COURSE OVERVIEW
This course introduces statistical techniques to evaluate health outcomes, focusing primarily on causal inference. Using observational data to establish causality — the conclusion that one thing causes another — has important applications in science, medicine, policy, economics, and business, as it allows us to measure and predict the true effectiveness of our decisions.

TOPICS

EXPERIMENTAL DESIGN
- “How Much Alcohol...” *NYT* article
- “Three or More Eggs...” *CNN* article
- “It’s Time as Academics...” *Forbes* article
- “The Evidence Supports...” *WSJ* article
- “AI Can’t Reason Why” *WSJ* article
- “Workplace Wellness...” *HA* article
- Chapter 1 in Mastering ‘Metrics

FIXED EFFECTS REGRESSIONS
- “Nike Says...” *NYT* article
- “Nike Vaporfly...” *NYT* article
- “The Costs of Low Birth Weight” *QJE* article
- Chapter 2 in Mastering ‘Metrics

MATCHING MODELS
- “Does Piped Water...” *World Bank* article
- “Too Much Ado...” *ViH* article
- “Comparison of Sales...” *JAMANO* article
- “Propensity Scores...” *SMMR* article

INSTRUMENTAL VARIABLES
- “Can Healthcare IT...” *JPE* article
- “Measuring Returns...” *JPE* article
- “Is There a Doctor...” *SIEPER* article
- Chapter 3 in Mastering ‘Metrics

REGRESSION DISCONTINUITY DESIGN
- “Regression Discontinuity Designs...” *BMJ* article
- “Subsidizing Health Insurance...” *MI* article
- “Estimating Marginal Returns...” *QJE* article
- “Does Medicare Save Lives?” *QJE* article
- “After Midnight...” *AEJ* article
- Chapter 4 in Mastering ‘Metrics
DIFFERENCES-IN-DIFFERENCES ESTIMATION
• Complete Article Review Questions
• "Regional Variation..." MI article
• "Do Large Health Insurance Subsidies..." MI article
• "Does Medicare Reimbursement..." RESrat article
• "Static and Dynamic..." QJE article
• "Hospital Discharges..." WSJ article
• Chapter 5 in Mastering ‘Metrics

CAUSAL INFERENCE
Comparisons made when “all else is equal” have a causal interpretation. That is, the change in a dependent variable $Y$ following a change in the explanatory variable $X$, holding all else equal, can be interpreted as the causal effect of $X$ on $Y$.

Potential outcomes are the outcomes associated with the possible choices individuals in the data can make. In the health insurance example, one potential outcome is the health status of someone who has health insurance. The other potential outcome is the health status that would come with not having it.

The counterfactual is what would have happened had an individual made another choice instead. For instance, if a person has health insurance, then the counterfactual would be having no health insurance. The counterfactual outcome, then, is what his or her health status would have been in the absence of health insurance. The fundamental issue for causal inference is that we do not observe the counterfactual outcome—we must infer it instead using statistical techniques.

Selection bias occurs when all else is not equal in our comparison of outcomes across individuals or groups. In the health insurance example, those with health insurance also tend to have higher incomes, so even without health insurance their health would likely have been better. The greatest challenge for making a valid causal inference using observational data is eliminating the selection bias that arises from unobserved differences across individuals in the data.

EXPERIMENTAL DESIGN
In a randomized controlled trial (RCT), researchers change the causal variable of interest (e.g., access to health insurance) for a randomly assigned treatment group and compare their outcomes (e.g., health status) to a control group that did not receive the treatment. This results in a valid causal inference because “all else is equal” due to the design of the experiment; that is, it eliminates selection bias (as long as the randomization worked properly). Checking for balance across the observable variables
of the treatment and control groups is the primary way to verify that randomization worked as intended.

Although RCTs are the gold standard for causal inference, they are not always feasible. For that reason, researchers instead use observational data and statistical methods that make all else as close to equal as possible and allow for a valid causal inference.

ARE EGGS, ALCOHOL, & NUTRASWEET ACTUALLY BAD FOR YOU?
In our first case discussion, we thought critically about the proper way to distinguish causation from correlation. The overarching theme of all three examples was that inferring causality from observational studies — even ones that include a large number of controls for important factors like income and obesity — can be fraught with challenges. For the study of alcohol consumption, for instance, the conclusion that “the safest level of drinking is none” appears unwarranted given the limitations of the data used to reach it. This can be seen most clearly in the fact that “drinking has been associated with conditions it couldn’t logically protect against: a lower risk of deafness, hip fractures, the common cold and even alcoholic liver cirrhosis.” It’s much more likely that health determines drinking rather than the other way around. This poses a severe problem for causal inference: if abstainers are predisposed towards poor health, then comparing them to those who drink will underestimate any negative effects that alcohol has — that is, there will be selection bias. The same issue also plagues the studies of eggs and artificial sweeteners. Unobservable differences between those who consume these products and those who don’t undermine any study that’s not an RCT. This was seen most clearly in the egg study’s appendix, where including additional controls brought down the effect of egg consumption on mortality and cardiovascular disease to a statistically insignificant, noisy zero.

AI CAN’T REASON WHY
This article provides a nice demonstration of why correlation does not imply causation, even in very large datasets and when using sophisticated methods like machine learning and AI. No matter how much data we have, without a formal model of underlying causal factors we “can’t tell whether a crowing rooster makes the sun rise, or the other way around.” An RCT, albeit a gruesome one, would immediately give us the answer: if we killed the rooster, the sun would still rise. Understanding such notions of cause and effect is the primary aim of this course.
**DO WORKPLACE WELLNESS PROGRAMS REALLY WORK?**

In our final case discussion of the first class, we again assessed the distinction between RCTs and observational studies, this time in the context of workplace wellness programs. Previous studies of these programs often found huge benefits, but they often suffered from fundamental flaws in their research designs. By observing voluntary, presumably motivated participants, or a “high-risk” cohort (meaning the previous period’s high utilizers), self-selection bias and regression to the mean are likely to be unavoidable. As stated in the article, “Self-improvers are likely to be drawn to self-improvement programs, and self-improvers are more likely to improve. Further, passive non-participants can be tracked all the way through the study since they cannot drop out from not participating, but dropouts from the participant group — whose results would presumably be unfavorable — are not counted and are considered lost to follow-up. So the study design is undermined by two major limitations, both of which would tend to overstate savings.” The policy implications are also worth noting here, as tens of millions of employees are subjected to these unpopular and expensive programs that have few demonstrated benefits.

