analyses. When including all states their estimate rises to 3.5%.

those aged 55 to 64—not just the low-income population, our first stage estimates attenuate as is seen in the top panel of Figure A4.

Our first-stage estimates are subject to several important limitations. First, the data on insurance coverage are subject to measurement error, which is likely worse in the county-level data. It is possible that this measurement error introduces attenuation bias. To combat this issue, we present results using state-level data from two different data sources and we use the largest measure of the first stage (2%) when interpreting power estimates later on. Second, survey based estimates of Medicaid enrollment have been shown to severely undercount the percent of true Medicaid enrollees. For example, using the Current Population Survey Davern et al. (2009) find that many Medicaid enrollees fail to report being enrolled with many reporting that they are uninsured. It is possible that newly insured individuals are more susceptible to under-reporting Medicaid enrollment status than those earlier enrolled in Medicaid. If this is the case then our first-stage estimates may be attenuated. Third, many who are eligible for Medicaid do not sign up until care is needed or are allowed to sign up retroactively after receiving healthcare (Marton and Yelowitz, 2015). Since prior work has shown that those with health insurance may have different health outcomes than those without (Doyle, 2005), this conditional/retroactive coverage may cause survey based estimates to undercount the share of those with "effective" Medicaid coverage.

For all of these reasons, it is possible that our first-stage is an underestimate of the true impact of Medicaid expansion on health insurance coverage. Thus when considering the mortality effect and the minimum mortality effect that is well powered, we will consider first stages of both 2% (as estimated) and 5%, which may account for some of these attenuating factors. As outlined in the next section, we will also explore analyses for various subgroups with larger first stages in the hopes that these analyses will correspond to larger—and thus more detectable—effects.

2.6.5 Evidence on Heterogeneous Effects

We conduct additional analyses of the effects of ACA-induced insurance variation on mortality, focusing on vulnerable subgroups or particular causes of death. These subgroups can potentially provide a stronger first stage, a stronger second stage, or both. However, moving to subgroup analysis also reduces the population composing each subgroup in each county or state, which affects the annual mortality rate variance. We consider subgroups based on gender, race/ethnicity, and specific cause of death. We present these results in the Appendix Table A3.

At the county level, our search for evidence of a significant effect of Medicaid expansion on mortality for particular subgroups also proves to be largely unsuccessful; we find statistically significant results for two of thirteen DD coefficients and one of thirteen triple difference coefficients. As with our full-sample results, we find stronger evidence for mortality effects when using state-level data—in particular, for respiratory mortality and mortality for men. We do find strong and

significant effects for Black mortality across both the county and state aggregates for both DD and triple difference designs; estimates suggest that Medicaid expansion induced reductions in Black mortality between 3.7 and 8.6 percent. In the next section we explore how these larger treatment effects and larger first stage estimates affect statistical power.

3 Simulated Power Analysis

We conduct a simulation-based power analysis by artificially introducing treatment effects of different sizes into the data in the pre-treatment period, and then assessing (over many iterations) how often our specifications of interest can detect these effects. The goal of this analysis is to determine the minimum effect of health insurance on amenable mortality that is reliably detectable with our data and research design. The alternative of a closed-form power analysis requires fully modeling the data generating process, including parameterizing the error term for both variance and covariance terms, and is especially hard to construct with panel data in which observations can be correlated over time (Burlig et al., 2020). Our use of regression weights and clustered standard errors further contributes to the difficulty in producing a tractable and credible form for an analytic power calculation.

Simulation avoids these challenges and lets us use the same research design and econometric specification as the main analysis. In our preferred simulations, we perturb real, but untreated data. For example, our simulation approach mechanically builds in correlations and "noise" from trends in the actual data. A simulation using entirely artificial data would have similar conceptual issues as the closed form analyses, requiring modeling the level and form of trends and correlations in the data.[18]

Our preferred simulation proceeds as follows. We exclude all data from the post-treatment period (2014 and after), opting to use only data from the pre-treatment period to ensure that it is not possible for any treatment effect to influence our results.

We maintain the same basic structure in our power analysis as was present in our earlier analysis. We still use ten years of "pre-treatment data" (from 2001 to 2010 instead of 2004 to 2013) and three years of "post-treatment data" (from 2011 to 2013 instead of 2014 to 2016), where a pseudo treatment is assigned in 2011. We restrict the sample of states to be the same ones used in our earlier analysis (i.e., the 41 states that either fully expanded or did not expand Medicaid).

We next repeat the following 1,000 times: we randomly assign a pseudo-expansion status to 22 of the 41 states, the same number of full-expansion states. In each case, we assume that the expansion occurred in 2011, giving us ten years of pre-pseudo-expansion and three years of post-pseudo-expansion data for each pseudo-treated state, ensuring that there are always the same number of

[18] We do explore some simulations that use artificial data. These results are presented in Table 3.

control state-years and treated state-years in our power analysis as in our earlier specification of interest.

For each randomly drawn set of pseudo-treated states, we impose a pseudo treatment effect of a reduction in amenable mortality (from 0% to 8%, in 0.25% increments) for all persons aged 55-64 living in a pseudo-treated state. We do this by randomly removing deaths from each pseudo-treated county-year (or state-year) using draws from a binomial distribution. For example, if a county-year has 100 healthcare-amenable deaths and the imposed treatment effect is 0.5%, we remove each death with probability 0.005. The expected number of remaining deaths is then 99.5, but the actual number must be a whole number and could be 100, 99, 98, etc. Each imposed treatment effect is randomly distributed across the pseudo-treated states and across counties in each state. Thus, as in the above example, it is unlikely that any pseudo-treated county will have its mortality rate decrease by exactly 0.5%, but the pseudo-treated counties will still experience the imposed treatment effect on average (subject to sample variation).[19]

We opt for this approach since it ensures that each simulated dataset used in our power analysis resembles data that could actually occur. Consider the alternative of simply reducing the *death rate* in each county by a fixed percent. In addition to enforcing an identical treatment effect across all treated units, this can produce negative death rates, an undesirable property in our setting since we are examining log death rates and bounding these rates at zero will affect the overall average treatment effect. Thus while a bit more cumbersome, we strongly prefer perturbing the data in a manner that produces plausible data.

Within each iteration, we re-estimate the inverse propensity score based weights so that the pseudo-control states that are the best match for the pseudo-treated states receive the most weight. Once we have introduced the artificial shocks, we run the DD model in Equation 1 and save the regression coefficient and standard error. The percentage of times a result is found to be statistically significant for a given effect size and significance level is the power for that effect size and significance level. For each specification, we report the smallest imposed effect size that can be detected at the 5% significance level with 80% power—an arbitrary, but common threshold used to denote "adequate power." We similarly assess power using the triple difference model in Equation 3.

In addition to statistical power, we also report three measures based upon Gelman and Carlin (2014) that inform the plausibility of any significant results obtained, given the underlying power of the study: 1) the percentage of times a significant, estimated treatment effect has the wrong

[19]Of note, when we impose zero treatment effect, we are not manipulating the data such that there is guaranteed to be zero effect. Similarly when we impose a specific effect (x%), there is no guarantee that exactly this effect will be found if a regression is run. This is because we are adding the imposed effect to the existing data. Since the data are untreated, there should be no effect on average. If we were to use data from the treated time period, we would first need to "enforce the null" before imposing our effect size. We opt not to follow this approach because it is unclear how best to enforce the null before adding in an imposed treatment effect.

sign—i.e., the opposite sign as the imposed treatment effect; 2) for the subset of cases where a significant effect is found, the mean ratio of the estimated treatment effect to the true, imposed effect (i.e., the exaggeration ratio); and 3) the percentage of significant treatment effect estimates that have the correct sign and an exaggeration ratio below two—which we term a "believable" coefficient.

Our empirical specifications of interest (Equations 1 and 3) and accompanying power simulations require a number of choices: for example, which, if any, weights to be used; the level of clustering for standard errors; the inclusion of control variables; the length of the pre-treatment period; the level of the analysis (e.g., a county-year vs. state-year analysis); whether or not to include data from the post-treatment period in the power simulation; and the use of perturbed—but real data—vs. entirely artificial data in the power simulation. Many of these decisions are made on an ad hoc basis and it is important to understand how these choices affect power and our estimate of the minimum detectable effect size. As such, we repeat our power simulation after altering each of these choices, reporting the minimum effect size that is detectable at the 5% significance level with 80% power for each.

Finally, it is possible that our analysis will lack power for the whole population of those aged 55-64 because not every member in this group experienced an insurance expansion. Indeed Medicaid expansion targeted lower-income individuals and an analysis solely focused on subgroups more likely to be treated could have more power than a whole group analysis. While the mortality data we use do not report income, the data do report other demographic information that is correlated with a larger first stage. The data also report cause of death and it is entirely possible that individuals with certain ailments (e.g., HIV or cancer) would disproportionately benefit from Medicaid enrollment, which would improve power. However, focusing on subgroups is not guaranteed to improve power, as zooming in on any subset of the data may increase variance more than the gains from an increased first stage or enlarged treatment effect. We explore this possibility by examining power, variance, and first-stage estimates in Section 3.3.

3.1 Full Sample Power Simulation Results

Figure 2 illustrates the results from our power simulations, using the log amenable mortality rate per 100,000 as the dependent variable. Each simulation uses data from 2001-2013 from our actual treated and control states. A pseudo-shock of gradually increasing magnitude is applied to 22 states chosen at random for 2011 to 2013. For a pseudo-shock of a given size, power is defined as the percent of the 1,000 simulations that result in a statistically significant pseudo-treatment effect at a given significance level. We display results for the 10%, 5%, 1% and .1% significance levels. The left panel shows simulation results using state-level data, while the right shows simulation

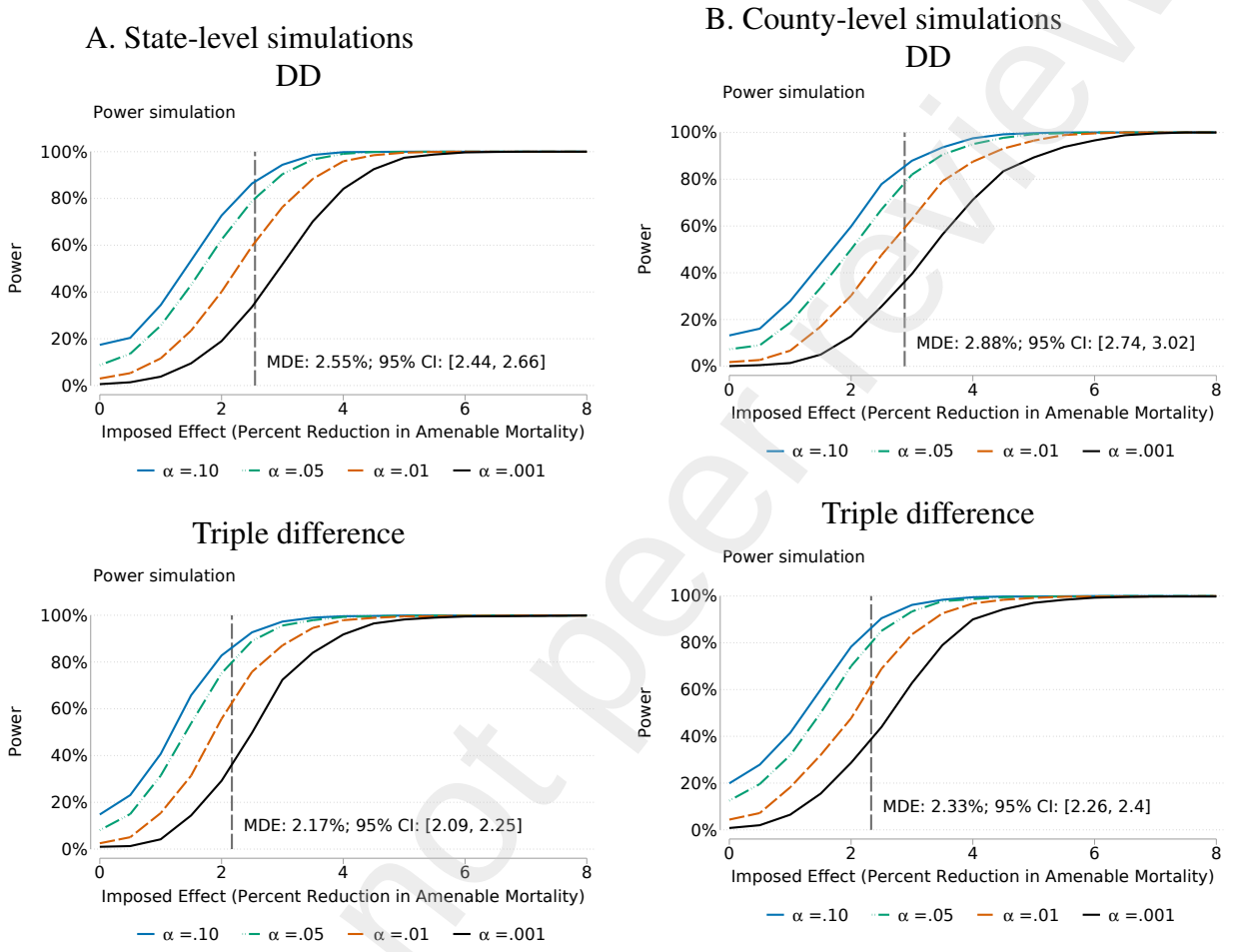
results using county-level data. The top row shows DD results that correspond to Equation 1 and the bottom row shows triple-difference results that correspond to Equation 3. When we impose an effect of 1.53% using the triple difference specification and state-level data, 51.2% of the 1000 simulations find a statistically significant effect at the 5% level. Simulation results indicate that the smallest mortality reduction that could be detected at the state (county) level —given our research design—80% of the time at the 5% significance level is 2.55% (2.88%) for the DD specification and 2.17% (2.33%) for the triple-difference specification.

For our triple difference specification, we depict measures of believability adapted from Gelman and Carlin (2014) in Figure 3. Panel A shows the ratio of the magnitude of the estimated effect—when found to be statistically significant—relative to the true magnitude, imposed in the simulation. For population effect sizes under 1%, the exaggeration ratio is high. That is, an effect large enough to be statistically significant is likely to be far from the truth, on average two to five times larger. In Panel B we show the proportion of statistically significant results that have the wrong sign. This proportion is also appreciable for the smaller imposed effect sizes, reaching almost 10%. As we increase the imposed effect size, prevalence of sign-error shrinks, and it becomes negligible for effect sizes above 1%; the exaggeration ratio also shrinks, but more slowly. We combine these two measures in Panel C, which considers a "believably" significant specification to be one that does not have a sign-error and that has a magnitude error no greater than two. The believability curves follow the power curves from Figure 2 quite closely. By the time the minimum detectable effect size is reached, issues of believability subside with more than 80% of specifications returning a significant and believable estimate.

As outlined earlier, our research design (represented by Equations 1 and 3), the data included in the analysis, and ad hoc choices made to implement our power simulation may affect power and thus our minimum detectable effect estimate. Table 3 demonstrates the sensitivity of this estimate to those choices. For each MDE estimate we report 95% confidence intervals that are computed using 1,000 bootstrapped draws from the 1,000 power simulations. Column (1) represents our preferred estimates: containing ATT$\times$population weights, clustering standard errors at the state level, and controlling for other county-year variables. Our preferred simulation uses data from treated and untreated states, but only from the pre-treatment period; with 2001-2010 being the pre-treatment period and the pseudo-treatment affecting 2011 through 2013.

The top panel presents simulation results using state-level data and the bottom panel presents analogous results at the county-level. Columns (1) through (6) all are power simulations that only use pre-Medicaid expansion data. In each column, we vary at least one feature from our preferred simulation, including: the level of clustering (county or state), the weights used (ATT$\times$population or population), the length of the pre-period (2001-2010, 2004-2010, or 2007-2010), and whether or not control variables are included. Columns (7) and (8) use some data that come from post-

Figure 2: Power simulation results



Note: Each sub-figure represents results from 1000 different power simulation as outlined in the text. Broadly, each simulation uses data from 2001-2013 from our actual treated and control states; a pseudo-shock of gradually increasing magnitude is applied to 22 states chosen at random for 2011 to 2013; and for a pseudo-shock of a given size, power is defined as the percent of the 1,000 simulations that result in a statistically significant pseudo-treatment effect at a given significance level. Results for the 10% significance level are displayed by the solid blue line; results for the 5% significance level are depicted by the dashed green line; 1% significance level estimates are represented by the dashed orange line; and the .1% significance level results are shown by the solid black line. MDE stands for minimum detectable effect size and is the smallest imposed treatment effect that is detectable with 80% power at the 5% significance level. The 95% confidence interval for the MDE is estimated through a bootstrapping procedure. The left column presents results using state-level data, while the right column presents results using county-level data. The DD results correspond to Equation 1, while the triple difference results correspond to Equation 3.