**FIXED EFFECTS REGRESSIONS**

Regressions are the bedrock of statistical analysis, as they allow us to estimate the various relationships among all of the variables in our data. Most importantly, when implemented so that the key observed variables have been made equal across treatment and control groups and selection bias from the things we can’t see has been mostly eliminated, regressions allow us to make a valid causal inference — precisely what we want to accomplish with our analysis.

One way in which a regression might fail to deliver a causal estimate, however, is if we have **omitted variable bias**. Omitted variable bias occurs when we omit a variable from the regression that affects both the treatment variable, \( X \), and the outcome variable, \( Y \). Omitting this variable — often denoted \( W \) and referred to as a **confound** — means that the error term is correlated with the regressors (a technicality in which one of our assumptions for OLS has been violated, so our results will be biased). By omitting \( W \), we will mistakenly conclude that all of the impact on \( Y \) comes from \( X \), even though part of it actually came from \( W \).

For example, consider a regression of health insurance on health outcomes that omits income. If we find a positive effect of health insurance on health, it could be that health insurance causes health to improve. Or, it could be that those with higher incomes are more likely to have insurance and that those with higher incomes are more likely to have better health irrespective of their insurance status (e.g., rich people can afford to
drink a lot of kale smoothies, which may lead to better health — or at least make them feel superior to everyone else).

When we have access to data with more than one observation per entity (e.g., we track several patients over multiple years at one hospital or we have several patients across multiple hospitals for just one year), we can employ a statistical technique called **fixed effects regression** (e.g., a patient fixed effect in the former, a hospital fixed effect in the latter). In a fixed effects regression, we can account for the time-invariant portion of the omitted factors that vary across individuals, like race, gender, family background, or innate ability, which we refer to as **unobserved heterogeneity**. The key idea behind fixed effects is this: if any of the omitted variables do not change over time, then any changes in Y that occur over time could not have been caused by these stable omitted variables. By using a fixed effect in our regression, we have successfully controlled for any constant unobserved variables (even though we don’t have data on them!) and therefore get closer to a causal estimate. If any of these unobserved variables change over time, however, a fixed effects regression will still suffer from selection bias.

**ARE NIKE VAPORFLY 4% REALLY 4% FASTER?**

In our first case, we discussed the various methods used by the *New York Times* to analyze the effectiveness of a groundbreaking new running shoe, the Nike Vaporfly 4%. In its marketing for the Vaporfly, Nike claims that it improves performance by up to 4% — a substantial effect that many experts viewed skeptically. To vet Nike’s assertions, the *NYT* cited both its own statistical analyses of observational data and an RCT conducted on a handful of elite runners.

This RCT is worth mentioning at the outset, because it can serve as a benchmark when we consider the other types of studies. In the experiment, researchers from the University of Colorado invited 10 fast, male runners to their lab, fitted them with motion-capture sensors, and filmed them as they wore the 4% shoe, a different Nike marathon shoe, and a similar Adidas model. The study measured gains of about 4%, as claimed by Nike, and was also able to document the science underlying these results. But even though an RCT is the gold standard for causal inference, we should remain circumspect about its external validity in this case: the gains experienced by elite male runners may not carry over to more pedestrian runners or even to elite female runners — unfortunately, gender bias in clinical trials remains a pervasive phenomenon.1 Despite some limitations, one key advantage of using observational data is that they often allow

---

1 See, for example, [https://www.nytimes.com/2019/05/30/health/gender-stereotypes-research.html](https://www.nytimes.com/2019/05/30/health/gender-stereotypes-research.html)
for a much larger sample and more nuanced findings, as they can include thousands of runners of all genders, races, ages, and other important characteristics, much more than could ever be included in a narrow RCT.

Moving on to the observational studies, the *NYT* first conducted a standard regression analysis using self-reported data from runners on Strava, a social network for endurance athletes. These regressions are a good first pass for testing whether the Vaporly is associated with faster running times, as they can control for many variables that might also affect performance, such as race conditions, weather, gender, age, past race times, and training mileage. Like any regression based on observational data, however, selection bias may prevent us from uncovering causal effects. For instance, motivation could be a key omitted variable in this case, as motivated runners may be more likely to purchase special $250 shoes (the treatment) and also more likely to have a breakthrough running performance (the outcome). Because Strava relies on user-reported data, the estimates from the first study may also suffer from measurement error if runners deliberately manipulate their entries in ways related to buying the Vaporfly (e.g., they may want to impress other users both by purchasing fancy shoes and by lying about having fast race times).

Extending this initial approach, the *NYT* also “matched” runners who completed the same pair of races based on their past performance and other observable characteristics, like gender and age. By matching runners in this way, the researchers aim to emulate an RCT by finding a treatment runner (the one who switches to Vaporflys in the second race) with a control runner (the one who doesn’t switch shoes). The hope is that this research design will make all else equal across the two runners so that the only difference between them is their shoes, but it cannot eliminate selection bias entirely: just as in the previous example, it is still possible — even likely — that a motivated runner selectively chooses to wear special shoes when he or she feels primed for a personal best.

Next, the *NYT* considers a runner fixed effect regression. Here, the analysis holds fixed any time-invariant unobservables within a runner, such as stride length or genetics, and identifies the effect of Vaporflys by comparing how a runner’s times change when he or she switches to the shoes. For this to be a credible research design, we would need to be confident that none of the unobserved time-varying factors are related to the runner’s shoe choice. As before, it seems likely that motivation could be such an unobserved time-varying confound: we don’t have data for a runners’ level of motivation, and they might switch to the expensive Vaporfly’s just for those races where they feel especially motivated to run fast.
Finally, the NYT estimated the likelihood of a runner achieving a personal best while wearing the Vaporflys. This type of analysis can be viewed as a complement to the previous fixed effects regressions, although in this case the outcome is binary (1 for personal best, 0 for not personal best). If that is the outcome runners care most about, then these regressions will be particularly useful. Nevertheless, they, too, suffer from the same limitations related to unobserved confounds.

Although each method mentioned in the article comes with its own strengths and weaknesses, that all four found runners improve their times by about 4% should reassure us that the effects are real, and this type of thorough robustness check represents the best practice for any rigorous scientific analysis.