Medicaid expansion. Importantly these data never included mortality rates from those states that expanded Medicaid in the years following expansion. Column (7) excludes data from expansion states and oversamples control (i.e., non-expansion) states so that 41 total states are selected. Column (8) uses exactly the same data as is used in Equation 1 and 3, except for the dependent variable. The dependent variable is entirely artificial and is constructed such that each county-year death rate is a draw from a normal distribution that has the same mean and variance as that given county's actual death rate from 2001 to 2013. We also allow the artificial data to exhibit the same time trends as the observed mortality data (see Figure A3). We do this by regressing a simple time trend on the raw mortality data in the pre-pseudo-treatment period and assuming that this time trend continues into the post-pseudo-treatment period.

The vast majority of these simulations yield similar minimum detectable estimates, indicating that the idiosyncratic choices in our preferred simulation do not drive our estimate. In general, most MDEs are just above 2%; the triple difference specifications have smaller minimum detectable effect sizes; and controlling for covariates improves power. The smallest MDEs use population weights instead of ATT$\times$population weights. These specifications are somewhat problematic as they exhibit pre-trends when analyzed in event-time.[20]

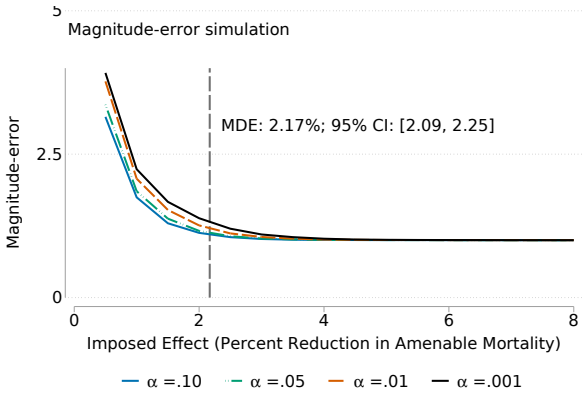
There could be a trade-off between identification (or a preferred specification) and the power of an estimate. For our analysis, we aim to first determine which specifications are reasonable and most likely to be well identified before we turn to concerns of power. An unidentified, but well-powered estimate may well be uninterpretable, while an identified but under-powered estimate is much easier to interpret. For example, consider the minimum detectable effect size estimates in Table 3. Among the specifications using only pre-expansion data, the set with the most power use population weights; however when we use population weights our identification assumptions are at a higher risk of being invalid since there are problematic pre-trends for these analyses. Therefore we opt to use the slightly lower powered, but better identified estimates that use inverse propensity score weights.

The largest minimum detectable effect sizes come from the specifications that use artificial death data. This is likely because these artificial data are not able to take advantage of the power gains from including covariates, county fixed effects, and year fixed effects. One could imagine simulating artificial data that is micromanaged and tweaked to maximize power, but it would be unclear what the optimal level would be. That is, the end goal is to produce a power estimate for the real analysis to be done, not to design a dataset that produces maximum power when the analysis is applied to it.

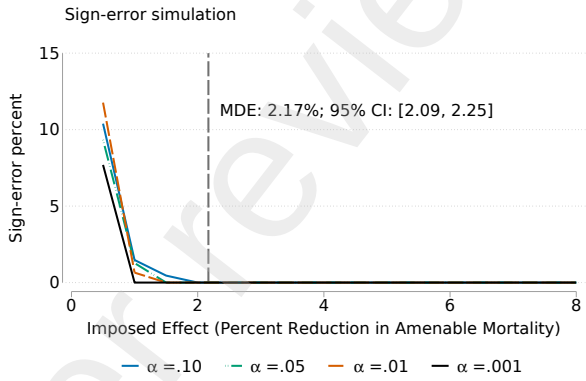
[20]These specifications were discussed extensively in an earlier version of this manuscript and are available by request. We omit them for brevity and have shifted towards the use of inverse propensity score weights to better ensure that the identifying assumption of parallel trends holds in our analyses.

Figure 3: As power increases so does believability (i.e., reduced magnitude error and chance of sign-error)

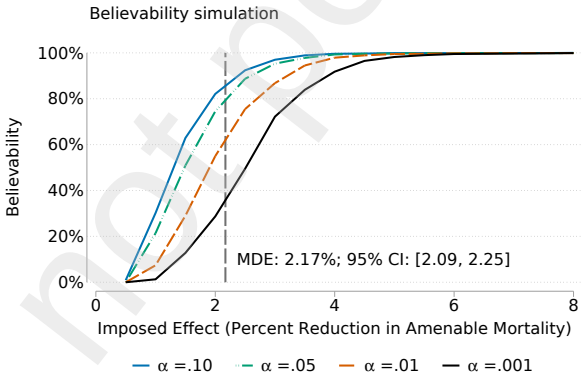
A. Magnitude error



B. Sign error



C. Believability % (no sign-error and magnitude error ≤ 2)



Note: Results here correspond to power simulations at state-level for the triple difference specification (the bottom right sub-figure from Figure 2). Each sub-figure here represents a different measurement in addition to power for that simulation. Magnitude-error is depicted in sub-figure A, which is the average of statistically significant estimate divided by the true effect size. Sign-error is depicted in sub-figure B, which is the % of times a statistically significant estimate has a different sign that the imposed treatment effect. Believability combines these two concepts and shows the % of simulations that report a statistically significant estimate without a sign-error and without a magnitude error. Results for the 10% significance level are displayed by the solid blue line; results for the 5% significance level are depicted by the dashed green line; 1% significance level estimates are represented by the dashed orange line; and the .1% significance level results are shown by the solid black line. MDE stands for minimum detectable effect size and is the smallest imposed treatment effect that is detectable with 80% power at the 5% significance level. The 95% confidence interval for the MDE is estimated through a bootstrapping procedure.

Table 3: The minimum imposed reduction (%) in health care amenable mortality rate that can be detected with 80% power at the 5% significance level varies with specification and other choices made in the power simulation.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Simulations using only pre-expansion data						Simulations including post-expansion data	
							Non-expansion states only	Using artificial death data
State-level								
MDE, DD 55-64	2.55 (2.44 to 2.66)	2.27 (2.21 to 2.33)	1.92 (1.81 to 2.03)	3.43 (3.30 to 3.56)	2.36 (2.30 to 2.43)		1.17 (1.11 to 1.24)	2.15 (2.00 to 2.29)
MDE, DDD	2.17 (2.09 to 2.25)	2.04 (1.94 to 2.13)	1.79 (1.73 to 1.85)	2.18 (2.11 to 2.26)	1.77 (1.73 to 1.81)		1.26 (1.21 to 1.32)	5.28 (5.11 to 5.45)
County-level								
MDE, DD 55-64	2.88 (2.74 to 3.02)	2.79 (2.70 to 2.87)	2.36 (2.23 to 2.50)	4.23 (4.08 to 4.38)	2.28 (2.21 to 2.34)	2.39 (2.24 to 2.53)	1.82 (1.72 to 1.92)	5.29 (4.70 to 5.88)
MDE, DDD	2.33 (2.26 to 2.40)	2.21 (2.13 to 2.29)	2.07 (1.99 to 2.16)	2.36 (2.29 to 2.44)	1.75 (1.70 to 1.80)	2.23 (2.15 to 2.31)	1.37 (1.25 to 1.50)	9.77 (9.01 to 10.53)
Weights	ATTxPop	ATTxPop	ATTxPop	ATTxPop	Pop	ATTxPop	ATTxPop	ATTxPop
Clustering	State	State	State	State	State	County	State	State
Controls	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Pre-treatment years	2001-2010	2004-2010	2007-2010	2001-2010	2001-2010	2001-2010	2001-2013	2001-2013
Post-treatment years	2011-2013	2011-2013	2011-2013	2011-2013	2011-2013	2011-2013	2014-2016	2014-2016

Note: Each number comes from a separate power simulation. The top set of results come from simulations that use state-level data. The bottom set of results come from simulations that use county-level data. The DD specifications are analogous to Equation 1, except with noted differences in each column. The DDD specifications are analogous to Equation 3, except with noted differences in each column. 95% confidence intervals in parentheses are calculated using a bootstrapping procedure. MDE refers to minimum detectable effect size at the population level.

3.2 Interpretation of results

Comparing the minimum detectable effect size to the size of the estimated first stage (i.e., the % increase in insured status due to Medicaid expansion) is a natural way to examine the plausibility of any MDE estimate. If done directly by simply comparing the 2.17% MDE to the 2% first stage, it appears that a well-powered result would also be unrealistic. This would imply that Medicaid expansion would have to reduce the average amenable mortality rate of all newly insured persons by 109% ($= \frac{0.0217}{0.0200}$). That is, the death rate would need to decrease by more than the % insured increased for the result to be well powered. However, this comparison bakes in the assumption that the underlying mortality rate of the newly Medicaid-insured is identical to other persons aged 55-64, and this need not be the case.

The mortality rate for the newly insured could be much higher (e.g., the newly insured tend to be low income, and thus may have higher mortality), or it may even be lower (e.g., the disabled are already insured and those in poor health could be more likely to already have insurance). Which is the case is unclear and cannot be evaluated using our data. Indeed, prior research has presents mixed evidence. Kowalski (2020) provides evidence that complier populations can be of better health relative to comparison populations. Finkelstein et al. (2012) study a likely lower-income, less-healthy population (persons who applied for the Oregon Medicaid expansion lottery), and report annual total mortality for the controls of 0.008, which is similar to the average total mortality rate we find for persons aged 55-64 in both full-expansion and non-expansion States.

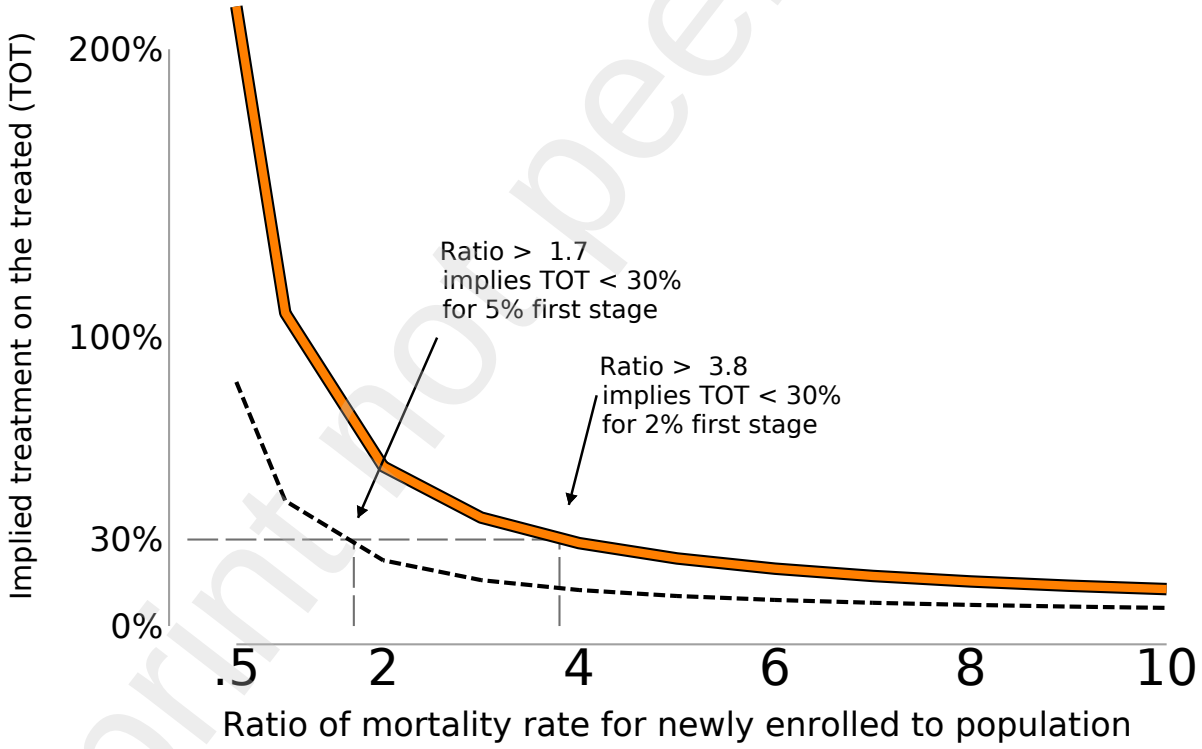
Conversely, Chetty et al. (2016) demonstrate a strong link between income and life-expectancy and income is directly tied to Medicaid eligibility. In a similar vein, Miller et al. (Forthcoming) use data from the National Health Interview to establish that Medicaid enrollees aged 55-64 in 2014 have a 2.3% probability of death in the next year, higher than the cohort average of 1.4%. Finally, Marton and Yelowitz (2015) show that conditional coverage does occur. That is, people may wait to sign up for Medicaid until they are sick. Thus, even if recent Medicaid enrollees appear to be of similar health to the general population, it could be the case that those eligible, but *not yet enrolled* are more likely to suffer a catastrophic health event or death without insurance.

Having a clear understanding of how much more likely the recently insured are to die relative to the general population of 55-64 year olds is crucial to understanding whether or not our minimum detectable effect size is of a plausible or an unrealistic magnitude. For example, if the mortality rate of the 55 to 64 year olds who gained insurance due to Medicaid expansions was twice that of the general population of 55 to 64 year olds, then the estimated treatment effect on the treated would drop from 109% to 58%. Since we cannot determine the mortality ratio between these two groups with certainty, we present a range of potential treatment on the treated estimates that correspond with different mortality ratios in Figure 4. Under the assumption that the first stage is 2%, to obtain a treatment on the treated of 30% or less the newly insured would have to be at least

3.8 times more likely to die than the general population without insurance. If the first stage was around 5%, then this ratio would only need to be 1.7 or greater to obtain a treatment on the treated of 30% or less.

Thus for our minimum detectable effect size to be of plausible magnitude at least one of two things must be true. First, it could be that those who recently obtained insurance are significantly more likely to die than those in the general population. Second, it could be that the true first stage is much larger than the estimated first stage of 2%. This is of course entirely plausible given the concerns of underreporting and conditional coverage previously documented (Davern et al., 2009; Marton and Yelowitz, 2015), but it is nonetheless an important consideration when determining the plausibility of our estimates.

Figure 4: For a given average treatment effect, the implied treatment on the treated (TOT) decreases as the mortality ratio of the newly insured relative to the remaining population changes grows or a larger share of the population receives insurance.



Note: The orange line represents the relationship between the implied treatment on the treated and the mortality ratio of the newly insured relative to the remaining population for a 2% first stage. The dashed black line represents the same relationship, but for a 5% first stage. Both consider an average treatment effect size of 2.17%, which is the minimum detectable effect size for the state-level triple difference research design.

3.3 Power for vulnerable subgroups

We also conduct power analyses for subgroups based on different demographics, education levels, and causes of death. We estimate the minimum detectable reduction in mortality across gender, race, cause of death, and education status that could be detected at the 5% significance-level 80% of the time. Table 2 demonstrates that many of these subgroups have larger first stage estimates than the overall population (e.g., Blacks aged 55-64 saw increases in insurance of around 2.6%, Hispanics saw increased coverage of 3.2 to 4%, and those who had not completed high-school saw increases of around 4.5%). We also estimate the MDE for specific causes of death (cancer, HIV, diabetes, cardiac, and respiratory ailments) that may be more directly preventable by insurance induced access to healthcare (e.g. HIV through antiretrovirals or cancer through chemo-therapies). It is possible that through increased size of the first-stage or through having a larger treatment effect (i.e., mortality reduction) there will be improved power.

Figure 5 displays results from these analyses. We construct each minimum detectable effect size estimate following an analogous power simulation as outlined earlier at the state-level in Panel A and at the county level in Panel B. Within each panel scatterplots from the DD specifications are on the left-side and scatterplots from triple-difference specifications are on the right. Of note, for both the DD and triple difference specifications the average and median MDE across our subgroup analyses are larger than the respective MDEs that use the whole sample.[21]

In the top row of each panel, the minimum detectable effect size is plotted against a measure of variance for each subgroup. As a simple measure of variance, we use the coefficient of variation, which is the average standard deviation of the death rate compared to the mean death rate by subgroup. There is a clear, positive relationship. That is, those sub-groups with the largest minimum detectable effect sizes (i.e., those whose analyses have the lowest power) have the most volatile mortality rates. This increased variance makes intuitive sense; while each subgroup analysis has the same number of observations—with the exception of a small number of counties that have no Black or no Hispanic residents—the number of people contributing to the death rate in each county-year is quite variant. That is, subgroups with smaller populations will naturally have more variation than the larger subpopulations. This is why average power among the subgroups is worse than when all groups are combined; breaking into subgroups increases noise in each group.