THE COST OF LOW BIRTH WEIGHT
In our final case discussion of the second class, we went through Almond, Chay, & Lee’s 2005 study “The Cost of Low Birth Weight.” Because birth weight has emerged as a leading indicator of infant health, many previous studies have attempted to estimate the benefits of interventions that increase the weight of newborns. These studies often used cross-sectional data, however, and may be biased by omitted variables that influence both birth weight and health outcomes and that cannot be influenced by policy, such as genetic factors. To address this limitation, the Almond et al. study compares the hospital costs, health at birth, and infant mortality rates between heavier and lighter infants from twin pairs. Specifically, they include a mother fixed effect in their regressions to capture any common unobserved factors between the twins (e.g., the mother’s smoking habits, prenatal stress levels, and so on), which then allows the authors to isolate how different birth weights across the twins with the same mother affect health and spending outcomes. Their analysis implies substantially smaller effects of low birth weight than previously thought, suggesting that existing estimates overstate the true costs and consequences of low birth weight by at least a factor of four. As such, policies aimed at increasing birth weights will be unlikely to result in substantial benefits.

MATCHING MODELS
Recall that in a randomized controlled trial, the treatment and control groups are chosen randomly so that any pre-treatment differences between the two groups are solely due to chance. This design allows us to estimate the causal effect of treatment because treatment and outcomes are not confounded by some unobserved variable. In observational studies, such confounding is likely to occur because participants self-select into the treatment or control group, and this decision may be related to unobserved factors correlated with their potential outcomes. Matching models help
correct for this by pairing individuals based on their observable characteristics, with the aim that selection bias from the things we cannot see will be mostly eliminated using this procedure.

For instance, if we compare the unadjusted death rates across different types of smokers, we will find that pipe and cigar smokers have a higher death rate than cigarette smokers. All else is not equal in this comparison, however, because cigar and pipe smokers are much older, on average, than cigarette smokers, so they would be expected to have a higher death rate anyway. When we match pipe smokers to cigarette smokers based on age, on the other hand, we find that cigarette smokers have a much higher death rate — the differences in age were biasing our results.

If we have multi-dimensional characteristics that explain treatment, we will want to use a propensity score to compare individuals who, based on their observable characteristics, had similar probabilities of being placed into the treatment group even though they differ in regards to their actual treatment. If two individuals have the same probability of being treated, conditional on X, then we say they have the same propensity score. And, if two individuals have the same propensity score but one is in the treatment group and the other is not, then any differences between their observed outcomes must be due to the treatment. The key assumption here is that treatment, conditional on the propensity score, is independent of potential outcomes (i.e., is "as good as random"); if it’s not, the results will be biased. We calculate the propensity score using either a probit or logit regression, and there are several different ways to implement these models, such as nearest neighbor or caliper. In all cases, we only want to include observations that share common support — that is, regions that have a sufficient number of treatment and control observations. This highlights a key advantage of matching models over OLS regressions: if the treated and control group have insufficient overlap in their multivariate covariate distributions, the groups may not be comparable to one another. OLS regressions deal with this issue through extrapolation, which may not be appropriate under certain conditions.

MEASURING THE COSTS OF ASTHMA TREATMENT

This study provides a nice overview of different matching models within the context of health expenditures for asthma patients. Those suffering from asthma have higher healthcare expenditures overall, but part of this difference could be due to the fact that they are sicker in general, not necessarily because asthma is particularly costly. In Table 2, we can see immediately why a simple comparison of means between the treatment and control groups would not represent a true causal effect of asthma treatment on healthcare expenditures: age and CCI have statistically significant differences across the
two groups, and there are some regional differences as well. Because the treatment group is younger, on average, we might suspect that their healthcare expenses would be biased downwards (younger people typically have fewer healthcare expenses). At the same time, however, the treatment group is generally sicker, as evidenced by their higher measure of CCI, so it is not clear that the difference in means is caused by the treatment, as it could be due to their comorbid conditions. In addition, regional differences in healthcare access and spending could be correlated in important, unobservable ways.

To account for these differences across the treatment and control group, the paper compares a series of matching models in which a treatment patient is matched to a control subject(s) using various techniques. For instance, because CCI is a strong predictor of having asthma, a matching model compares individuals with similar CCI measures but different exposure to asthma. The extent of this confounding bias can be seen in the unmatched difference of means of $7,053, which falls to $2040-$4463 in the matching models. Although the results from the matching models get us closer to a true causal effect, they may still suffer from omitted variable bias — a matching model is not a silver bullet in that regard.

DOES BREASTFEEDING IMPROVE INFANTS’ NEURODEVELOPMENT?

Breastfeeding has long been believed to confer benefits to infants. As mentioned in the study, however, randomizing breastfeeding is not feasible or ethical, and confounding by parental education and socio-economic status (SES) is highly likely in the absence of randomization. For that reason, any differences in IQ we observe across those who were and were not breastfed may be due to factors other than breastfeeding itself. For example, breast feeding is easier for mothers who work in an office where frequent breaks and a space for pumping are available, and these types of jobs are more common for wealthier and more highly educated mothers. The article also mentions unobserved gestational age as another example of a potential unobserved confounder that could impact the expected treatment outcome, which would break the strongly ignorable treatment assignment (SITA) assumption that there are no unobserved confounders. The article concludes with a comparison of various matching models, where the estimated IQ benefits from breastfeeding are smaller than the unadjusted mean would suggest.

COMPARISON OF SALES INCOME AND R&D COSTS FOR CANCER DRUGS

One justification for high drug prices is that their financial return needs to be large enough to finance the R&D of new drugs. An observational study of 99 cancer drugs
approved by the FDA from 1989 to 2017 suggests the returns are indeed sufficient to provide such incentives: the median income by the end of 2017 was found to be $14.50 for every $1 spent on R&D. This measure of ROI may be inflated by a fundamental flaw in the study’s design, however, as the sample does not include a proper accounting of failed drug development. This can be seen most prominently in a sensitivity analysis that included 50 failed drugs and brought down the effective ROI by nearly half. To measure the true ROI for drug development, a potential research strategy would be to collect a sample of failed drugs and drugs that have reached the market and then apply a propensity score matching method based on their observable characteristics. Without a true control group of failed drug launches, this study suffers from survivorship bias, a form of selection bias in which focusing exclusively on those who made it past some selection process and overlooking those who did not, typically because of their lack of visibility, leads to biased findings.

**INSTRUMENTAL VARIABLES**

In our previous classes, we discussed how to measure the causal effect of a treatment, T, by comparing the outcomes, Y, of individuals who received the treatment (T = 1) to those who received a placebo or no treatment at all (T = 0). In regression form, we have

\[ Y_i = \alpha X_i + \beta T_i + \varepsilon_i \]

where \( \beta \) measures the causal effect of treatment if the treatment was randomly assigned, as in an RCT. If individuals self-select treatment, however, our results may suffer from selection bias and not reflect a causal effect. For instance, if more-motivated individuals are more likely to sign up for a wellness program, then any improvement in health among those who participate in the program may be due to this unobserved factor rather than the program itself. In addition, the treatment may be endogenous, meaning that individuals were assigned to a treatment group based on some specific characteristics related to their potential outcomes, such as the likelihood that a particular drug will be effective for them compared to others. Technically speaking, our results will be biased because unobserved characteristics in the error term will contain variables that are also correlated with the treatment, T, so that \( \text{cov}(T, \varepsilon) \neq 0 \), which violates one of the key assumptions of OLS for obtaining unbiased estimates.

An instrumental variable (IV), Z, breaks this link between unobserved confounds and treatment under the following conditions:

1. Z is correlated with T: \( \text{cov}(Z, T) \neq 0 \)
2. $Z$ is uncorrelated with $\varepsilon$: $\text{cov}(Z, \varepsilon) = 0$

These conditions stipulate that the instrument must be relevant, in the sense that it affects selection into treatment, and also exogenous, in the sense that it is not correlated with the unobserved factors that affect outcomes (we refer to this as the exclusion restriction). As an example, consider a drug trial that assigns patients to the treatment or control group based on a coin flip. The coin flip is relevant, because it determines which group a participant is placed into, so directly affects treatment. And the coin flip is exogenous, because whether it lands on heads or tails is independent of any patient characteristics, both observed and unobserved. Therefore, the coin flip is a valid instrument.

We can use a valid IV to estimate the causal effect of treatment in a two-stage least squares regression, where in the first stage we regress $T$ on $Z$ and predict $\hat{T}$, and in the second stage we regress $Y$ on $\hat{T}$, which then gives us the causal effect if treatment because $\hat{T}$ is uncorrelated with the unobserved factors that influence $Y$. Although implementing this technique is straightforward using software, the real challenge is finding a valid instrument that satisfies our two conditions (some professors’ entire careers have been made just by finding a compelling instrument!).

**IS THERE A DOCTOR IN THE HOUSE?**

In our first case discussion, we considered a study of informal health advice. Having a health-care provider in your family may yield important benefits: a simple reminder to take your medicine, for example, may lead to greater adherence, or getting a second opinion from a trusted loved one might precipitate a life-saving — but unpleasant — procedure. The statistics bear this out, as those with relatives in a health profession are 10 percent more likely to live beyond age 80 and are also significantly less likely to have chronic lifestyle-related conditions, such as heart attacks and diabetes. Determining whether having a doctor in your household actually causes these dramatic health benefits, however, is complicated by the fact that our family members are not randomly assigned to us. That is to say, families with medical professionals in them might be different from other families in ways that are also correlated with better health outcomes, such as differences in education, income, or self-discipline.

To measure the causal effect of having a familial medical source on health outcomes, the researchers used data from Sweden, where lotteries were used in the early 2000s to break ties among equally qualified applicants for admission into medical schools. They then compared the health of the family members of lottery winners against lottery losers — a setup similar to a randomized control trial. The med school lottery is a great
example of a valid IV. It is clearly relevant, as those who win the lottery are more likely to go to medical school, and it is also exogenous, because the lottery is the single causal channel through which the family member would be more likely to attend medical school and ultimately improve the rest of the family’s health (i.e., it satisfies the exclusion restriction). On this last point, the exclusion restriction, the authors also test whether income and education drive family health outcomes, in which case the exclusion restriction would be violated because these other determinants, not informal medical advice, drive the difference. To do this, they compared outcomes for a profession with similar income and education covariates, lawyers, and find their conclusion stands: the parents of doctors were 16 percent more likely to be alive than the parents of lawyers 20 years after their children matriculated.

**MEASURING RETURNS TO HOSPITAL CARE USING AMBULANCE REFERRAL PATTERNS**

Next, we discussed research that studied whether hospitals that provide more care and accrue higher Medicare reimbursements actually achieve better health outcomes — or whether the additional spending at high-cost hospitals is mostly wasted. A key challenge when estimating performance differences across hospitals is patient selection. Patients choose or are referred to hospitals on the basis of the hospital’s capabilities, so the highest-quality hospital in an area may treat the sickest patients. Alternatively, better-educated or higher-income patients may be in better health and also more likely to choose what is perceived to be a higher-quality hospital. For that reason, efforts to provide report cards for hospitals are often criticized for their inability to fully control for the underlying differences in patients that may bias these ratings.

In light of such challenges, this paper develops an empirical framework that allows the authors to compare hospital performance using plausibly exogenous variation in hospital assignment. The key ingredient of their approach is that the locus of treatment for emergency hospitalizations is, to a large extent, determined byprehospital factors: ambulance transport decisions and a patient’s location. Because ambulance companies are pseudo-randomly assigned to patients in an emergency, the authors can develop convincing measures of the impact hospitals have on patient outcomes.

The authors use two complementary identification strategies to exploit variation in ambulance transports. The first uses the fact that in areas served by multiple ambulance companies, the company dispatched to the patient is effectively random because of rotational assignments. Moreover, the authors demonstrate that ambulance companies serving the same small geographic area have preferences for which hospital they take patients. These facts suggest that the ambulance company dispatched to emergency patients may serve as a random assignment mechanism across local hospitals — that
is, ambulance referral patterns are a valid IV that allows us to recover a causal estimate of how hospital spending affects health outcomes.

Their second strategy considers contiguous areas on opposite sides of ambulance service area boundaries in the state of New York. In New York, each state-certified EMS provider is assigned to a territory through a certificate of need process in which it is allowed to be first due for response; other areas may be entered when that area's local provider is busy. This allows the authors to compare those living on either side of an ambulance service area boundary. To the extent that these neighbors are similar to one another, the boundary can generate exogenous variation in the hospitals to which patients are transported, a second type of instrument.

In the IV analysis, the key underlying assumption of the two approaches is that the sources of variation across the different types of hospitals have been purged of patient-specific factors that affect costs or outcomes. To assess whether this is true, at least along observable dimensions, the authors show the balance of patient characteristics across those whose ambulances tend to transport patients to relatively high-spending or low-spending hospitals. All of their measures appear to be well balanced.

To implement this estimation approach, the authors then set up a standard two-stage least squares regression. In the first stage, they estimate the relationship between average hospital spending, $H$, and the instrument, $Z$, which are the hospital costs associated with the ambulance assigned to the patient. A similar setup applies to the boundary analysis, and in both cases the instrument is clearly relevant.