The second row illustrates that groups with larger first-stages tend to have larger minimum detectable effect sizes. This finding may initially be surprising, but it is driven by the fact that the power improvements expected by "zooming-in" on sub-groups with larger first-stages and larger

[21] For the state-level (county) DD specifications the average MDE across subgroups is 5.45 (7.46) and the median is 3.33 (4.73); compared to an MDE of 2.55 (2.88) using the whole population of 55-64 year olds. For the triple-difference specification the average MDE across subgroups is 6.06 (6.75) and the median is 3.82 (4.35); compared to an MDE of 2.17 (2.33) using the whole population of 55-64 year olds.

treatment effects are offset by the increased variance of average mortality associated with these more narrowly defined groups. Finally the third row demonstrates that larger measures of volatility are associated with larger first-stage estimates. Thus, in our setting it does not appear that we can improve power by focusing on smaller subgroups since the increase in volatility appears to dominate the increased first-stage and potentially larger treatment effects.

3.4 What data would be needed for reasonable power?

We turn in this section to a different question—what combination of a stronger first stage and a reduction in amenable mortality for the newly insured would be detectable with reasonable power, if we could use a richer dataset, with data on mortality linked to data on income and family status (to determine eligibility for expanded Medicaid coverage) and pre-ACA insurance status (to exclude the always-insured from the sample)?

These hypothetical data are similar to those used in recent work by Miller et al. (Forthcoming) and Goldin et al. (2021). At a minimum one could improve the first stage to 7% by studying only adults with incomes $\leq$ 138% of FPL and one could reach almost 10% as in Simon et al. (2017) by focusing on childless adults with incomes $\leq$ 100% of FPLs. In these analyses, we consider the triple-difference design, since it has been shown to induce more statistical power in our setting.

In this scenario, we imagine that we can identify in each county both a treated subsample and a similar *within-county* control subsample, both aged 55-64. For example, if the treated subsample is childless adults with income $\leq$ 138% of FPL, the within county control subsample could be childless adults with incomes from 138% to 250% of FPL. We assume hypothetical first stages varying from 1% to 20% and hypothetical second stages varying from 0% to 10%. For, say, a 5% first stage and a 10% second stage, we assign "insurance due to Medicaid expansion" to 5% of the persons in a "5% first stage" subsample of each expansion county, and then remove 10% of the amenable mortality deaths from the treated persons in this subsample (thus applying an overall mortality reduction to the subsample of .005). We again use data from 2001 to 2013 and a pseudo-treatment beginning in 2011, and assess whether we could detect this mortality effect if we did not know which specific individuals within this subsample would have gained insurance due to this pseudo-treatment. Since the treated and control samples are drawn at random from the same county and age range, they have the same expected mortality rates, by construction.

In Figure 6, we show power curves only for the 95% significance level. We vary (i) the assumed first stage (we show curves for 1%, 3%, 5%, 10%, 15%, and 20% first stages) and (ii) the imposed mortality reduction for the newly insured (from 0% to 10%) for the 5% significance level. With this hypothetical richer data, we need a smaller number of avoided deaths to be able to reliably detect a treatment effect. For example, with a 10% first stage, we could reliably detect mortality

reductions of 2.4% or more within this subsample. This shows that individual level data with known treatment status can improve power, and highlights the contribution of recent studies that use such data (Miller et al., Forthcoming; Goldin et al., 2021).

4 Conclusion

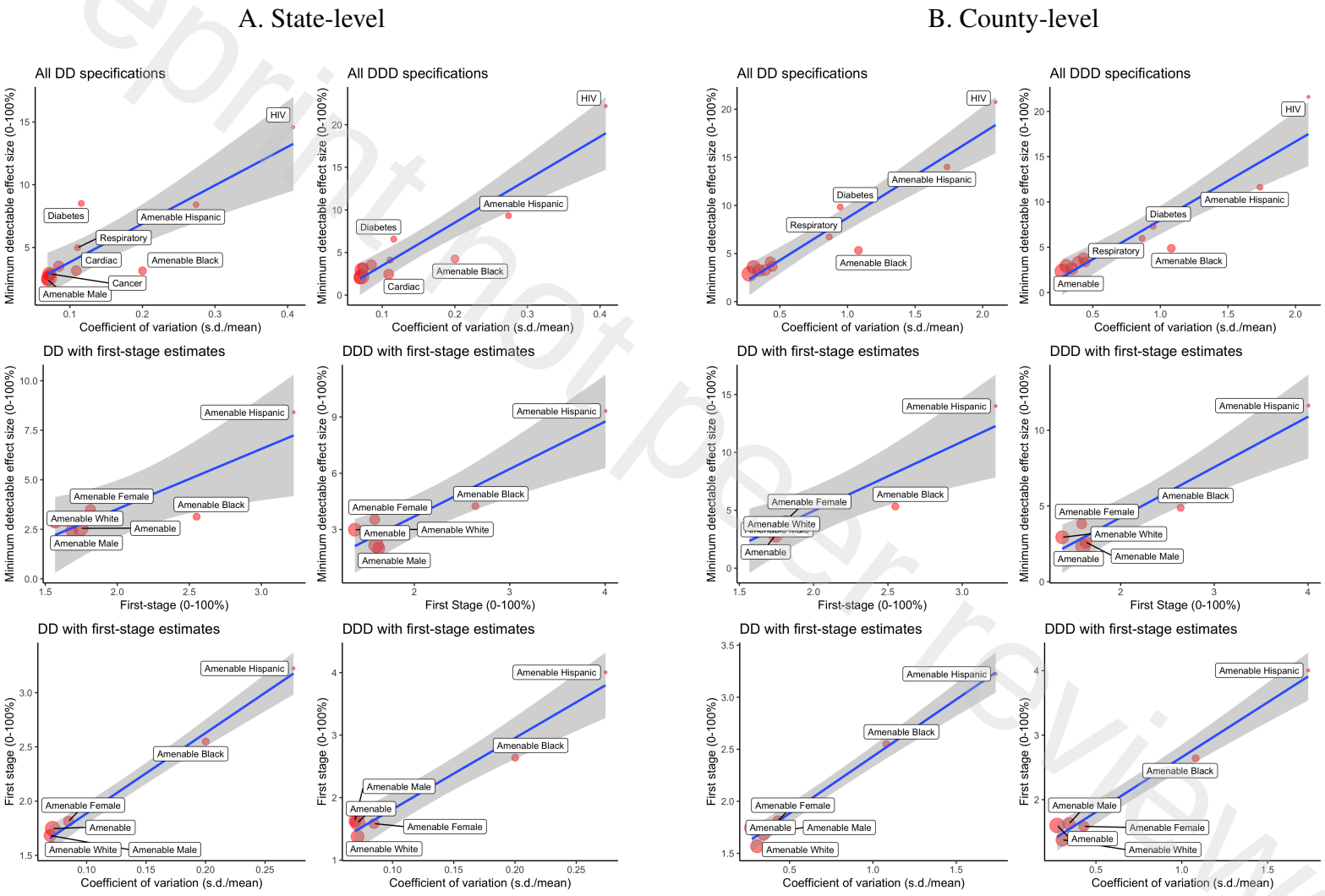
In this manuscript, we explore issues of statistical power using an applied example; the relationship between mortality and health insurance. We focus on whether or not true effects can be reliably detected using geographically aggregated data. Using county-level data we uncover no statistically significant evidence of an ACA-induced decline in mortality in Medicaid expansion states, but with state level data we find evidence that mortality declined for the young elderly by around 1.53%. Using simulated power analyses, we show that our research designs would need to find at least a 2.17% effect to be well powered; when the true effect is 1.53%, statistical power is just above 50%. We end with a discussion of how individual level data, like those used in recent work linking health insurance to mortality, can be used to overcome issues of power.

Power is an important metric that can be used to evaluate the strength of an empirical result. Despite this it is under reported in empirical work. Studies with low statistical power are more likely to find null results even when a true effect exists, and conditional on finding a statistically significant effect they are likely to overstate the magnitude of the true treatment effect. Moreover, meta-analyses of economics have shown that power may be an important issue for economists, with median power being below 18% (Ioannidis et al., 2017).[22]

One impediment to considering power is the large amount of observational work in economics. Another is the lack of closed form power formulas for many common research designs. Here we present a simple method of estimating power using simulation. While we are not the first to estimate power using simulation, we take very seriously the actual impediments a modern researcher faces. There is no uniform approach that will work to estimate power in every setting. Each project will face its own idiosyncratic choices and trade-offs. In this regard, we attempt to outline common issues that other researchers may face when seeking to conduct their own simulated power analyses.

[22]We caution against the interpretation that low-powered studies contain no important information. We suggest instead that results from such studies should be interpreted with more caution than results from otherwise identical, but well-powered studies.

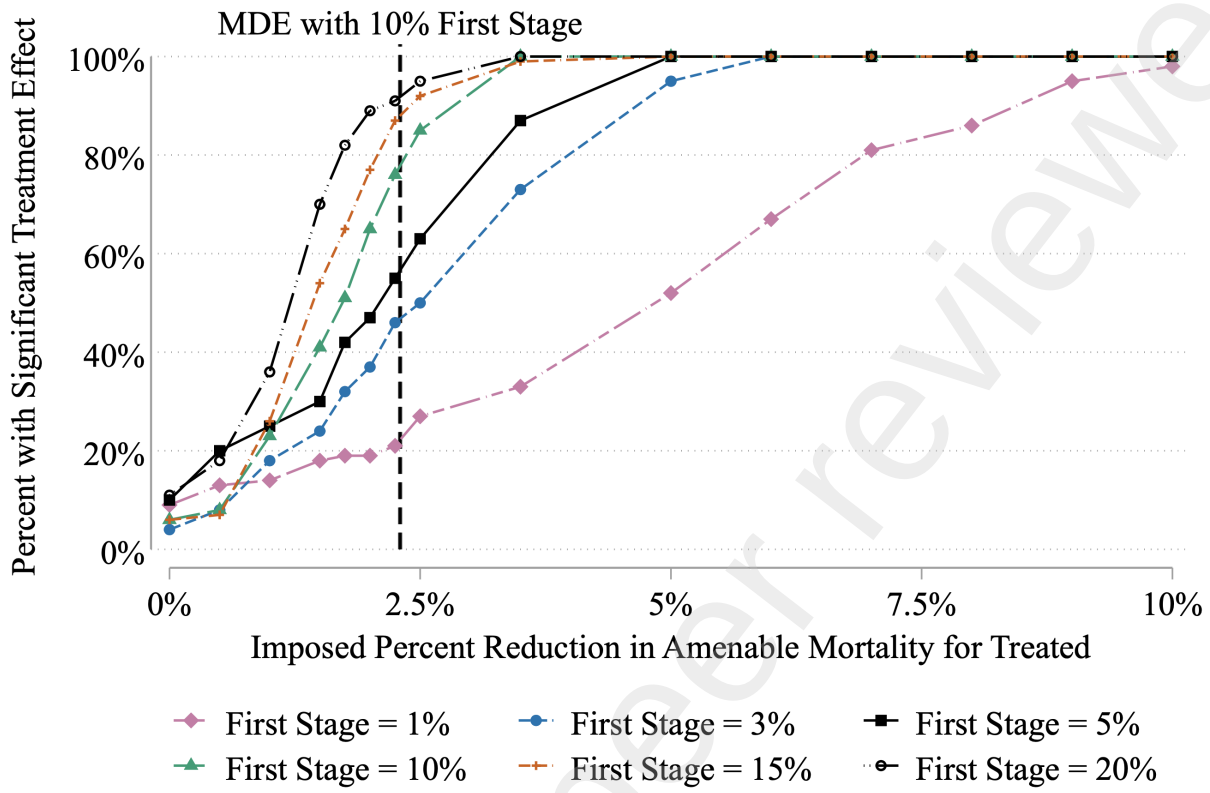
Figure 5: In our setting, focusing on sub-groups with larger first-stages reduces power because the sub-groups tend to have more volatile mortality rates.



Note: The left panel presents simulation results using state-level data. The right panel presents simulation results using county-level data. Within each panel the left column shows results from the DD simulations and the right shows results from triple difference simulations. The top row shows a scatter of the minimum detectable effect size for each subgroup against a measure of variance (the coefficient of variation). The middle row shows a scatter of the minimum detectable effect size against the first stage estimate. The bottom row shows a scatter of the first stage against the coefficient of variation.

42

Figure 6: Additional data on treatment status can greatly reduce minimum detectable effect size



References

Arnold, Benjamin F., Daniel R. Hogan, John M. Colford, and Alan E. Hubbard (2011) "Simulation methods to estimate design power: An overview for applied research," *BMC Medical Research Methodology*, Vol. 11.

Baicker, Katherine, Sarah L. Taubman, Heidi L. Allen, Mira Bernstein, Jonathan H. Gruber, Joseph P. Newhouse, Eric C. Schneider, Bill J. Wright, Alan M. Zaslavsky, and Amy N. Finkelstein (2013) "The Oregon Experiment — Effects of Medicaid on Clinical Outcomes," *New England Journal of Medicine*, Vol. 368, No. 18, pp. 1713–1722.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman (2015) "Six Randomized Evaluations of Microcredit: Introduction and Further Steps," *American Economic Journal: Applied Economics*, Vol. 7, No. 1, pp. 1–21.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) "How Much Should We Trust Differences-In-Differences Estimates?," *The Quarterly Journal of Economics*, Vol. 119, No. 1, pp. 249–275.

Borgschulte, Mark and Jacob Vogler (2020) "Did the ACA Medicaid Expansion Save Lives?" *Journal of Health Economics*, Vol. 72, p. 102333.

Brook, Robert H., John E. Ware, William H. Rogers, Emmett B. Keeler, Allyson R. Davies, Cathy A. Donald, George A. Goldberg, Kathleen N. Lohr, Patricia C. Masthay, and Joseph P. Newhouse (1983) "Does Free Care Improve Adults' Health?: Results from a Randomized Controlled Trial," *New England Journal of Medicine*, Vol. 309, No. 23, pp. 1426–1434.

Burlig, Fiona, Louis Preonas, and Matt Woerman (2020) "Panel data and experimental design," *Journal of Development Economics*, Vol. 144, p. 102458.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò (2013) "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, Vol. 14, No. 5, pp. 365–376.

Card, David, Carlos Dobkin, and Nicole Maestas (2004) "The Impact of Nearly Universal Insurance Coverage on Health Care Utilization and Health: Evidence from Medicare,"Technical Report w10365, National Bureau of Economic Research, Cambridge, MA.

——— (2009) "Does Medicare Save Lives?" *Quarterly Journal of Economics*, Vol. 124, No. 2, pp. 597–636.

Case, Anne and Angus Deaton (2015) "Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century," *Proceedings of the National Academy of Sciences*, Vol. 112, No. 49, pp. 15078–15083.

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler (2016) "The Association Between Income and Life Expectancy in the United States, 2001-2014," *JAMA*, Vol. 315, No. 16, p. 1750.

Courtemanche, Charles, James Marton, Benjamin Ukert, Aaron Yelowitz, and Daniela Zapata (2017) "Early Impacts of the Affordable Care Act on Health Insurance Coverage in Medicaid Expansion and Non-Expansion States," *Journal of Policy Analysis and Management*, Vol. 36, No. 1, pp. 178–210.

Croke, Kevin, Joan Hamory Hicks, Eric Hsu, Michael Kremer, and Edward Miguel (2016) "Does Mass Deworming Affect Child Nutrition? Meta-analysis, Cost-Effectiveness, and Statistical Power,"Technical Report 22382, National Bureau of Economic Research, Cambridge, MA.

Damiano, Peter C., Suzanne E. Bentler, , Erin Robinson, Ki H. Park, and Elizabeth T. Momany (2013) "Evaluation of the IowaCare Program: Information about the Medical Home Expansion."

Davern, Michael, Jacob Alex Klerman, David K. Baugh, Kathleen Thiede Call, and George D. Greenberg (2009) "An Examination of the Medicaid Undercount in the Current Population Survey: Preliminary Results from Record Linking," *Health Services Research*, Vol. 44, No. 3, pp. 965–987.

De Long, J. Bradford and Kevin Lang (1992) "Are all Economic Hypotheses False?," *Journal of Political Economy*, Vol. 100, No. 6, pp. 1257–1272.

Doyle, Joseph J. (2005) "Health Insurance, Treatment and Outcomes: Using Auto Accidents as Health Shocks," *Review of Economics and Statistics*, Vol. 87, No. 2, pp. 256–270.

Dunn, Abe and Adam Hale Shapiro (2019) "Does Medicare Part D Save Lives?" *American Journal of Health Economics*, Vol. 5, No. 1, pp. 126–164.

Finkelstein, A. (2007) "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare," *The Quarterly Journal of Economics*, Vol. 122, No. 1, pp. 1–37.

Finkelstein, Amy and Robin McKnight (2008) "What Did Medicare Do? The Initial Impact of Medicare on Mortality and out of Pocket Medical Spending," *Journal of Public Economics*, Vol. 92, No. 7, pp. 1644–1668.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group (2012) "The Oregon Health Insurance Experiment: Evidence from the First Year," *The Quarterly Journal of Economics*, Vol. 127, No. 3, pp. 1057–1106.

Gallet, Craig A. and Hristos Doucouliagos (2017) "The impact of healthcare spending on health outcomes: A meta-regression analysis," *Social Science & Medicine*, Vol. 179, pp. 9–17.

Gates, Alexandra and Robin Rudowitz (2014) "Wisconsin's BadgerCare Program and the ACA,"Technical report, Kaiser Family Foundation.

Gelman, Andrew and John Carlin (2014) "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science*, Vol. 9, No. 6, pp. 641–651.

Gelman, Andrew and Jennifer Hill (2006) "Sample size and power calculations," in *Data Analysis Using Regression and Multilevel/Hierarchical Models*, pp. 437–456.

Gerin, William, Christine Kapelewski, and Niki L. Page (2017) *Writing the NIH Grant Proposal: A Step-by-Step Guide*, Los Angeles, CA: SAGE Publications, Inc, 3rd edition, pp.232.

Gilbert, Gregory E. and Susan Prion (2016) "Making Sense of Methods and Measurement: The Danger of the Retrospective Power Analysis," *Clinical Simulation in Nursing*, Vol. 12, No. 8, pp. 303–304.

Goldin, Jacob, Ithai Z Lurie, and Janet McCubbin (2021) "Health Insurance and Mortality: Experimental Evidence from Taxpayer Outreach," *Quarterly Journal of Economics*, p. 53.

Gouveia, Nelson and Tony Fletcher (2000) "Time Series Analysis of Air Pollution and Mortality: Effects by Cause, Age and Socioeconomic Status," *Journal of Epidemiology & Community Health*, Vol. 54, No. 10, pp. 750–755.

Hannon, Susan J., Kathy Martin, Len Thomas, and Jim Schieck (1993) "Investigator Disturbance and Clutch Predation in Willow Ptarmigan : Methods for Evaluating Impact," *Journal of Field Ornithology*, Vol. 64, No. 4, pp. 575–586.

Heberlein, Martha, Tricia Brooks, Jocelyn Guyer, Samantha Artiga, and Jessica Stephens (2011) "Holding Steady, Looking Ahead: Annual Findings of a 50-State Survey of Eligibility Rules, Enrollment and Renewal Procedures, and Cost-Sharing Practices in Medicaid and CHIP, 2010–2011,"Technical report, Kaiser Commission on Medicaid and the Uninsured.

Hoenig, John M. and Dennis M. Heisey (2001) "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," *The American Statistician*, Vol. 55, No. 1, pp. 19–24.

Hsiang, Solomon M, Marshall Burke, Edward Miguel, Kyle Meng, and Mark Cane (2009) "Analysis of statistical power reconciles drought-conflict results in Africa."

Ioannidis, John P A (2005) "Why Most Published Research Findings Are False," *PLoS Medicine*, Vol. 2, No. 8, p. e124.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos (2017) "The Power of Bias in Economics Research," *The Economic Journal*, Vol. 127, No. 605, pp. F236–F265.

Johnson, Garrett A., Randall A. Lewis, and David H. Reiley (2017) "When Less Is More: Data and Power in Advertising Experiments," *Marketing Science*, Vol. 36, No. 1, pp. 43–53.

Kaiser Family Foundation (2018) "Status of State Action on the Medicaid Expansion Decision."
Keeler, E. B., Brook, Robert H., Goldberg, George A., Kamberg, Caren J., and Newhouse, Joseph P. (1985) "How Free Care Reduced Hypertension in the Health Insurance Experiment," *JAMA: The Journal of the American Medical Association*, Vol. 254, No. 14, pp. 1926–1931.
Kenney, Genevieve M., Victoria Lynch, Jennifer Haley, and Michael Huntress (2012) "Variation in Medicaid Eligibility and Participation among Adults: Implications for the Affordable Care Act," *Inquiry*, Vol. 49, No. 3, pp. 231–253.
Kowalski, Amanda (2020) "Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform,"Technical Report w24647, National Bureau of Economic Research, Cambridge, MA.
Levy, Helen G. and David Meltzer (2004) "What Do We Really Know about Whether Health Insurance Affects Health?" in Catherine G. McLaughlin ed. *Health Policy and the Uninsured*, Washington, D.C: Urban Institute Press.
——— (2008) "The Impact of Health Insurance on Health," *Annual Review of Public Health*, Vol. 29, No. 1, pp. 399–409.
Lewis, Randall A. and Justin M. Rao (2015) "The unfavorable economics of measuring the returns to advertising," *Quarterly Journal of Economics*, Vol. 130, No. 4, pp. 1941–1973.
Lewis, Randall A. and David H. Reiley (2014) "Online Ads and Offline Sales: Measuring the Effect of Retail Advertising via a Controlled Experiment on Yahoo!," *Quantitative Marketing and Economics*, Vol. 12, No. 3, pp. 235–266.
Lindo, Jason M. (2015) "Aggregation and the Estimated Effects of Economic Conditions on Health," *Journal of Health Economics*, Vol. 40, pp. 83–96.
Lou, Moghtaderi, A. Markus, and A. Dor (2018) "Impact of Medicaid Expansion on Total Revenue of Community Health Centers by Funding Sources."
Marton, James and Aaron Yelowitz (2015) "Health Insurance Generosity and Conditional Coverage: Evidence from Medicaid Managed Care in Kentucky: Medicaid Conditional Coverage," *Southern Economic Journal*, Vol. 82, No. 2, pp. 535–555.
Maxwell, Scott E. (2004) "The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies.," *Psychological Methods*, Vol. 9, No. 2, pp. 147–163.
McClellan, Chandler (2017) "The Affordable Care Act's Dependent Care Coverage and Mortality," *Medical Care*, Vol. 55, No. 5, p. 6.
McCloskey, Deirdre N and Stephen T Ziliak (1996) "The Standard Error of Regressions," *Journal of Economic Literature*, Vol. 34, No. 1, pp. 97–114.
McCloskey, Donald (1985) "The Loss Function Has Been Mislaid : The Rhetoric of Significance Tests," *American Economic Review: Papers and Proceedings*, Vol. 75, No. 2, pp. 201–205.
Miller, Sarah, Norman Johnson, and Laura Wherry (Forthcoming) "Medicaid and Mortality: New Evidence from Linked Survey and Administrative Data," *The Quarterly Journal of Economics*.
Newhouse, Joseph P. (1993) *Free for All? Lessons from the Rand Health Insurance Experiment*, Cambridge, Mass: Harvard University Press.
NIH (2016) "Reviewer Guidance on Rigor and Transparency : Research Project Grant and Mentored Career Development Applications."
Nolte, Ellen and Martin McKee (2003) "Measuring the health of nations: analysis of mortality amenable to health care," *BMJ*, Vol. 327, No. 7424, p. 1129.
Polsky, Daniel, Michael Richards, Simon Basseyn, Douglas Wissoker, Genevieve M. Kenney, Stephen Zuckerman, and Karin V. Rhodes (2015) "Appointment Availability after Increases in Medicaid Payments for Primary Care," *New England Journal of Medicine*, Vol. 372, No. 6, pp. 537–545.
Powell, David (2018) "Imperfect Synthetic Controls: Did the Massachusetts Health Care Reform Save Lives?", p. 44.
Senn, S. J (2002) "Power is indeed irrelevant in interpreting completed studies," *BMJ*, Vol. 325, No. 7375, pp. 1304–1304.

Simon, Kosali, Aparna Soni, and John Cawley (2017) "The Impact of Health Insurance on Preventive Care and Health Behaviors: Evidence from the First Two Years of the ACA Medicaid Expansions: Impact of Health Insurance on Preventive Care and Health Behaviors," *Journal of Policy Analysis and Management*, Vol. 36, No. 2, pp. 390–417.

Sommers, Benjamin D., Katherine Baicker, and Arnold M. Epstein (2012) "Mortality and Access to Care among Adults after State Medicaid Expansions," *New England Journal of Medicine*, Vol. 367, No. 11, pp. 1025–1034.

Sommers, Benjamin D., Thomas Buchmueller, Sandra L. Decker, Colleen Carey, and Richard Kronick (2013) "The Affordable Care Act Has Led To Significant Gains In Health Insurance And Access To Care For Young Adults," *Health Affairs*, Vol. 32, No. 1, pp. 165–174, PMID: 23255048.

Sommers, Benjamin D., Sharon K. Long, and Katherine Baicker (2014) "Changes in Mortality After Massachusetts Health Care Reform: A Quasi-Experimental Study," *Annals of Internal Medicine*, Vol. 160, No. 9, p. 585.

Stigler, Stephen (1977) "Do robust estimators work with real data?," *Annals of Statistics*, Vol. 5, No. 6, pp. 1055–1098.

Taylor-Robinson, David C., Nicola Maayan, Karla Soares-Weiser, Sarah Donegan, and Paul Garner (2015) "Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance," *Cochrane Database of Systematic Reviews*, Vol. 2015, No. 7.

Weathers, Robert R. and Michelle Stegman (2012) "The Effect of Expanding Access to Health Insurance on the Health and Mortality of Social Security Disability Insurance Beneficiaries," *Journal of Health Economics*, Vol. 31, No. 6, pp. 863–875.

Wehby, George L. and Wei Lyu (2018) "The Impact of the ACA Medicaid Expansions on Health Insurance Coverage through 2015 and Coverage Disparities by Age, Race/Ethnicity, and Gender," *Health Services Research*, Vol. 53, No. 2, pp. 1248–1271.

Wherry, Laura R. and Sarah Miller (2016) "Early Coverage, Access, Utilization, and Health Effects Associated With the Affordable Care Act Medicaid Expansions: A Quasi-Experimental Study," *Annals of Internal Medicine*, Vol. 164, No. 12, p. 795.

Zhang, Le and Andreas Ortmann (2013) "Exploring the Meaning of Significance in Experimental Economics,"Technical report.

Ziliak, Stephen T. and Deirdre N. McCloskey (2004) "Size matters: the standard error of regressions in the American Economic Review," *The Journal of Socio-Economics*, Vol. 33, No. 5, pp. 527–546.

Appendix

Table A1 includes Medicaid expansions through 2016. It is based on combining and reconciling the classification of states as "full expansion," "None," or in- between ("mild" or "substantial" expansion), by Simon et al. (2017), Lou et al. (2018), and Kaiser Family Foundation (2018). Most states could be classified based on their rules for when and to what level they expanded Medicaid for all adults. Arizona required special care; see detailed analysis below. Because our mortality data are annual, we consider New Hampshire to be a 2015 expansion, Alaska to be a 2016 expansion, and Louisiana to be a 2017 expansion, hence beyond our study period. In the expansion details column, ACA Expansion means regular expansion to 138% of FPL, on the date stated in the Effective Date column. In the inclusion/exclusion column, C = control (non-expansion), T = treatment (full expansion), and other states are excluded. Simon et al. (2017) classify early expansion states as mild or substantial expansion, based on their assessment of the extent to which enrollment increased with full Affordable Care Act expansion in 2014. This classification of states based on expansion status is also used in our manuscript (referred to as BHNS). % change in uninsured enrollees (2013-20156) come from SAHIE estimates for ages 18-64 and considering all income groups.

Table A1: Medicaid expansion states: part 1

State	Abbr.	Expansion Details	Effective Date	% change in uninsured enrollees (2013-2016)	Inclusion/ Exclusion	Expansion type	Compare to BHNS
Alabama	AL	None		-5.3	C [.]	None	Consistent
Alaska	AK	Medicaid Expansion	9/1/15	-6.9	T [2016]	Full	Consistent for 2014-2015 (expanded late 2015)
Arizona[1]	AZ	§ 1115 Waiver (100% FPL, but closed to new enrollees in 2011)	2000	-9.1	T[2014]	Full	Consistent
Arkansas[2]	AR	ACA Expansion § 1115 Waiver	1/1/14 1/1/14	-10.8	T [2014] Private Option	Full	Consistent
California[3]	CA	§ 1115 Waiver (LA county) § 1115 Waiver (200% FPL) ACA Expansion	1/1/95 11/1/10 1/1/14	-12.8	Excluded (Early expansion)	Substantial	Consistent
Colorado[4]	CO	§ 1115 Waiver (to 10% of FPL) ACA Expansion	4/1/12 1/1/14	-7.2	T [2014]	Full	Consistent
Connecticut[5]	CT	State Plan Amendment (56% FPL) ACA Expansion	4/1/10 1/1/14	-5.7	Excluded (Early Expansion)	Substantial	Consistent
Delaware[6]	DE	ACA Expansion	1/1/96 1/1/14	-5.8	Excluded (Early Expansion)	Mild	Consistent
District of Columbia[7]	DC	State Plan Amendment (133% FPL) § 1115 Waiver ACA Expansion	7/1/10 12/1/10 1/1/14	-3.4	Excluded (Early expansion)	Mild	Consistent
Florida	FL	None		-9.5	C [.]	None	Consistent
Georgia	GA	None		-7.0	C [.]	None	Consistent
Hawaii[8]	HI	ACA Expansion	8/1/94 1/1/14	-4.5	Excluded (Early expansion)	Substantial	Consistent
Idaho	ID	None		-7.8	C [.]	None	Consistent
Illinois	IL	ACA Expansion	1/1/14	-8.4	T [2014]	Full	Consistent
Indiana	IN	§ 1115 Waiver	2/1/15	-8.1	T [2015]	Full	Consistent
Iowa[9]	IA	§ 1115 Waiver	1/1/14	-6.2	T [2014]	Full	Consistent
Kansas	KS	None		-5.6	C [.]	None	Consistent
Kentucky	KY	ACA Expansion	1/1/14	-12.9	T [2014]	Full	Consistent
Louisiana	LA	ACA Expansion	7/1/16	-9.1	C [.]	None	Consistent
Maine	ME	None		-4.4	C [.]	None	Consistent
Maryland	MD	ACA Expansion	1/1/14	-5.2	T [2014]	Full	Consistent

Notes: [1] Arizona used a § 1115 waiver to expand Medicaid coverage to childless adults up to 100% FPL during 2000-2011. In 2011, the state started to phase out that program (transitioning into Medicaid expansion). Which category Arizona belongs in was unclear based on its rules, so we also examined the extent to which Medicaid enrollment increased in 2014. See details below. [2] Arkansas operated a limited-benefit premium-assistance program for childless adults who worked for small uninsured employers (ARHealthNetworks waiver) prior to the ACA. Arkansas's Medicaid expansion includes a "private option" under which Medicaid-eligible persons receive health insurance from the state insurance exchange, with a small monthly premium. [3] California expanded Medicaid in 2010-2011, in selected counties. [4] Colorado conducted early expansion through a § 1115 waiver in 2012, but only to persons with income ¡10 (ten) % of FPL, and also capped new enrollment at 10,000 individuals. It expanded Medicaid program fully in 2014. We ignore the small expansion in 2012, and treat Colorado as a full expansion state. [5] Connecticut, elected to enact the Medicaid expansion in 2010 through a state amended plan at 56%. Connecticut expanded its Medicaid program fully in 2014. [6] In Delaware, childless adults with incomes up to 100% FPL were eligible for Medicaid through the Diamond State Health Plan waiver, effective on 01/01/1996. [7] DC expanded its Medicaid program at 133% of FPL in 2010. [8] In Hawaii, childless adults with incomes up to 100% FPL were eligible for the state's QUEST Medicaid managed care waiver program, effective on 08/01/1994. [9] Under the IowaCare program, chilDless adults with income below 200% FPL were eligible for health insurance since 2005. However, IowaCare provided limited services in a limited network, so low-income adults in Iowa received a substantial coverage expansion in 2014 (Damiano et al., 2013). During 2014-2015, Iowa residents with income ¡ 100% of FPL were enrolled in Medicaid managed care plans, while those with income of 100-138% of FPL received private insurance obtained through the Iowa health exchange, with premiums waived (a partial "private option"). See `https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/ia/Market-Place-Choice-Plan/ia-marketplace-choice-plan-state-term-app-06012016.pdf`.