In the second stage, the authors estimate that a 10 percent increase in spending is associated with a 2.4 percentage point lower mortality rate, or about 6 percent of baseline mortality. Unlike the OLS results, these 2SLS results are more robust to the inclusion of controls; the estimates fall by only 20 percent from the first to the last column and are statistically indistinguishable, albeit partly because of larger 2SLS standard errors. To put the estimate in context, this result implies that a one standard deviation increase in average hospital spending is associated with a 3.7 percentage point reduction in mortality, or 10 percent of the sample mortality rate. Thus, these results provide compelling evidence that higher-spending hospitals have significantly lower patient mortality, at least for emergency admissions.

For the border analysis, the authors first show that patients are more likely to attend a hospital located within their area. Specifically, for patients with a hospital located in their area, 72 percent of patients living within 1 mile of an ambulance service area border are treated at a hospital within this same area, and this rate increases to 75 percent and
78 percent for those living within 2 and 5 miles of a border, respectively. This is borne out in the first-stage relationship reported in table 7. As with the national Medicare results, the New York State OLS results are sensitive to controls; indeed, adding controls moves the coefficients from an insignificant positive estimate to a significant negative estimate. These findings indicate substantial selection bias in the OLS regression.

Using the border strategy, a 10 percent increase in costs is associated with a 0.005 percentage point reduction in mortality, or about 2 percent of the baseline mortality rate. This result implies that a one standard deviation increase in average hospital costs is associated with a 2.1 percentage point reduction in mortality, or 8.9 percent of the baseline mortality rate. These results reinforce the findings from the ambulance company preference strategy, as once again we see that higher-cost hospitals have lower mortality for emergency admissions.

**CAN HEALTHCARE IT SAVE BABIES?**

In our final discussion, we considered whether the transition to electronic medical records (EMR) caused health outcomes to improve. Many believe that the digitization of health information will lead to better health-care quality and reduce overall costs compared to paper records, with one study finding that replacing paper obstetric records with electronic ones reduced the incidence of missing charts from 16% to 2%. Within this context, more than 18,000 babies die each year in the United States during their first 28 days of life, and part of this may be due to poor monitoring technologies.

To determine whether health-care IT causes neonatal mortality to decline, the authors use a twelve-year panel data set containing the birth and death records for all U.S. counties combined with data on the adoption of EMRs by U.S. hospitals. With their data, the authors first estimate regressions that include multiple controls for hospital and county characteristics, as well as county and year fixed effects, to measure the impact of EMR adoption, finding a negative association between the proportion of hospitals that have adopted EMRs and neonatal death rates.

Despite the rich set of controls and fixed effects in their main regressions, the estimated effect may not be causal if there are unobserved and confounding changes in county or hospital characteristics over time that are correlated with EMR adoption. For example, if a hospital decides to start specializing in high-risk cases, it may invest in more technology and also experience worse health outcomes. Or, if patients become wealthier, hospitals may adopt EMRs due to better finances, and at the same time experience improved health outcomes because they treat patients who exercise more
and eat less fast food, which would lead the researchers to over-estimate the effects of healthcare IT on health outcomes.

To overcome this potential identification challenge, the authors estimate an IV specification that exploits variation in health privacy laws across states and over time. These regulations restrict the ability of hospitals to exchange and use electronic patient information and consequently reduce the attractiveness of EMRs. That is, exogenous variation in privacy laws affects the likelihood that a hospital will adopt EMRs (the IV is relevant, because the privacy laws make EMRs less useful and, therefore, hospitals less likely to adopt them), which breaks the link between the hospital’s endogenous decision to adopt EMRs and any confounding unobservable characteristics (the IV satisfies the exclusion restriction, because privacy laws should have no direct effect on health outcomes other than through the use of EMRs).

From the first stage of the IV estimates, we see that the largest single predictor of EMR adoption is the presence of a rule limiting the disclosure of patient information, which makes sense because this is a type of privacy safeguard that intentionally tries to prevent the easy flow of patient information that the digitization of health records is meant to promote. Given the strong negative association between privacy regulations and EMR adoption, the instrument is clearly relevant.

In the second stage, we see that EMR adoption does have a causal impact on infant mortality. The larger estimates in the IV model may be the result of an upward bias in the OLS model, where hospitals adopt EMRs as they specialize in more-complicated cases and consequently experience higher mortality rates. It may also be the result of heterogeneous effects of EMR adoption on health outcomes. Specifically, the larger estimates may be caused by the local average treatment effect of EMR adoption being largest for hospitals whose adoption is most influenced by the privacy regime in place. In short, the IV estimate provides strong support for the main OLS finding that EMR adoption improves neonatal health outcomes, with a rough cost-effectiveness calculation suggesting that health-care IT is associated with a cost of $531,000 per infant saved.

REGRESSION DISCONTINUITY DESIGNS

Although randomized controlled trials allow us to measure the causal effects of a given treatment or policy, many important research questions cannot be studied using such experimental methods. To get around this limitation, a regression discontinuity design (RDD) uses observational data from a setting in which a clinical or policy rule differentially assigns individuals to a treatment if they fall above or below an arbitrary
cut-off, and a control group otherwise. Because individuals were as good as randomly assigned to either side of the cut-off, omitted variables should be balanced across the treatment and control groups, allowing us to estimate an unbiased, causal effect of treatment. As we discussed in class, such situations are ubiquitous throughout health care, like the age 65 threshold for Medicare eligibility and the 1500 gram threshold for being classified as having very low birthweight.

Implementing an RDD is straightforward. In a **sharp RD**, the probability of receiving treatment jumps deterministically from 0 to 1 at the cut-off, with those on one side of the cut-off receiving treatment and those on the other side not receiving it. Given this setup, we can measure the treatment effect by comparing outcomes for those just above the cut-off to those just below it. We do this by regressing outcomes on a dummy variable for being above (or below) the threshold along with the **running variable** — the variable on which the threshold is based — for observations within a narrow bandwidth of the cut-off. It’s best to report results for several different bandwidths to show that the estimates are robust, with the fundamental tradeoff being that a narrow bandwidth ensures that unobserved confounds are balanced across the treatment and control group, whereas one that’s too narrow might result in imprecise estimates (a narrow bandwidth means you have to throw out many of your observations).

In contrast to a sharp RD, a **fuzzy RD** shifts the probability of treatment at the cut-off even though it’s not deterministic. That is, some individuals on either side of the cut-off receive treatment, although the probability of treatment increases at the cut-off. This setup is very similar to an IV regression, with similar strengths and weaknesses.

RDDs do come with several limitations. First, they may have limited external validity, as those within a narrow window of the threshold may be very different from those farther away — and hence the treatment effect for them may be very different as well. Second, manipulation around the treatment cut-off may invalidate this approach, so it is important to always test, either visually or statistically, whether it appears as if this has occurred, which can often be seen in irregular heaps on either side of the cut-off.