Table A1: Medicaid expansion states, part 2

State	Abbr.	Expansion Details	Effective Date	% change in uninsured enrollees (2013-2016)	Inclusion/ Exclusion	Expansion type	Compare to BHNS
Massachusetts[10]	MA	"Romneycare" ACA Expansion	4/12/06 1/1/14	-1.7	Excluded	Mild	Consistent
Michigan	MI	ACA Expansion	4/1/14	-7.9	T [2014]	Full	Consistent
Minnesota[11]	MN	State Plan Amendment (75% FPL) § 1115 Waiver (200% FPL)	3/1/10 8/1/10	-4.6	Excluded (Early Expansion)	Substantial	Consistent
Mississippi	MS	ACA Expansion None	1/1/14	-6.5	C [.]	None	Consistent
Missouri	MO	§ 1115 Waiver (St. Louis County Only) (200% FPL) None	7/1/12	-4.9	C [.]	None	Consistent
Montana	MT	ACA Expansion	1/1/16	-10.7	T [2016]	Full	Consistent for 2014-2015 (expanded in 2016)
Nebraska	NE	None		-2.3	C [.]	None	Consistent
Nevada	NV	ACA Expansion	1/1/14	-11.3	T [2014]	Full	Consistent
New Hampshire[12]	NH	§ 1115 Waiver	8/15/14	-6.3	T [2015]	Full	Consistent (expanded late 2014)
New Jersey[13]	NJ	§ 1115 Waiver (23% FPL) ACA Expansion	4/1/11 1/1/14	-7.2	T [2014]	Full	Consistent
New Mexico	NM	ACA Expansion	1/1/14	-14.5	T [2014]	Full	Consistent
New York[14]	NY	§ 1115 waiver ACA Expansion	10/1/01 1/1/14	-6.1	Excluded (Early expansion)	Mild	Consistent
North Carolina	NC	None		-6.4	C [.]	None	Consistent
North Dakota	ND	ACA Expansion	1/1/14	-4.4	T [2014]	Full	Consistent
Ohio	OH	ACA Expansion	1/1/14	-7.6	T [2014]	Full	Consistent
Oklahoma	OK	None		-4.5	C [.]	None	Consistent
Oregon	OR[15]	ACA Expansion	1/1/14	-11.1	T [2014]	Full	Consistent

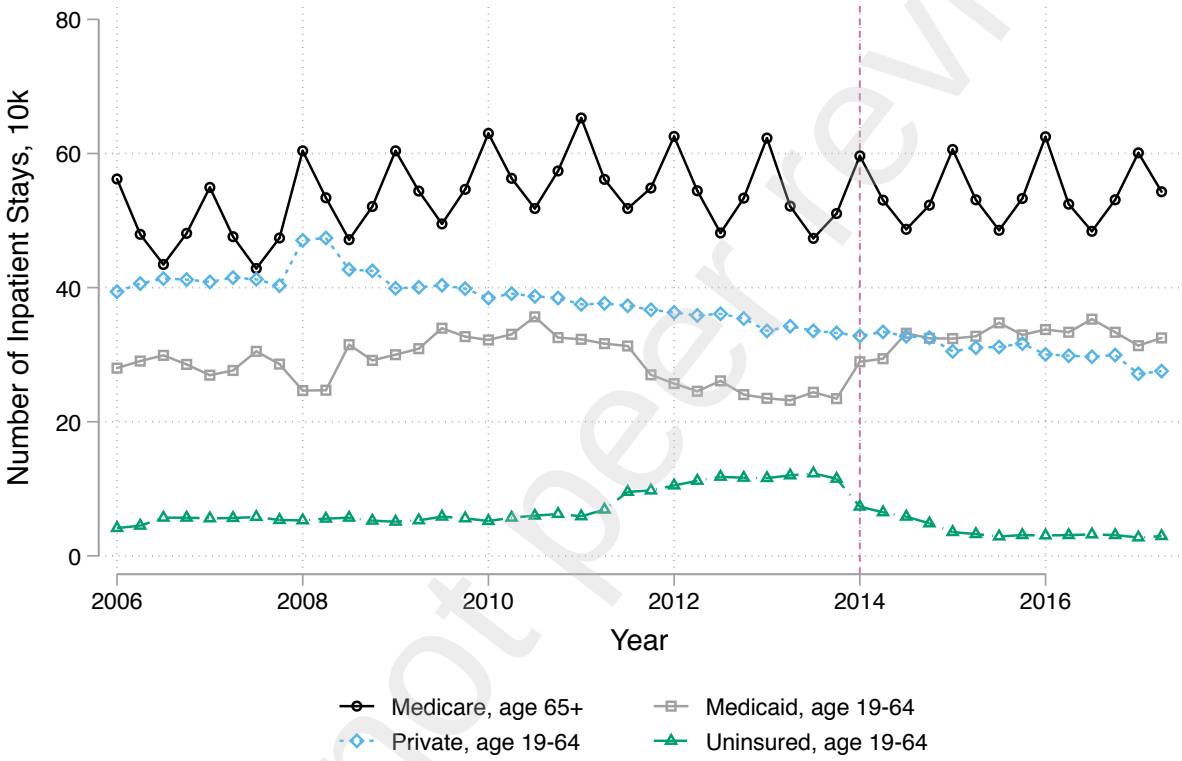
Notes: [10] Massachusetts implemented reforms to expand insurance coverage to low-income adults in 2006. [11] Minnesota conducted early expansion in 2010 two ways. Persons with income $\leq$ 75%FPL were insured through Medical Assistance Medicaid, funded through a State Plan Amendment, persons with income from 75 200% of FPL were insured through MinnesotaCare, funded through a § 1115 Waiver, which had limited benefits and cost-sharing. [12] New Hampshire implemented a "private option" (mandatory purchase of subsidized private insurance, instead traditional Medicaid, in 2016. See `https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/nh/health-protection-program/nh-health-protection-program-premium-assistance-appvl-amend-req-06232015.pdf`. [13] New Jersey's expansion in 2011 only extended to 23% FPL; we therefore treated it as a full expansion state. [14] In New York, childless adults up to 78% FPL were eligible for the Medicaid (Home Relief) waiver program and childless adults up to 100% FPL were eligible for the Family Health Plus waiver program (Heberlein et al., 2011). [15] In 2008, Oregon enacted a small Medicaid expansion for low-income adults through a lottery among applicants. However, less than one-third of the 90,000 people on the waitlist were selected to apply for Medicaid in 2008 (Baicker et al., 2013), some of the denied applicants were then enrolled in 2010. We treat Oregon as full expansion due to the small size of this earlier expansion.

Table A1: Medicaid expansion states, part 3

State	Abbr.	Expansion Details	Effective Date	% change in uninsured enrollees (2013-2016)	Inclusion/ Exclusion	Expansion type	Compare to BHNS
Pennsylvania	PA	ACA Expansion	1/1/15	-5.3	T [2015]	Full	Consistent
Rhode Island	RI	ACA Expansion	1/1/14	-9.1	T [2014]	Full	Consistent
South Carolina	SC	None		-7.2	C [.]	None	Consistent
South Dakota	SD	None		-4.4	C [.]	None	Consistent
Tennessee	TN	None		-6.1	C [.]	None	Consistent
Texas	TX	None		-6.4	C [.]	None	Consistent
Utah	UT	None		-6.1	C [.]	None	Consistent
Vermont	VT[16]	§ 1115 Waiver ACA Expansion	1/1/96 1/1/14	-5.0	Excluded (Early expansion)	Mild	Consistent
Virginia	VA	None		-4.3	C [.]	None	Consistent
Washington[17]	WA	§ 1115 Waiver (133% FPL) ACA Expansion	1/3/11 1/1/14	-10.3	T [2014]	Full	Consistent
West Virginia	WV	ACA Expansion	1/1/14	-12.0	T [2014]	Full	Consistent
Wisconsin[18]	WI	New eligibility for BadgerCare but not ACA Expansion	2009	-5.2	Excluded	Substantial	Consistent
Wyoming	WY	None		-2.5	C [.]	None	Consistent

Notes: [16] In Vermont, childless adults up to 150% FPL were eligible for Medicaid equivalent coverage through the Vermont Health Access Plan waiver program (Heberlein et al., 2011). Vermont Health Access Plan (Sec. 1115 waiver) was approved in 1995 and effective in 1996. [17] Washington's early expansion was limited to prior state plan enrollees (Sommers et al., 2013). [18] Although Wisconsin was not an ACA expansion state, the state received federal approval to offer Medicaid to childless adults below 100% FPL through the BadgerCare program as of 2009 (Gates and Rudowitz, 2014).

Figure A1: Number of Arizona inpatient stays by payor and year



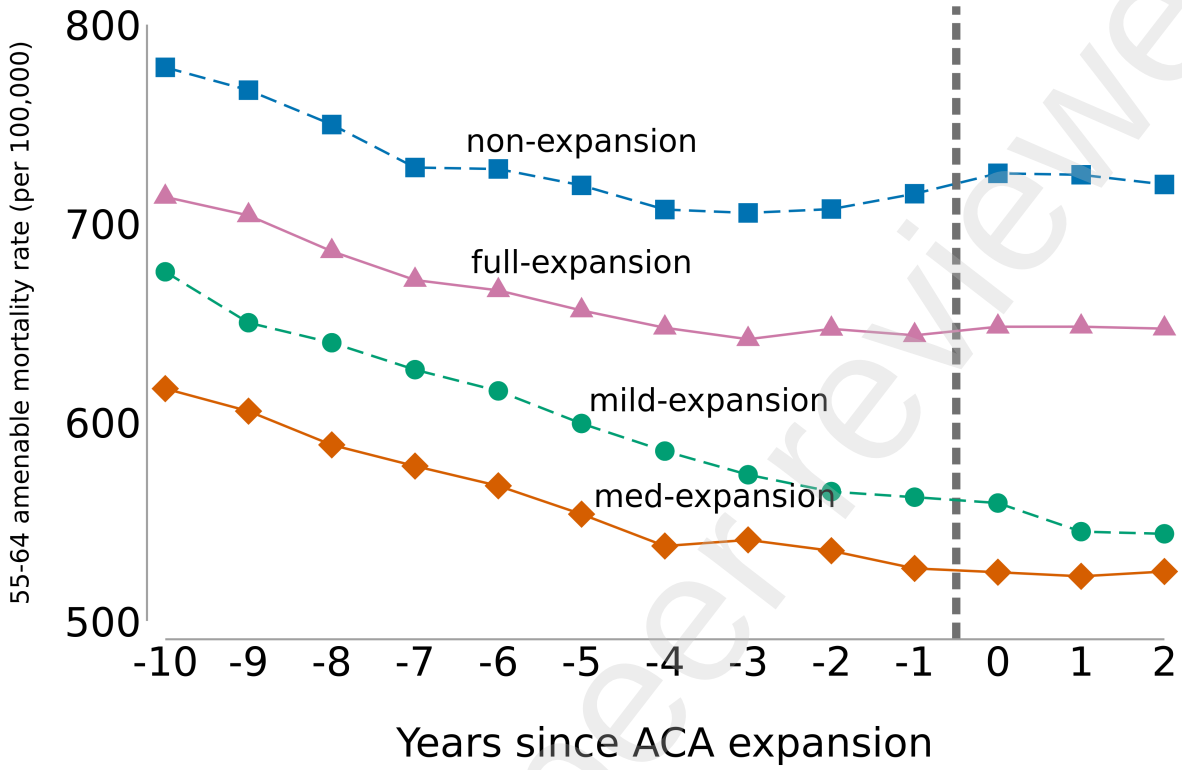
Author reproduction of HCUP figure using HCUP Fast Stats at `https://www.hcup-us.ahrq.gov/faststats/StatePayerServlet?state1=AZ`. Accessed June 2018. Arizona had a S.1931 program providing Medicaid up to 106% FPL for parents. It also had a limited program for 7 childless adults, under a § 1115 waiver, starting in 2001, which was closed to new entrants since 2011.23 Whether to treat Arizona as a full expansion state or an early expansion state turns on how many childless adults were still covered at the ACA onset in 2014, given churn in eligibility. The tail off in hospital admissions with Medicaid payment, and jump at the start of 2014 (with uninsured admissions showing the opposite pattern), persuades us that Arizona should be treated as a regular expansion state.

Table A2: Covariate Balance for Full-Expansion and Non-Expansion States

	Full Expansion States (1)	Non-Expansion States (2)	Difference t-stat (3)	Normalized Difference (4)
% 55-64 years	12.68	12.35	0.31	0.08
% 65-74 years	16.61	16.85	0.67	–0.05
% Female	50.23	50.43	0.76	–0.03
% Female, 55-64 years	4.96	4.85	0.21	0.05
% Female, 65-74 years	7.27	7.35	0.22	–0.03
% White	84.15	87.17	0.81	–0.18
% White, 55-64 years	9.97	10.23	0.94	–0.06
% White, 65-74 years	13.81	14.53	0.94	–0.14
% Black	9.59	8.29	0.62	0.10
% Black, 55-64 years	1.79	1.53	0.63	0.11
% Black, 65-74 years	1.77	1.55	0.66	0.09
% Hispanic	4.03	2.98	0.52	0.11
% Hispanic, 55-64 years	0.59	0.37	0.81	0.14
% Hispanic, 65-74 years	0.63	0.46	0.54	0.09
% In Poverty	14.92	13.42	1.44	0.24
% Managed Care Penetration	24.86	22.92	0.40	0.18
% Disabled (ages 18-64)	16.36	16.83	0.59	–0.08
Mean Per Capita Income	40364.57	41696.06	0.52	–0.15
Median Household Income	51284.70	52565.73	0.29	–0.12
Unemployment Rate, 16+	8.68	7.55	2.81	0.38
% with Diabetes	8.91	8.67	0.45	0.12
% Physically Inactive	22.59	22.38	0.13	0.04
% Obese	27.97	27.74	0.04	0.05
% Smoker	21.80	21.51	0.42	0.07
Physicians/1,000 people	3.12	3.04	0.48	0.05
% Uninsured (18-64 years)	18.69	18.27	0.32	0.07
% Uninsured (50-64 years)	12.88	12.91	0.02	–0.01
% Uninsured (18-64 years), $\leq$138% FPL	37.80	39.57	0.33	–0.23
% Uninsured (50-64 years), $\leq$138% FPL	32.87	34.97	0.81	–0.32
Death Rate	859.58	845.20	0.34	0.05
Death Rate 55-64 years	852.93	797.00	1.41	0.16
Death Rate 65-74 years	1861.27	1786.72	0.93	0.12
Non-amenable Death Rate	589.41	576.92	0.38	0.06
Non-amenable Death Rate 55-64 years	644.03	607.40	1.11	0.12
Npn-amenable Death Rate 65-74 years	1498.39	1436.08	0.89	0.12
Amenable Death Rate	270.16	268.29	0.16	0.02
Amenable Death Rate 55-64 years	208.91	189.60	1.71	0.14
Amenable Death Rate 65-74 years	362.88	350.64	0.50	0.06

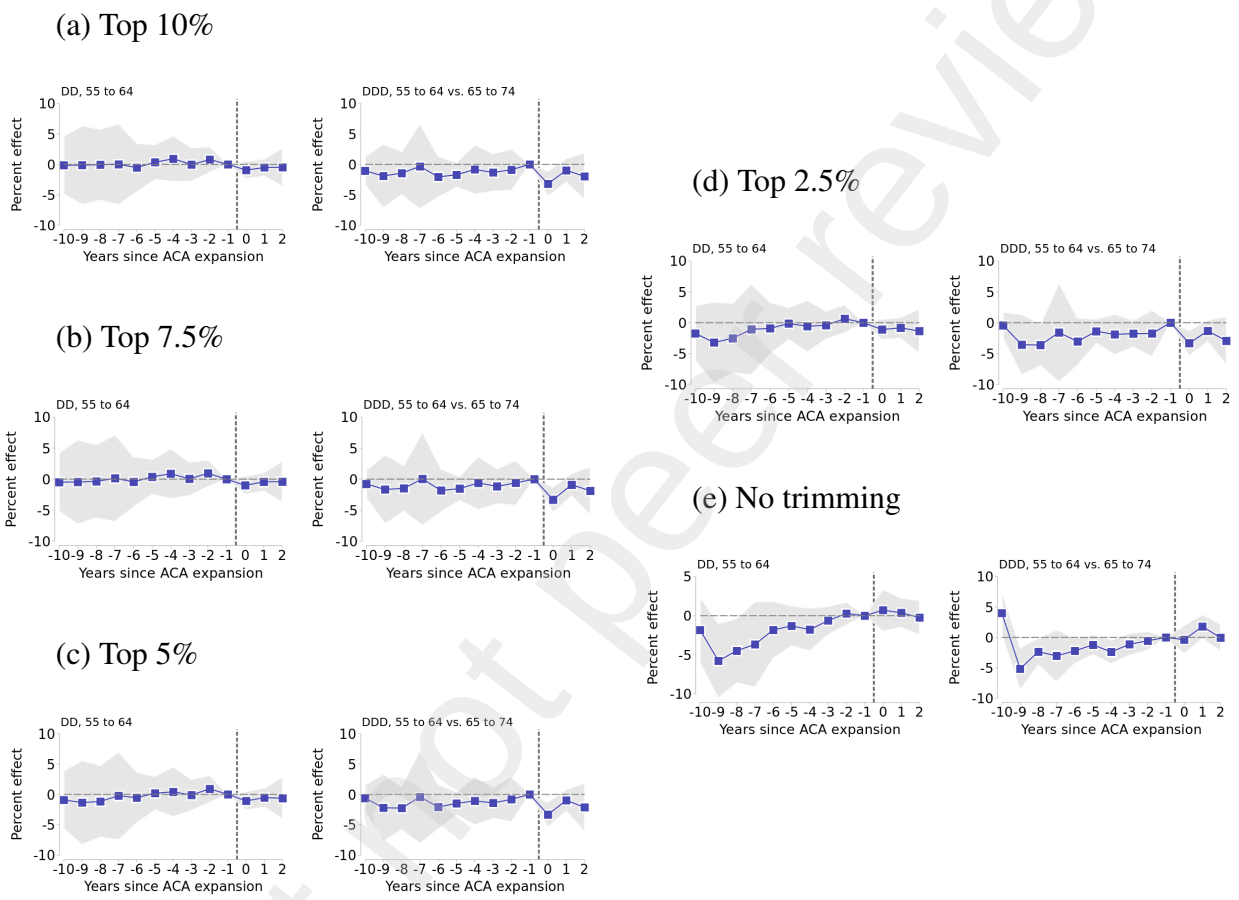
Note: Table shows summary statistics for county-level covariates and mortality for Full-Expansion and Non-Expansion states during pre-expansion period, using ATT$\times$ population weights. T-statistics use two-sample t-test for difference and robust standard errors with state clusters. Normalized difference is a sample-size independent measure of the difference between two means, scaled by standard deviation). State groups are defined in Table A1. Mortality rates are per 100,000 persons. Dollar amounts are in 2010 $. See Table A5 in the Appendix to see the same Table, but with population weights instead of ATT$\times$Pop weights.

Figure A3: Time trends in amenable mortality for persons aged 55-64



Note: Number along the x-axis indicates years since most Medicaid expansion. For non-expansion states, $t = 0$ corresponds to 2014. Years included in this figure are 2004 to 2016. Health care amenable causes of death follow Nolte and McKee (2003).

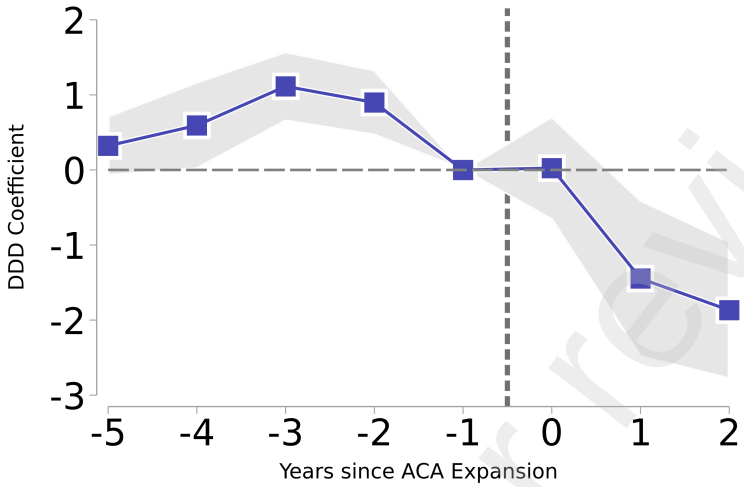
Figure A2: Event studies with different windsorization of weights



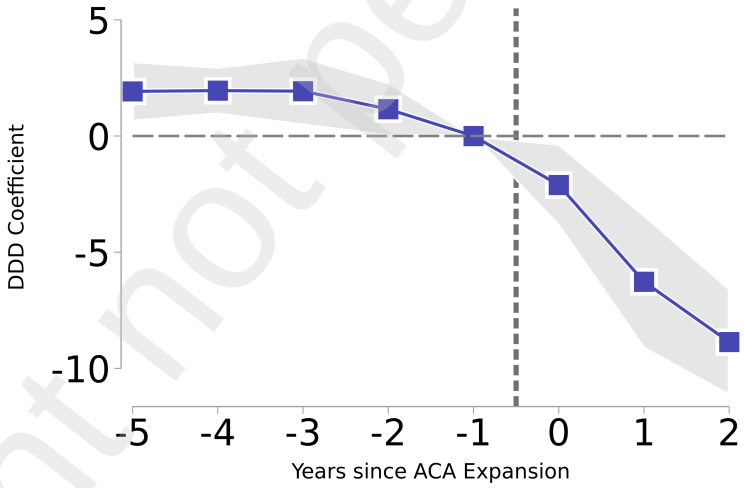
Note: Number along the x-axis indicates years since most Medicaid expansion. For non-expansion states, $t = 0$ corresponds to 2014. Years included in this figure are 2004 to 2016. Since each dependent variable is a natural log transform of the healthcare amendable death rate per 100,000, we transform each regression coefficient by 100(exp(coef.)-1) and the standard errors using the delta method. Thus each reported coefficient can be interpreted as the percent effect of Medicaid expansion t years ago on the healthcare amenable mortality rate for a given age group (i.e., A coefficient of -2 would mean that the mortality rate would decrease by 2%). Point estimates are depicted by blue squares and come from regressions analogous to Equation 2, except that in each sub-figure a different trimming is used. 95% confidence intervals are displayed by gray area and are calculated using robust standard errors clustered at the state level.

Figure A4: The impact of Medicaid expansion on the percentage uninsured aged 55 to 64 is the largest for lower income individuals

A. No income restrictions (as in mortality data)



B. Restricting to those with income $\leq$ 138% FPL



Note: Number along the x-axis indicates years since most Medicaid expansion. For non-expansion states, $t = 0$ corresponds to 2014. Data come from the American Community Survey and are at the county-level covering 2008 to 2016. Each reported coefficient can be interpreted as the percent effect of Medicaid expansion t years ago on the percentage of those without health insurance in each group. Point estimates are depicted by blue squares and come from regressions analogous to Equation 2, but for the triple difference specification. 95% confidence intervals are displayed by gray area and are calculated using robust standard errors clustered at the state level.

Table A3: Treatment effect estimates by subgroup

	State-level			County-level		
	DD 55-64	DD 65-74	DDD	DD 55-64	DD 65-74	DDD
	(1)	(2)	(3)	(4)	(5)	(6)
Healthcare amenable mortality	-2.32**	-0.98	-1.53*	-0.65	0.41	-0.85
	(1.01)	(0.89)	(0.77)	(1.74)	(0.80)	(0.94)
Non-amenable mortality	1.45	3.32***	-1.73	2.60	0.90	1.58
	(1.35)	(0.76)	(1.12)	(2.36)	(1.12)	(1.79)
All mortality	-1.27	-0.18	-1.32*	0.28	0.57	-0.18
	(0.97)	(0.77)	(0.72)	(1.70)	(0.68)	(1.01)
Respiratory mortality	-5.34***	-4.04**	-2.69*	-7.19	-2.64	-5.75
	(1.56)	(1.54)	(1.47)	(4.28)	(2.49)	(6.73)
Cancer mortality	-1.50*	0.25	-1.71	0.73	1.73***	-0.52
	(0.82)	(0.86)	(1.05)	(1.83)	(0.60)	(1.72)
Cardiac mortality	-0.97	-0.55	-0.80	0.58	-0.15	0.93
	(1.17)	(1.13)	(0.91)	(1.77)	(0.83)	(1.52)
HIV mortality	-2.40	0.97	-5.33	0.57	10.63	-7.16
	(3.94)	(5.71)	(9.17)	(4.18)	(7.69)	(9.51)
Diabetes mortality	-3.42	-2.06	-1.25	-4.44	-2.35	-2.66
	(2.58)	(2.50)	(1.26)	(2.74)	(5.05)	(2.42)
Male, amenable mortality	-1.48	-0.10	-1.88**	-0.15	1.24	-1.01
	(1.10)	(0.86)	(0.83)	(1.91)	(1.02)	(1.13)
Female, amenable mortality	-3.51***	-2.14*	-1.18	-1.45	-0.71	-0.31
	(1.09)	(1.07)	(1.01)	(1.64)	(0.87)	(1.47)
White, amenable mortality	-2.18**	-1.41	-0.89	0.31	0.82	-0.08
	(0.93)	(0.90)	(0.89)	(1.68)	(0.76)	(1.04)
Black, amenable mortality	-3.72***	0.99	-5.24***	-5.82*	0.44	-8.55***
	(1.15)	(1.11)	(1.21)	(3.17)	(2.82)	(2.85)
Hispanic, amenable mortality	-9.44**	-3.39	-5.67	-14.81**	-5.61	-5.33
	(3.75)	(3.41)	(3.73)	(6.66)	(6.18)	(4.61)
Weights	ATTxPop	ATTxPop	ATTxPop	ATTxPop	ATTxPop	ATTxPop
Unit fixed-effects (state or county)	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed-effects	Yes	Yes	Yes	Yes	Yes	Yes
Covariates	Yes	Yes	Yes	Yes	Yes	Yes

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors clustered at the state level in parentheses. Each DD regression corresponds to Equation 1 while each triple-difference regression corresponds to Equation 3, except that we change the subgroup or cause of death as noted. Years included in this table are 2004 to 2016. Each dependent variable is a natural log transform of the death rate per 100,000 population for the respective group. We transform each regression coefficient by 100(exp(coef.)-1) and the standard errors using the delta method. Thus each reported coefficient can be interpreted as the percent effect of Medicaid expansion on the specific mortality rate for a given age group (i.e., a coefficient of -2 would mean that the mortality rate would decrease by 2%). Covariates include % Medicare Advantage Penetration, % Medicare enrollees receiving social security disability benefits, Per Capita Income, Median Household Income, % Persons in Poverty, Unemployment Rate, 16+, Number of individuals with diabetes as percent of the population, Age-adjusted percent of the population that is physically inactive, Percent of population that is obese, Smoking Prevalence (all genders), Non-Federal Active MDs per 1000 people.

Table A4: The effect of Medicaid expansion on the percentage uninsured by age group and demographic, using population weights

	ACS (County-level)			SAHIE	
	DD 55-64	DD 65-74	DDD	State DD 50-64	County DD 50-64
	(1)	(2)	(3)	(4)	(5)
All	–1.19***	–0.01	–1.16**	–1.23**	–1.21**
	(0.42)	(0.06)	(0.47)	(0.51)	(0.47)
Female	–1.32***	0.01	–1.29**	–1.35**	–1.31***
	(0.46)	(0.07)	(0.50)	(0.53)	(0.48)
Male	–1.07**	–0.03	–1.03**	–1.14**	–1.10**
	(0.43)	(0.06)	(0.48)	(0.49)	(0.47)
White (non-Hispanic)	–1.51***	–0.02	–1.37***	–1.54***	
	(0.28)	(0.04)	(0.35)	(0.37)	
Black (non-Hispanic)	–2.25***	0.09	–2.63**	–2.09**	
	(0.80)	(0.21)	(1.03)	(0.78)	
Other (non-Hispanic)	–1.44	–0.57	–2.12*		
	(0.97)	(0.45)	(1.11)		
Hispanic	–0.04	0.12	0.09	–0.38	
	(1.47)	(0.52)	(1.59)	(1.18)	
Elementary school	–2.74**	–0.08	–2.93**		
	(1.32)	(0.39)	(1.09)		
High school incomplete	–4.24***	–0.27	–4.26***		
	(0.71)	(0.20)	(0.83)		
High school complete	–1.48**	0.08	–1.37**		
	(0.56)	(0.07)	(0.65)		
Some college	–0.87**	–0.07	–0.79**		
	(0.36)	(0.05)	(0.38)		
Below 138% FPL	–6.54***	–0.09	–6.26***	–6.38***	–6.43***
	(0.88)	(0.25)	(1.00)	(0.95)	(0.95)
138%-400% FPL	–1.01	–0.01	–0.91	–1.00	–1.27*
	(0.60)	(0.06)	(0.67)	(0.76)	(0.71)
Observations	369	369	738	451	19628
Unit fixed-effects (state or county)	Yes	Yes	Yes	Yes	Yes
Year fixed-effects	Yes	Yes	Yes	Yes	Yes
Covariates	Yes	Yes	Yes	Yes	Yes
Weights used	Pop	Pop	Pop	Pop	Pop

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors clustered at the state-level are reported in parentheses. Each estimate comes from a separate regression and the point estimate can be interpreted as the effect of Medicaid expansion on the percent uninsured for a given group (i.e., a coefficient of -2 would mean that those without insurance in that group decreased by 2 percentage points). ACS data refers to the American Community Survey and are at the county-level covering 2008 to 2016. SAHIE data refers to the Census Bureau's Small Area Health Insurance Estimates and are reported at both state and county-levels covering 2006 to 2016. See Table 2 in main text to see the same Table, but with ATT×Pop weights rather than population weights.

Table A5: Covariate Balance for Full-Expansion and Non-Expansion States Using Population Weights.

	Full Expansion States (1)	Non-Expansion States (2)	Difference t-stat (3)	Normalized Difference (4)
% 55-64 years	12.68	13.72	2.61	–0.26
% 65-74 years	16.61	17.64	2.59	–0.22
% Female	50.23	49.37	2.12	0.15
% Female, 55-64 years	4.96	5.38	2.11	–0.19
% Female, 65-74 years	7.27	7.62	1.56	–0.12
% White	84.15	75.02	3.11	0.56
% White, 55-64 years	9.97	9.44	0.87	0.14
% White, 65-74 years	13.81	13.17	1.45	0.12
% Black	9.59	15.52	2.69	–0.48
% Black, 55-64 years	1.79	2.96	2.69	–0.48
% Black, 65-74 years	1.77	2.91	2.62	–0.45
% Hispanic	4.03	7.85	1.09	–0.39
% Hispanic, 55-64 years	0.59	1.06	0.90	–0.30
% Hispanic, 65-74 years	0.63	1.25	1.06	–0.32
% In Poverty	14.92	17.16	3.14	–0.35
% Managed Care Penetration	24.86	23.64	0.77	0.11
% Disabled (ages 18-64)	16.36	17.58	1.59	–0.20
Mean Per Capita Income	40364.57	37594.69	2.13	0.30
Median Household Income	51284.70	46819.56	2.29	0.42
Unemployment Rate, 16+	8.68	8.12	1.47	0.19
% with Diabetes	8.91	9.77	2.78	–0.43
% Physically Inactive	22.59	24.52	1.97	–0.39
% Obese	27.97	29.09	1.29	–0.26
% Smoker	21.80	21.49	0.44	0.08
Physicians/1,000 people	3.12	2.67	2.63	0.27
% Uninsured (18-64 years)	18.69	25.06	3.23	–1.09
% Uninsured (50-64 years)	12.88	17.44	3.33	–1.03
% Uninsured (18-64 years), $\leq$138% FPL	37.80	46.97	3.46	–1.21
% Uninsured (50-64 years), $\leq$138% FPL	32.87	39.48	3.73	–1.01
Death Rate	859.58	825.58	0.50	0.12
Death Rate 55-64 years	852.93	933.79	2.20	–0.23
Death Rate 65-74 years	1861.27	1939.68	0.71	–0.13
Non-amenable Death Rate	589.41	557.11	0.81	0.16
Non-amenable Death Rate 55-64 years	644.03	701.74	1.84	–0.20
Non-amenable Death Rate 65-74 years	1498.39	1540.86	0.35	–0.08
Amenable Death Rate	270.16	268.47	0.40	0.02
Amenable Death Rate 55-64 years	208.91	232.05	2.60	–0.17
Amenable Death Rate 65-74 years	362.88	398.81	1.71	–0.17

Note: Table shows summary statistics for county-level covariates and mortality for Full-Expansion and Non-Expansion states during the pre-expansion period using county population weights. t-statistics use two-sample t-test for difference and robust standard errors with state clusters. Normalized difference is a sample-size independent measure of the difference between two means, scaled by standard deviation). State groups are defined in Table A1. Mortality rates are per 100,000 persons. Dollar amounts are in 2010 $. See Table A2 in the main text to see the same Table, but with ATT$\times$Pop weights instead of population weights.

A Example code

This is an example of the type of simulated power analysis done in our paper. This example is done with publicly available data. You can find the code, data, and output for this example hosted on https://github.com/hollina/simulated-power-analysis.