**SUBSIDIZING HEALTH INSURANCE FOR LOW-INCOME ADULTS**

In our first case, we discussed a study that considers how much low-income individuals are willing to pay for health insurance on Massachusetts’ health insurance exchange, called CommCare. The authors’ analysis exploits the discontinuous drops in CommCare’s subsidies that correspond to higher incomes, with the subsidy amount changing discretely at 150%, 200%, and 250% of the federal poverty line. These sharp drops provide the identifying variation required to estimate the causal effect of
premiums on coverage choices through an RDD. At each threshold where premiums increase, the authors find significant reductions in enrollment (they also provide evidence that individuals do not manipulate their income around these thresholds), so enrollment declines reflect individuals’ decisions not to buy insurance given the higher price.

Part of the reason that enrollment remains relatively low despite the large subsidies is adverse selection: the enrollees who drop coverage are healthier (that is, lower cost) than the average enrollee, so do not find it worthwhile to take up coverage. Adverse selection is not the entire story, however, as the insurers’ costs of covering marginal consumers are three to four times higher than these individuals’ willingness to pay. Because low-income individuals pay only 20-30% of their total medical expenditures (the remaining balance is either provided as charitable care or left unpaid), the uncompensated care can explain the low willingness to pay and low enrollment even at highly subsidized prices. Enrollees’ willingness to pay is much closer to their own costs net of the uncompensated care that they would have received while uninsured, meaning that the subsidies benefit not only the recipient directly but also have spillover benefits to the providers of uncompensated care. The primary beneficiary of health insurance expansions, therefore, may be the providers of uncompensated care, as opposed to the previously uninsured.

ESTIMATING MARGINAL RETURNS TO MEDICAL CARE

Next, we discussed research that studied whether additional medical care actually provides any additional benefits. The paper was motivated by the limitations of previous studies that have used cross-sectional data to compare “high-spending” and “low-spending” geographic areas, which tend to find large differences in spending yet remarkably similar health outcomes. Omitted variables may have biased such studies, however, so the authors consider an empirical strategy that can address such unobserved confounds and deliver a causal estimate for the returns to marginal care.

To do so, the authors focus on newborns with very low birthweight. Due to arbitrary treatment guidelines, infants weighing less than 1500 grams are eligible to receive more-intensive care, so the authors can use an RDD for their estimation. It will be a valid approach because this threshold is unlikely to represent a break in underlying health risk: the cutoff of 1500 grams reflects medical convention rather than medical criteria, so those infants just above 1500 grams should be essentially identical to those just below the cut-off. One concern is that physicians might manipulate reported birthweights to provide additional care to infants whom they feel would benefit from it, though the lack of heaping at the threshold suggests that physicians and hospitals do
not do this. The authors also compare the main covariates of interest in the 5 ounces around the threshold to show that they are stable on either side, which verifies that the setting satisfies this requirement for using an RDD.

The authors’ main regression shows a 22% reduction in mortality compared to a mean mortality rate of 5.53% in the three ounces above the threshold (the untreated group in the RDD) — a substantial effect from extra care! The OLS estimate in the second column mimics the local linear regression but now places equal weight on the observations up to three ounces on either side of the threshold. The point estimate is slightly smaller here but still large. Importantly, the authors test different bandwidths and find that their results are robust.

The authors then report the mean hospital charges in one-ounce bins. The measure appears fairly flat at $94,000 for the three ounces prior to the threshold, then falls discontinuously to $85,000 after the threshold and continues on a downward trend, consistent with heavier newborns having fewer complications. This is likely to be directly related to an infant’s length of stay, as Figure IIIB shows that average length of stay drops from just over 27.3 days immediately prior to the threshold to 25.7 days immediately after it.

Following this analysis, the authors examine several possible treatment mechanisms at the discontinuity. They find some weak evidence of differences for operations on the heart and diagnostic ultrasounds, for which they estimate an approximately 10% increase in usage just prior to the threshold, though these differences are often not statistically significant. All in all, the differences in these summary measures are consistent with medical care driving the decline in mortality, but the study lacks the statistical power to detect differences in any particular procedures. Finally, the authors estimate that the cost per newborn life saved is $527,083, which is well within conventional benchmarks for cost-effectiveness.

DOES MEDICARE SAVE LIVES?
Continuing with our case discussions, we considered whether Medicare improves patients’ health. The authors use an intuitive estimation strategy to answer this question: because the elderly can only enroll in Medicare after turning 65, this eligibility rule creates an opportunity to employ a regression discontinuity design to compare the outcomes of those on Medicare to those just younger than 65 and not yet eligible for the program. Associated with the rise in Medicare coverage at age 65 is an increase of about 9 percentage points in the fraction of people with any coverage, leaving only
about 3 percent of the population over 65 uninsured, compared with about 13% of those under 65.

The study focuses on unplanned admissions through the emergency room for non-deferrable conditions – those with similar weekend and weekday admission rates – so the crucial RD assumptions hold in this case (i.e., the admissions are not manipulable). The authors also do some other checks to ensure their study design is valid, finding no discernible evidence of a drop in the severity of co-morbidities at age 65 or any other noticeable differences in characteristics on either side of the cut-off.

Given this setup, the authors’ main findings are striking. First, all-cause admissions jump 7% once people reach 65 and emergency room admissions rise by 3%. Second, they find a small increase in the number of procedures performed and total list charges, as well as a large increase in the chance of being transferred to skilled nursing facilities. Third, they find a reduction in the 28-day readmission rate, a 1 percentage point lower death rate (or roughly 20 percent), and a similar absolute reduction in mortality at 28 days and 90 days, with this pattern persisting for at least two years after admission. The authors argue that this pattern is consistent with an insurance generosity channel, reflecting increased services (or the more-timely delivery of services) for patients who are covered by Medicare and supplemental insurance relative to the typical insurance package held by people just under 65. So, once again, a very nice application of how an RDD can uncover the causal effects of a policy.

**AFTER MIDNIGHT**

Finally, we discussed a paper that tests for moral hazard in the care of newborns by comparing treatment and health outcomes across essentially identical newborns who happen to differ in their insurance coverage. A main challenge when testing for moral hazard is that insurance coverage and treatment levels are chosen in response to the underlying health of the patients (i.e., adverse selection). The authors address this endogeneity problem by exploiting the rules of hospital reimbursement schedules: hospitals are reimbursed based on the number of days a patient is in the hospital, with days counted as the number of midnights in care. That is, a newborn delivered at 12:01 am will have nearly a full day in care before being counted as present in the hospital, whereas a newborn delivered at 11:59 pm will be counted as present only sixty seconds after deliver, so those born just after midnight are covered for an additional night of care compared to those born just prior to midnight.