This set-up is designed to mimic a typical DiD setting. Here we will compare 23 randomly chosen treated states to 18 randomly chosen control states. We will impose a series of treatment effects that gradually increase in magnitude and report whether or not these imposed treatment effects are detectable. We will vary the set of randomly chosen treated states. We will calculate the minimum detectable effect size at various power and significance levels. We will also explore a measure of believability, which is based upon Gelman and Carlin (2014) measures of sign and magnitude error.

In this simple design we used 5 years of pre-expansion data and 3 years of post-expansion data. Both state and year fixed-effects are included. Regressions are weighted by state-population and standard errors will be clustered at the state-level. The dependent variable will be the natural log of the all-cause non-elderly mortality rate per 100,000.

This stata code is simply an example of our simulated power analysis and is not an attempt to identify the impact of Medicaid expansion on mortality. Importantly, changing the research design (e.g. adding control variables, shifting to the county-level, changing the cause of death, using propensity score weights, or using a synthetic control estimator) will impact power. Our approach could be easily modified to accommodate any of these alternative research designs. Any improvements to the research design will very likely increase power and decrease the minimum detectable effect size.

Stata code

```
// Change directory
cd "~/Documents/GitHub/health_insurance_and_mortality"

////////////////////////////////////////////////////////////////////////////
// Initial Set-up
// ====================
```

```
// Here we will set-up the power analysis and choose various required parameters/options

// First we clear the memory

clear all

set matsize 10000

// Choose the number of datasets we want to compose each estimate. For example, if we ch

local max_dataset_number  = 500


// Pick the number of psuedo-post-expansion years

local number_post_years = 3
local last_year = 2013-`number_post_years'+1

// Set number of psuedo-pre-expansion years

local number_pre_years = 5
local first_year = `last_year'-`number_pre_years'

// Set effect size step and max value in percent terms (0-1)

local step_size = .0025 // Quarter of a percent
local end_value = .05 // End at 5%

// Create a local macro from the choices above

local step_macro
forvalues x = 0(`step_size')`end_value' {
   local step_macro `step_macro'  `x'
}
```

```
// Determine the length of the macro above, so percent complete can be displayed later

local num :  word count ‘step_macro’
local num = ‘num’
local max_steps = ‘num’
di ‘max_steps’

// Calculate the max number of rows so percent complete can be displayed later

local max_row = ‘max_dataset_number’*‘num’


////////////////////////////////////////////////////////////////////////////
// Import and clean mortality data
// ====================

// Import data extracted from [CDC wonder](https://wonder.cdc.gov/). All cause mortality

import delimited "state_level_public_data_example/data/Multiple Cause of Death, 1999-201


// Drop total variables

drop if missing(year)


// Drop unneeded variables from CDC Wonder

drop notes


// Drop years after expansion

drop if year>=2014
```

```
// Drop if year before first desired year

drop if year<‘first_year’


// Change state name to be state postal code

replace state ="AL" if state=="Alabama"
replace state ="AK" if state=="Alaska"
replace state ="AZ" if state=="Arizona"
replace state ="AR" if state=="Arkansas"
replace state ="CA" if state=="California"
replace state ="CO" if state=="Colorado"
replace state ="CT" if state=="Connecticu "
replace state ="DE" if state=="Delaware"
replace state ="DC" if state=="District of Columbia"
replace state ="FL" if state=="Florida"
replace state ="GA" if state=="Georgia"
replace state ="HI" if state=="Hawaii"
replace state ="ID" if state=="Idaho"
replace state ="IL" if state=="Illinois"
replace state ="IN" if state=="Indiana"
replace state ="IA" if state=="Iowa"
replace state ="KS" if state=="Kansas"
replace state ="KY" if state=="Kentucky"
replace state ="LA" if state=="Louisiana"
replace state ="ME" if state=="Maine"
replace state ="MD" if state=="Maryland"
replace state ="MA" if state=="Massachusetts"
replace state ="MI" if state=="Michigan"
replace state ="MN" if state=="Minnesota"
replace state ="MS" if state=="Mississippi"
replace state ="MO" if state=="Missouri"
replace state ="MT" if state=="Montana"
replace state ="NE" if state=="Nebraska"
replace state ="NV" if state=="Nevada"
```

```
replace state ="NH" if state=="New Hampshire"
replace state ="NJ" if state=="New Jersey"
replace state ="NM" if state=="New Mexico"
replace state ="NY" if state=="New York"
replace state ="NC" if state=="North Carolina"
replace state ="ND" if state=="North Dakota"
replace state ="OH" if state=="Ohio"
replace state ="OK" if state=="Oklahoma"
replace state ="OR" if state=="Oregon"
replace state ="PA" if state=="Pennsylvania"
replace state ="RI" if state=="Rhode Island"
replace state ="SC" if state=="South Carolina"
replace state ="SD" if state=="South Dakota"
replace state ="TN" if state=="Tennessee"
replace state ="TX" if state=="Texas"
replace state ="UT" if state=="Utah"
replace state ="VT" if state=="Vermont"
replace state ="VA" if state=="Virginia"
replace state ="WA" if state=="Washington"
replace state ="WV" if state=="West Virginia"
replace state ="WI" if state=="Wisconsin"
replace state ="WY" if state=="Wyoming"

// Add expansion status to each state

gen expansion4=0
label define expansion4 0 "0. Non-expansion" 1 "1. Full expansion" ///
    2 "2. Mild expansion" 3 "3. Substantial expansion"
label values expansion4 expansion4



local full AZ AR CO IL IA KY MD NV NM NJ ND OH OR RI WV  WA
foreach x in ‘full’ {
    replace expansion4=1 if state=="‘x’"
}
local mild DE DC MA NY VT
```

```
foreach x in ‘mild’ {
    replace expansion4=2 if state=="‘x’"
}
local medium CA CT HI MN WI
foreach x in ‘medium’ {
    replace expansion4=3 if state=="‘x’"
}

// Account for mid-year expansions

replace expansion4=1 if state=="MI"  //MI expanded in April 2014
replace expansion4=1 if state=="NH"  //NH expanded in August 2014
replace expansion4=1 if state=="PA"  //PA expanded in Jan 2015
replace expansion4=1 if state=="IN"  //IN expanded in Feb 2015
replace expansion4=1 if state=="AK"  //AK expanded in Sept 2015
replace expansion4=1 if state=="MT"  //MT expanded in Jan 2016
replace expansion4=1 if state=="LA"  //LA expanded in July 2016

// Keep only full or non-expansion states

drop if expansion4==2 | expansion4==3


// Store number of expansion states

distinct statecode  if expansion4==1
scalar number_expand = r(ndistinct)

// Save data to be called in power analysis
// ====================
// Save temporary dataset to be called

compress
save "state_level_public_data_example/temp/temp_data.dta", replace

////////////////////////////////////////////////////////////////////////////
```

```
// Run simulated power analysis
// ====================
// Start a timer to show how long this takes

timer on 1



//  Create matrix to store results

matrix b_storage = J(‘max_dataset_number’,‘max_steps’,.)
matrix se_storage = J(‘max_dataset_number’,‘max_steps’,.)
matrix p_storage = J(‘max_dataset_number’,‘max_steps’,.)

// Start run counter for % update

local run_count = 1



// Run a loop. Performing the power analysis once for each of the desired number of data

forvalues dataset_number = 1(1)‘max_dataset_number’    {
    // Display the dataset number
    qui di "‘dataset_number’"


  // Open main dataset for analysis
    qui use "state_level_public_data_example/temp/temp_data.dta", clear

    // Set seed for reproducibility. We want the seed to be the same within a dataset.
    qui local rand_seed = 1234 + ‘dataset_number’
    qui set seed   ‘rand_seed’

    ////////////////////////////////////////////////////////////////////////////
    // Generate a random variable for each state, then the first N in rank will be
    // considered expansion states. Where N is # of expansion states
```

```
qui bysort statecode: gen random_variable = runiform() if _n==1
qui bysort statecode: carryforward random_variable, replace

// Rank the states
qui egen rank = group(random_variable)

// Given this random ordering of states, assign expansion status to the # set above
qui gen expansion = 0
qui replace expansion=1 if rank <=number_expand

// Do this same thing for the treatment variable
qui gen treatment = 0
qui replace treatment = 1 if expansion==1 & year>=`last_year'

// Create Post variable
qui gen post = 0
qui replace post =1 if year>=`last_year'

// Generate a death rate with no effect
qui gen death_rate = (deaths/population)*100000

// Gen order variable
qui gen order = _n

////////////////////////////////////////////////////////////////////////////
// Create a reduced deaths variable by a given percentage using the binomial for eac
qui local counter = 1

foreach x in `step_macro' {
    qui gen reduced_deaths_`counter' = 0
    qui replace reduced_deaths_`counter' = rbinomial(deaths,`x') if treatment==1
    qui replace reduced_deaths_`counter'=0 if missing(reduced_deaths_`counter')

    qui gen deaths_`counter' = deaths - reduced_deaths_`counter'
    qui replace deaths_`counter'=0 if missing(deaths_`counter')
```

```
    qui gen death_rate_‘counter’= ln((deaths_‘counter’/population)*100000+1)

    // Move the row and counter one forward
    qui local counter = ‘counter’ + 1
}

////////////////////////////////////////////////////////////////////////
// Run regression of treatment on reduced deaths variable for each effect size

// Reset the counter
qui local counter = 1

forvalues counter = 1(1)‘num’ {

    qui reghdfe death_rate_‘counter’ ///
        i.treatment ///
        i.post i.expansion ///
        [aweight=population] ///
        ,  absorb(statecode year)  vce(cluster statecode)

    //Evaluate Effect using nlcom. Since we will do a tranform of the log results
    qui nlcom 100*(exp(_b[1.treatment])-1)

    // Store in matrices
    mat b = r(b)
    mat V = r(V)

    scalar b = b[1,1]
    scalar se_v2 = sqrt(V[1,1])
    scalar p_val = 2*ttail(‘e(df_r)’,abs(b/se_v2))

    mat b_storage[‘dataset_number’, ‘counter’] = b
    mat se_storage[‘dataset_number’, ‘counter’] = se_v2
    mat p_storage[‘dataset_number’, ‘counter’] = p_val

    // Display Percent Complete
```

```
        di "////////////////////////////////////////////////////////////////////"
        di "//////////////////////////Percent Complete//////////////////////////"
        di ((`run_count'-1)/`max_row')*100
        di "////////////////////////////////////////////////////////////////////"

        qui local run_count = `run_count' + 1
        qui local counter = `counter' + 1
    }
}
// Stop timer
timer off 1
timer list



// Erase temporary dataset used for analysis

erase "state_level_public_data_example/temp/temp_data.dta"



// Save power results as csv
quietly {
    clear
    svmat b_storage
    format * %20.5f

    svmat se_storage
    format * %20.5f

    svmat p_storage
    format * %20.5f
    ds p_*
    foreach x in `r(varlist)' {
        replace `x' = 0 if `x' < .00001
    }
```

```
    export delimited using "state_level_public_data_example/temp/power_simulation_storag
}

////////////////////////////////////////////////////////////////////////////
// Clean results from simulated power analysis
// ====================


// Calculate a count variable

gen count = 1


// Make an indicator if powered at a certain level

local counter = 1

foreach x in `step_macro' {
    // Calculate indicator for power threshold for an observation
    gen power_10_`counter' = 0
    gen power_05_`counter' = 0
    gen power_01_`counter' = 0
    gen power_001_`counter' = 0

    replace power_10_`counter' = 1 if p_storage`counter' <= .1
    replace power_05_`counter' = 1 if p_storage`counter' <= .05
    replace power_01_`counter' = 1 if p_storage`counter' <= .01
    replace power_001_`counter' = 1 if p_storage`counter' <= .001

    // Move the counter one forward
    local counter = `counter' + 1
}


// Make an indicator if sign error at a certain level
```

```
local counter = 1
foreach x in ‘step_macro’ {
    local power_list 10 05 01 001
    foreach y in ‘power_list’ {
        gen s_error_‘y’_‘counter’ = 0
        replace s_error_‘y’_‘counter’ = 1 if power_‘y’_‘counter’ == 1 & b_storage‘counte
    }
    // Move the counter one forward
    local counter = ‘counter’ + 1
}
```

```
// generate  magnitude-error

local counter = 1
foreach x in ‘step_macro’ {
    gen m_error_‘counter’ = abs(b_storage‘counter’/(‘x’*100))

    local power_list 10 05 01 001
    foreach y in ‘power_list’ {
        gen m_error_‘y’_‘counter’ = m_error_‘counter’
        replace m_error_‘y’_‘counter’ = . if power_‘y’_‘counter’ == 0
    }
    drop m_error_‘counter’

    // Move the counter one forward
    local counter = ‘counter’ + 1
}

// Generate Beliveabilitiy
```

```
local counter = 1
foreach x in ‘step_macro’ {
    local power_list 10 05 01 001
    foreach y in ‘power_list’ {
        gen believe_‘y’_‘counter’ = 0
        replace believe_‘y’_‘counter’ = 1 if power_‘y’_‘counter’ == 1 & m_error_‘y’_‘cou
    }
    // Move the counter one forward
    local counter = ‘counter’ + 1
}


// Collapse by effect size

gcollapse (sum) count power_* s_error_* believe_* (mean) m_error_*


// Reshape data by effect size

reshape long ///
    power_10_@ power_05_@ power_01_@ power_001_@ ///
    s_error_10_@ s_error_05_@ s_error_01_@ s_error_001_@ ///
    m_error_10_@ m_error_05_@ m_error_01_@ m_error_001_@ ///
    believe_10_@ believe_05_@ believe_01_@ believe_001_@, ///
    i(count) j(effect_size)

// Remove hanging _

rename *_ *


// Turn into a percent

local power_list 10 05 01 001
foreach y in ‘power_list’ {
    replace s_error_‘y’ = (s_error_‘y’/power_‘y’)*100
```

```
    replace s_error_‘y’ = . if effect_size == 1 // Cannot have a sign-error when no trea
}


// Turn into a percent

qui ds power_* believe_*
foreach x in ‘r(varlist)’ {
    replace ‘x’ = (‘x’/count)*100
}

// Cannot have a sign-error when no treatment effect

local power_list 10 05 01 001
foreach y in ‘power_list’ {
    replace believe_‘y’ = . if effect_size == 1
}


// Fix effect size

local counter = 1


foreach x in ‘step_macro’ {
    replace effect_size = ‘x’*100 in ‘counter’

    // Move the counter one forward
    local counter = ‘counter’ + 1
}


// Drop unneeded variables

drop count
```

```
////////////////////////////////////////////////////////////////////////////
// Plot power curves
// ====================


// Create a variable that is the gap between desired power level and closest estimates

capture drop gap
gen  gap = abs(80 - power_05)

// Sort on this variable

sort gap


// Run a regression on the two closest observations

qui reg power_05 effect_size in 1/2


// Predict MDE using these two points (more accurate for finer grid)

scalar mde = (power_05-_b[_cons])/_b[effect_size]
local mde = mde

// Add label to graph with this MDE

capture drop mde_label
gen mde_label = ""
set obs `=_N+1'
replace mde_label = "MDE" in `=_N'
replace effect_size = `mde' in `=_N'

capture drop full_power
gen full_power = 102.5
```