The authors find that the discontinuity in insurance coverage associated with the minute of birth generates a substantial difference in the average stay length, despite nearly
identical observable characteristics for the mothers and infants. Infants born shortly after midnight spend an additional 0.25 nights in the hospital, on average, suggesting that this will be a valid RDD. The authors do not find any corresponding health benefits to go along with these longer stays, however, suggesting that this extra care is mostly wasteful and represents a clear case or moral hazard. But we should be somewhat circumspect in reaching this conclusion: the authors do not have data on many important measures, such as long-term outcomes or maternal stress and well-being, that may justify the longer hospital stays following the birth of a child.

DIFFERENCES-IN-DIFFERENCES ESTIMATION

For our final empirical method, we discussed differences-in-differences (DiD) estimations. As with our other methods throughout the term, DiD emulates an RCT by comparing the differential effect of a treatment on a treatment group versus a control group. That is, by comparing the average change in outcomes over time for the treatment group compared to the average change over time for the control group, we can estimate the causal effect of treatment under the assumption that the counterfactual trend for the treatment group would have been the same as the control group’s had it not been for the treatment, which we call the parallel trends assumption.

Set up as a regression, a DiD controls for observable differences across the treatment and control groups, as well as a time-invariant unobserved factors in the form of group-specific fixed effects. Common time trends are accounted for with a time fixed effect. Structured this way, we can take the difference of two differences to estimate the treatment effect. The first difference compares the outcomes for the treatment group before and after the treatment, which will include both a treatment effect and a time effect. The second difference compares the outcomes for the control group before and after the treatment group receives the treatment; this difference just includes a time effect. Taking the difference between the first difference and second difference then gives us the treatment effect: (Treatment Effect + Time Effect) – (Time Effect) = Treatment Effect.

A DiD estimation may still suffer from selection bias, however, depending on how the treatment and control groups are chosen. For that reasons, care must be taken to verify the timing and assignment of treatment were as good as random — verifying that the outcomes for both the treatment and control groups evolved similarly in the pre-treatment period is an important check on the parallel trends assumption here. The treatment may also be correlated with other changes occurring at the same time, so conducting a placebo test of other outcomes not affected by treatment should provide some reassurance that the DiD isn’t picking up a spurious correlation.
REGIONAL VARIATION IN U.S. HEALTHCARE USE

In our first case, we discussed why the use of health care by Medicare beneficiaries varies so widely across the U.S. — the average enrollee in Miami costs $14,423, for example, but costs just $7,819 in Minneapolis. This variation has puzzled policy makers for some time, as higher area-level use is not generally correlated with better patient outcomes. Because it remains unclear whether providers or patients (or both) drive these differences, these puzzled policy makers can’t even be certain about which policies they should target. Variation stemming from the provider side would suggest implementing different reimbursement policies than variation that comes from patients, who may require a different set of incentives to change their behavior.

To address this question, the authors have the clever idea to use the migration of patients to separate variation due to the demand-side factors of patients’ characteristics, such as their health or health-care preferences, from variation due to place-specific variables or supply-side factors, such as doctors’ incentives and beliefs, endowments of physical or human capital, and hospital market structure. To see the intuition for this approach, consider a patient who moves from high-use Miami to low-use Minneapolis. If all of the difference in health-care use between these cities arises from supply-side differences, we would expect the migrant’s use to drop immediately following the move, to a level similar to other patients of the low-use doctors in Minneapolis. But if all of the difference reflects the demand-side, in the sense that residents of Miami are sicker, we would expect the migrant’s use to remain constant after the move, at a level similar to the typical person in Miami. Where the observed change falls between these two extremes identifies the relative importance of demand- and supply-side factors.

With this empirical strategy, the authors find that nearly 50% of the spending differences across geographical areas stems from the characteristics of patients, meaning both their basic health and their varying preferences for the intensiveness of care. The rest of the spending differences stem from place-specific factors, potentially due to disparities in provider practices and incentives. They also find that demand-side factors matter more for outcomes such as emergency room visits, where 71% is attributable to patients, who are likely to make more of the decisions about whether or not to seek care in those situations. In the other direction, demand-side factors matter less for outcomes like diagnostic and imaging tests, where the physician is the main decision-maker: patients account for only 9% of the regional variation in diagnostic tests and 14% of the variation in imaging tests. In these cases, the variation by geographical region may be due to
differing provider practices, with healthcare institutions in some places consistently spending more money on testing than other providers do elsewhere.

**DO LARGER HEALTH INSURANCE SUBSIDIES FAVOR PATIENTS OR PRODUCERS?**

Next, we discussed research that studied the effect of health insurance subsidies for private Medicare plans, which typically offer a narrower network in exchange for lower cost sharing compared to traditional Medicare. These private plans receive a capitation subsidy to take on Medicare beneficiaries and may charge a supplemental premium, with about 31% of Medicare beneficiaries using private plans that account for over $200 billion in spending. In 2000, Medicare instituted the Benefits Improvement and Protection Act (BIPA), which created subsidy floors in urban and rural counties that effectively increased payments to many Medicare Advantage plans. The authors use this payment change to study how much of the larger subsidies private plans pass through to their policy holders, and how much they keep for themselves as profits.

Proponents of larger subsidies argue that they result in lower premiums or more generous benefits for Medicare beneficiaries. Opponents, on the other hand, argue that they lead to large profits for insurance companies and healthcare providers, with few gains for patients. Economic theory suggests that either of these two possibilities could occur. Extensive market power, for instance, would result in very little of the subsidies being passed through, as plans with few competitors have little incentive to lower their premiums to attract more policy holders. Moreover, if a plan that lowers its premiums attracts a more costly group of newly enrolled consumers, then it must increase its premiums to break even, which would result in a less than complete pass-through even in the absence of market power.

In their analysis, the authors examine the incidence of subsidies to private Medicare Advantage plans by studying the sharp change in these payments brought about by BIPA. Prior to BIPA, payments were largely determined by historical traditional Medicare per capita expenditures in the county. BIPA reformed these payments by instituting a system of rural and urban payment floors that raised payments in many counties, with the floor in urban counties set at $525 per member per month, and $475 in rural counties. About 72% of counties had payments below the floor, so plans operating in these areas faced an immediate payment increase equal to the difference between their prior capitation level and the new floor.