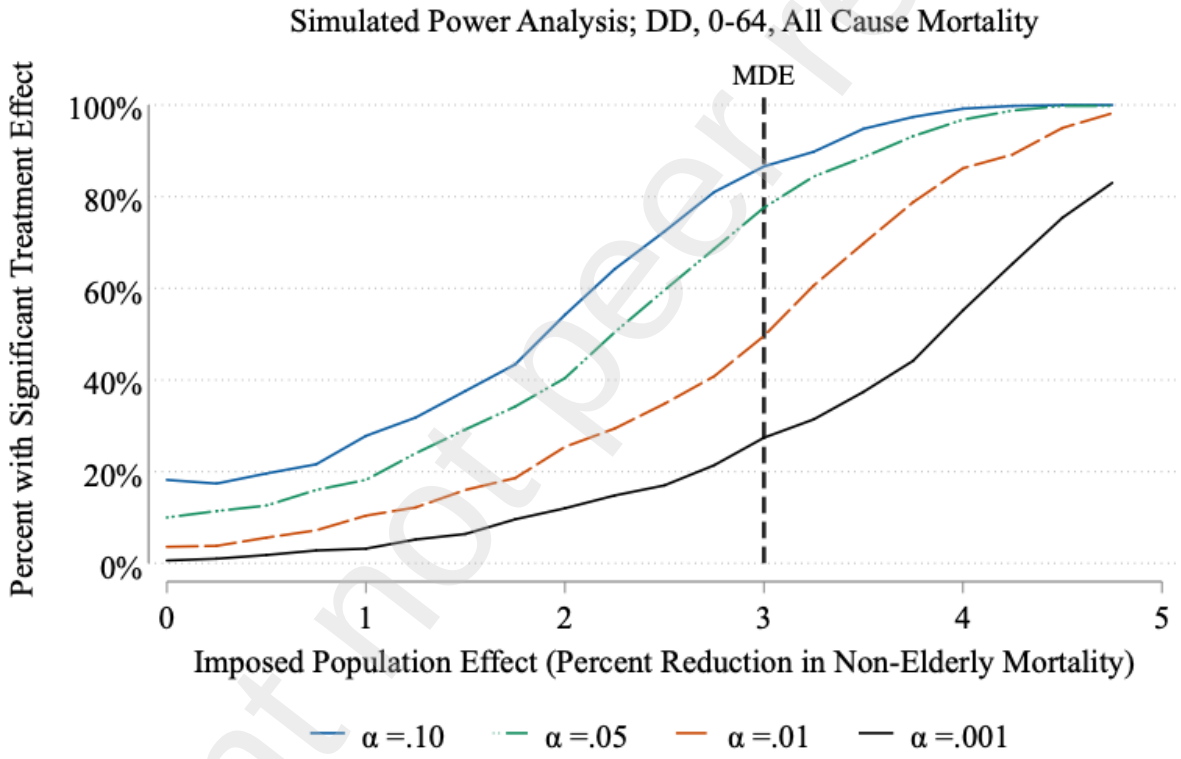
```
// Plot power curve

sort effect_size
twoway connected power_10 effect_size ,  lpattern("l") color(sea) msymbol(none) mlabcolo
    || connected  power_05 effect_size ,  lpattern("..._") color(turquoise) msymbol(none)
    || connected  power_01 effect_size , lpattern("_") color(vermillion) msymbol(none) m
    || connected  power_001  effect_size ,  lpattern("l") color(black) msymbol(none) mla
    || scatter full_power effect_size , mlabel(mde_label) msymbol(none) mlabpos(12) mlab
        xline(‘mde’, lpattern(dash) lcolor(gs3) lwidth(.5) noextend) ///
        ytitle("Percent with Significant Treatment Effect", size(4)) ///
        xtitle("Imposed Population Effect (Percent Reduction in Non-Elderly Mortality)",
        xscale(r(0 5)) ///
        xlabel(, nogrid labsize(4)) ///
        ylabel(0 "0%" 20 "20%"  40 "40%" 60 "60%" 80 "80%" 100 "100%",gmax noticks labsi
        legend(order( 1 2 3 4) pos(6) col(4) ///
            label(1 "{&alpha} =.10") label(2 "{&alpha} =.05") ///
            label(3 "{&alpha} =.01") label(4 "{&alpha} =.001") size(4)) ///
            title("Simulated Power Analysis; DD, 0-64, All Cause Mortality" " ", size(4)


    graph export "state_level_public_data_example/output/simulated_power_analysis.png",

// ![Simulated Power Analysis; DD, 0-64, All Cause Mortality](simulated_power_analysis.p
```

Figure A5: Power



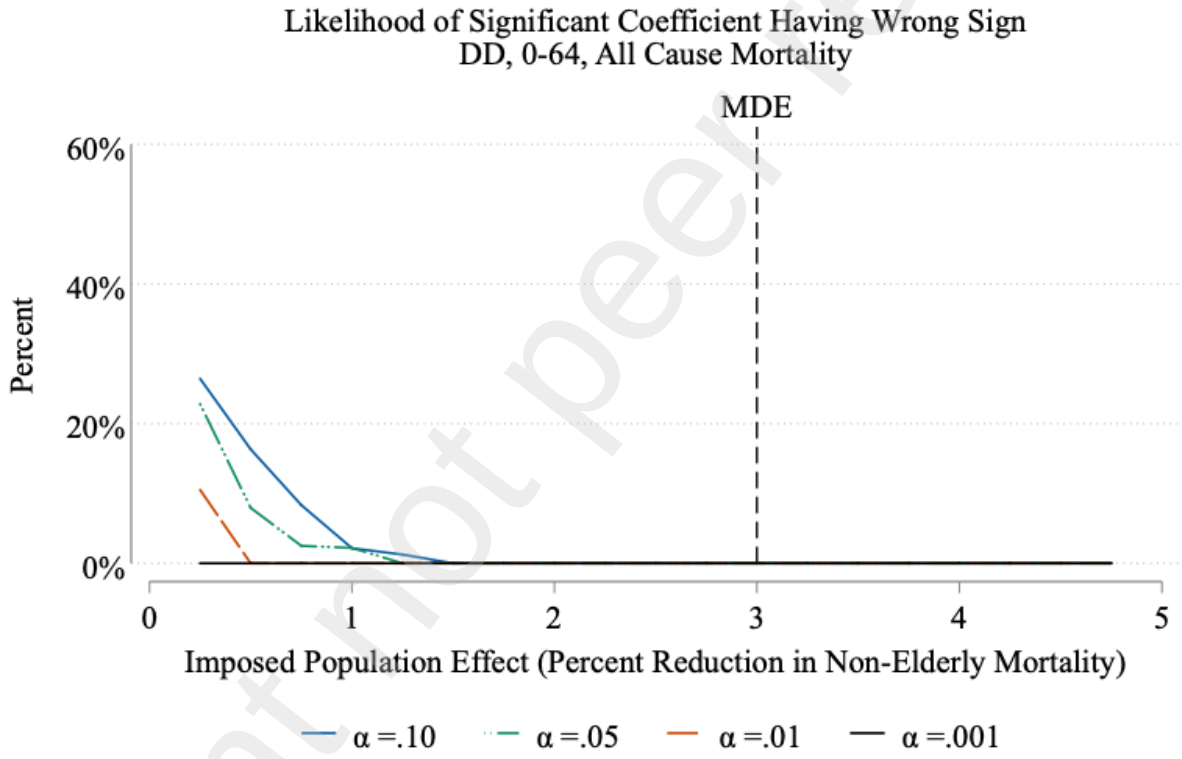
```
// Plot sign error

sum s_error_10
gen  s_error_label= 62.5
twoway connected s_error_10 effect_size ,  lpattern("l") color(sea) msymbol(none) mlabco
    || connected s_error_05 effect_size  ,  lpattern("._") color(turquoise) msymbol(non
    || connected s_error_01 effect_size  , lpattern("_") color(vermillion) msymbol(none)
    || connected s_error_001  effect_size  ,  lpattern("l") color(black) msymbol(none) m
    || scatter s_error_label effect_size  , mlabel(mde_label) msymbol(none)  mlabpos(12)
    ytitle("Percent", size(4)) ///
        xtitle("Imposed Population Effect (Percent Reduction in Non-Elderly Mortality)",
        legend(size(4) order(1 2 3 4) pos(6) col(4) label(1 "{&alpha} =.10") label(2 "{&
        xscale(r(0 5)) ///
        xline(‘mde’, lpattern(dash) lcolor(grey) noextend) ///
        xlabel( , nogrid labsize(4)) ///
        ylabel(0 "0%" 20 "20%"  40 "40%" 60 "60%",gmax noticks labsize(4)) ///
        title("Likelihood of Significant Coefficient Having Wrong Sign" "DD, 0-64, All C

    graph export "state_level_public_data_example/output/s_error.png", replace width(800

// ![Likelihood of Significant Coefficient Having Wrong Sign DD, 0-64, All Cause Mortali
```

Figure A6: Sign error

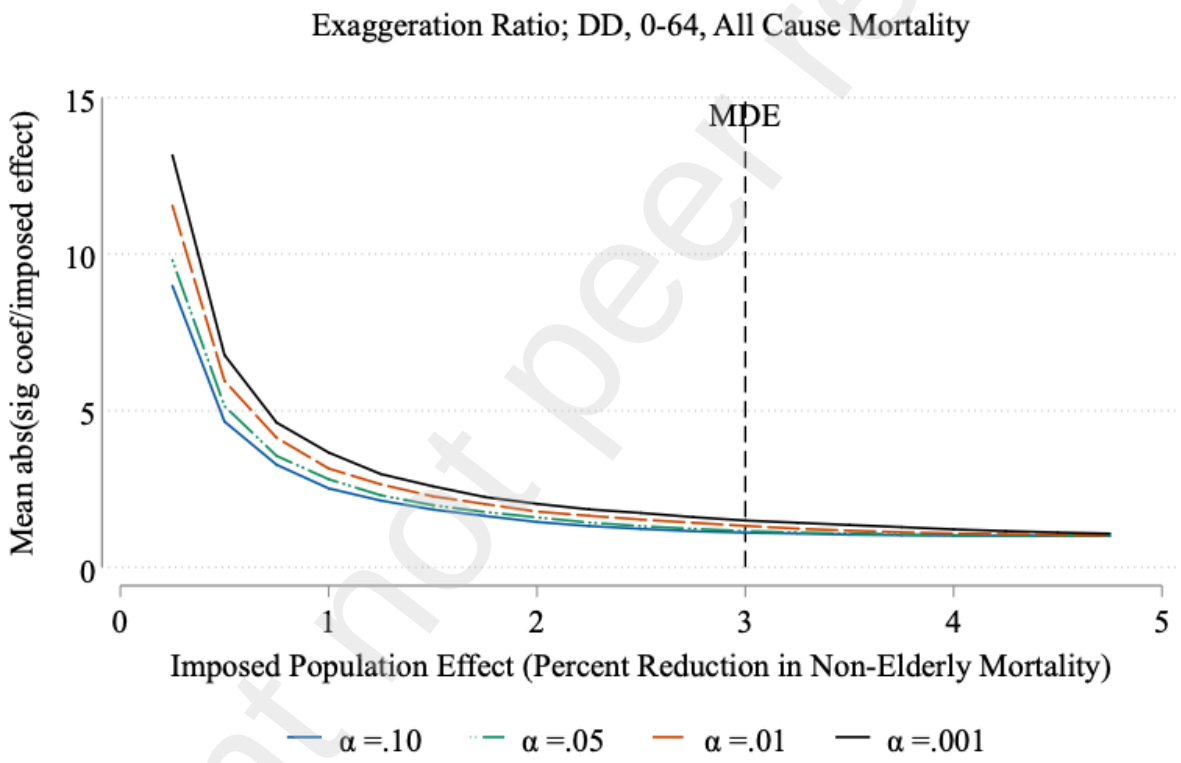


```
// Plot magnitude error

sum m_error_001
gen  height= `r(max)'*1.05
twoway connected m_error_10 effect_size ,  lpattern("l") color(sea) msymbol(none) mlabco
    || connected m_error_05 effect_size ,  lpattern("..._") color(turquoise) msymbol(none
    || connected m_error_01 effect_size , lpattern("_") color(vermillion) msymbol(none)
    || connected m_error_001  effect_size ,  lpattern("l") color(black) msymbol(none) ml
    || scatter height effect_size , mlabel(mde_label) msymbol(none)  mlabpos(12) mlabsiz
    ytitle("Mean abs(sig coef/imposed effect)", size(4)) ///
        xtitle("Imposed Population Effect (Percent Reduction in Non-Elderly Mortality)",
        legend(size(4) order(1 2 3 4) pos(6) col(4) label(1 "{&alpha} =.10") label(2 "{&
        xscale(r(0 5)) ///
        xline(`mde', lpattern(dash) lcolor(grey) noextend) ///
        xlabel(, nogrid labsize(4)) ///
        ylabel(, gmax noticks labsize(4)) ///
        title("Exaggeration Ratio; DD, 0-64, All Cause Mortality"  " ", size(4))

        graph export "state_level_public_data_example/output/m_error.png", replace width
// ![Exaggeration Ratio; DD, 0-64, All Cause Mortality](m_error.png){width="100%"}
```

Figure A7: Magnitude error



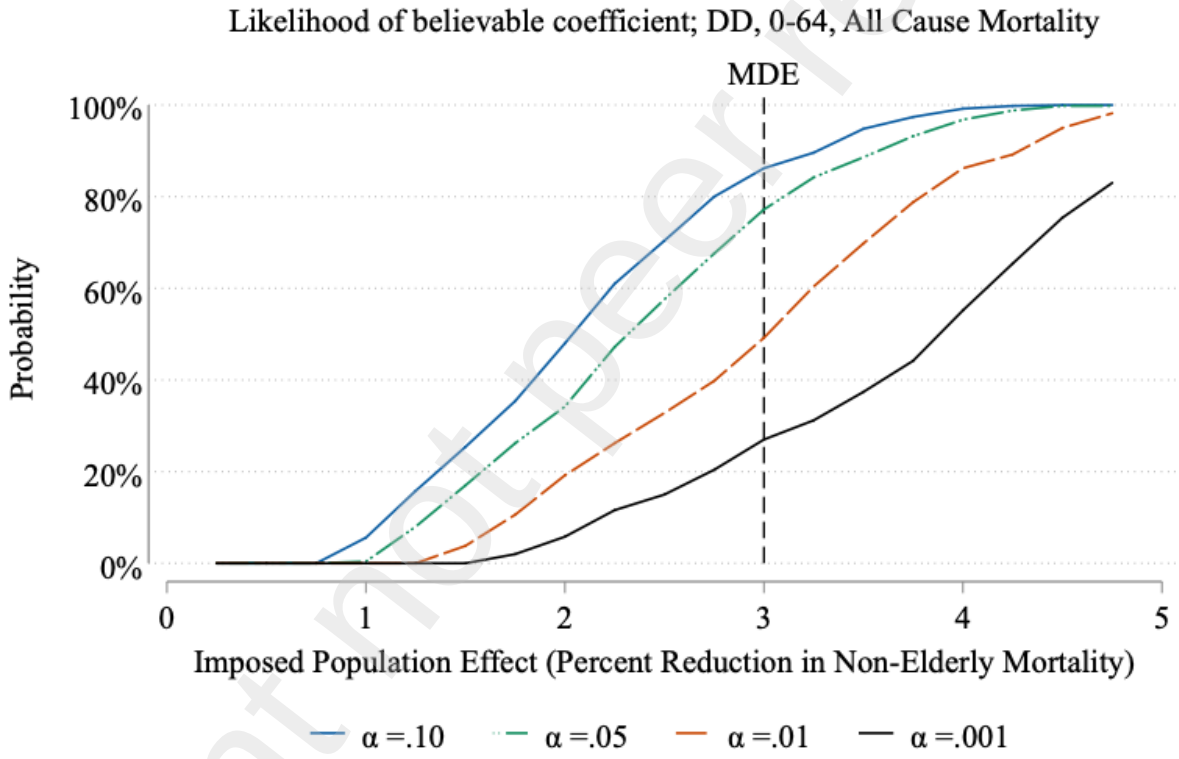
```
// Plot believability

twoway connected believe_10 effect_size ,  lpattern("l") color(sea) msymbol(none) mlabco
    || connected believe_05 effect_size ,  lpattern("._") color(turquoise) msymbol(none
    || connected believe_01 effect_size , lpattern("_") color(vermillion) msymbol(none)
    || connected believe_001  effect_size ,  lpattern("l") color(black) msymbol(none) ml
    || scatter full_power effect_size , mlabel(mde_label) msymbol(none) mlabpos(12) mlab
    xtitle("Imposed Population Effect (Percent Reduction in Non-Elderly Mortality)", siz
        legend(size(4) order(1 2 3 4) pos(6) col(4) label(1 "{&alpha} =.10") label(2 "{&
                ytitle("Probability", size(4)) ///
        xscale(r(0 5)) ///
        xline(‘mde’, lpattern(dash) lcolor(grey) noextend) ///
        xlabel(, nogrid labsize(4)) ///
        ylabel(0 "0%" 20 "20%"  40 "40%" 60 "60%" 80 "80%" 100 "100%",gmax noticks labsi
        title("Likelihood of believable coefficient; DD, 0-64, All Cause Mortality" " ",

    graph export  "state_level_public_data_example/output/believable.png", replace width


// ![Likelihood of believable coefficient; DD, 0-64, All Cause Mortality](believable.png
```

Figure A8: Believability



Using this simple example, we can see that for this simple research design the minimum mortality reduction that is believable, well-powered (80%), and significant at the 5% level is around 3%. Changing the research design (e.g. adding control variables, shifting to the county-level, changing the cause of death) would certainly impact power.

This simple research design is a DiD comparing 23 random treated states to 18 random control states. In this simple design we used 5 years of pre-expansion data and 3 years of post-expansion data. Both state and year fixed-effects were included. Regressions were weighted by state-population and standard errors were clustered at the state-level. The dependent variable was the natural log of the all-cause non-elderly mortality rate per 100,000.