The authors show that Medicare Advantage capitation payments in the counties with binding floors were following the same trend as places unaffected by the floors in the period before the payment reform, but they increased by an average of about $50 per
month (12%) when BIPA was implemented. This provides the variation necessary for a DiD estimation of capitation payments and suggests the parallel trends assumption holds.

The authors find a pass through of about 30-45%. In addition to lowering premiums, plans may have responded to the increased payments by raising the generosity of their coverage (benefits pass-through). They find some evidence that this also occurred: the increase in payments led to a decline in average personal physician and specialist co-payments.

**DOES MEDICARE REIMBURSEMENT DRIVE UP DRUG LAUNCH PRICES?**

Continuing with our case discussions, we considered whether Medicare’s reimbursement policies lead to higher prices for new drugs. Before 2005, Medicare reimbursed providers based on the average wholesale price of the drug, which may not reflect an actual sales price. After 2005, Medicare began reimbursing based on the past price of the drug, or average sales price (ASP). But because a new drug doesn’t have a past price on which to base its average price, their reimbursement begins as a markup of the wholesale acquisition cost.

Medicare’s choice to base reimbursement on past prices could motivate manufacturers to set higher launch prices for two reasons. First, providers are less sensitive to the price when reimbursement covers a portion of it (i.e., moral hazard). Second, firms may set a high launch price, higher than what would be the short-run profit-maximizing price, in order to secure higher reimbursement for the provider in the future. That is, by setting a higher launch price, the future reimbursement is higher, so future demand from providers is also higher and the firm can charge a higher price in the future. In response to this, one Medicare administrator claimed that the new pricing policy “creates a perverse incentive for manufacturers to set higher prices,” which is what this paper seeks to study.

To answer this question, the authors compare drug launch prices before and after the reimbursement change. This is a compelling empirical strategy for a few reasons. First, they can use launch-price variation within a molecule for drugs launching different dosage forms of the same molecule in different years. Second, they can use launch-price variation across molecules for drugs in the same class. Third, they can compare launch prices for drugs reimbursed under different payment schemes. Given this set up, the authors can estimate a DiD in which the control group will be the drugs for which Medicare reimbursement did not change to a markup on ASP.
As the basis for this estimation, figure 3 shows that average launch prices are much higher after the price change. Nevertheless, other factors could be causing these price increases, so the authors control for novelty, class, and year in their regression analysis. In the DiD estimation, the authors compare drug launch prices before and after the 2005 policy change. The treatment group consists of liquid drugs administered by outpatient providers, for which Medicare and private insurers began reimbursing based on ASP in 2005, while the control group varies depending on the specification. In one, the control group is prescription drugs with 70% or more sales in the retail market. In another, it’s prescription drugs administered by providers in an outpatient setting but not reimbursed based on ASP. These drugs include vaccines, blood products, and infusion drugs furnished through a covered item of durable medical equipment.

As with the previous regressions, the data show a clear pattern of a rise in prices for the treated group relative to the control group. For both control groups, the results are economically and statistically significant: coefficients vary from 1.7 to 1.8 for new molecules, indicating that launch prices are 450% to 500% higher after the log transformation. We can also see that the effect of the policy is strongest for new molecules, where manufacturers have more freedom to launch at higher prices.

As a whole, this paper provides very compelling evidence that Medicare's reimbursement change in 2005 is partly responsible for the steep rise in drugs prices over the past decade, although the analysis does have some limitations. First, the authors use list prices rather than prices net of discounts and rebates. Second, they focus only on drug launch prices, ignoring changes in the prices of existing drugs. Finally, in the difference-in-differences analysis, the control groups might be flawed. Although one control group is outpatient provider drugs exempt from ASP-based reimbursement, the other control group is retail drugs, and the retail drug environment changed considerably during the sample period with the creation of Medicare Part D, which might have driven up the prices of some drugs sold at retail due to a moral hazard problem. Alternatively, Medicare Part D might have driven down prices if insurers used their buying power to drive down prices.

STATIC & DYNAMIC EFFECTS OF HEALTH POLICY
Finally, we discussed a paper that compares the static and dynamic welfare effects of a public health policy. Public policies designed for the static purpose of increasing utilization of an existing technology may also affect incentives to develop new technologies. In the medical sector, such technological progress has been a key contributor to the dramatic health improvements over the past fifty years, but these improvement may come at a considerable cost.
This paper provides the first empirical evidence that health policies aimed at affecting health care utilization also have large effects on technological change. It examines the introduction of three different public health policies designed to increase vaccination rates against six specific diseases. These policies effectively increased the expected return to developing new vaccines against the respective diseases.

The author uses DiD to estimate the change in the investments in vaccines affected by the policies compared to selected diseases that were not affected by the policies, which controls for underlying secular trends in vaccine industry R&D. The author uses three different policy changes to measure these effects. The first two policies increased the economic returns to vaccine development by enlarging the expected market for the vaccine: the 1991 CDC recommendation that all infants be vaccinated against Hepatitis B and the 1993 decision for Medicare to cover the cost of influenza vaccination for Medicare recipients without any copayments or deductibles. A third policy increased incentives for vaccine development by reducing expected liability costs. At the time of their enactment, the industry expected these policies to substantially increase the return to developing vaccines against the affected diseases.

The author finds that the number of new clinical trials per year increased for vaccines affected by one of the policies. This increase occurs with about a one-year lag after the policy’s introduction, and persists through the latest years of available data. The one-year lag is consistent with industry opinion that it would take six to eighteen months to start a new clinical trial.

A central limitation of this simple time series analysis, however, is that R&D was increasing more generally throughout the entire time period. To distinguish any potential effect of the policies from the increase in new clinical trials that would have occurred anyway, the author compares changes in the number of new vaccine clinical trials for affected diseases after the policy goes into effect with changes in the number of new vaccine clinical trials for diseases that were not affected by the policies. To do so, she uses a DiD specification. The choice of appropriate control diseases for this estimation is an important and difficult one, where the author needs to control for factors that affect the trend in vaccine development, particularly exogenous changes in technology.

The author’s results indicate that, across all specifications, the policies are associated with a statistically significant increase of 1.2 to 1.3 new vaccine clinical trials per year for each affected disease. Between 1983 and 1986, each affected disease had on average 0.5 new clinical trials per year, so the results suggest that the policies are associated
with about 2 and a half times more new vaccine clinical trials per year per affected disease. Moreover, the OLS estimates imply that the economic incentives embodied in these three policies alone account for almost one-third of the 260 total new vaccine clinical trials for all diseases during the entire seventeen-year period. The Hepatitis B and flu policies alone were associated with an annual increase in expected market revenue of $518 million and $98 million, respectively.

REFERENCES


