Journal of Empirical Legal Studies
Volume 4, Issue 1, 69–102, March 2007

The Impact of Damage Caps on Malpractice Claims: Randomization Inference with Difference-in-Differences

*John J. Donohue III and Daniel E. Ho**

We use differences-in-differences (DID) to assess the impact of damage caps on medical malpractice claims for states adopting caps between 1991–2004. We find that conventional DID estimators exhibit acute model sensitivity. As a solution, we offer (nonparametric) covariance-adjusted randomization inference, which incorporates information about cap adoption more directly and reduces model sensitivity. We find no evidence that caps affect the number of malpractice claims against physicians.

I. Introduction

Compared to the average OECD country, the United States has fewer per-capita doctors, nurses, and hospital beds, but ends up spending vastly more for medical services. In 2002, U.S. health expenditures per capita were $1,821–5,267 more than the next highest spending country (Switzerland) and $3,074 more than the median OECD country (Anderson et al. 2005). Unfortunately, it is not clear that substantially greater U.S. spending is buying better experiences with the health-care system (with the exception of

*John J. Donohue III, Leighton Homer Surbeck Professor of Law, Yale Law School, P.O. Box 208215, New Haven, CT 06520; email: j.donohue@yale.edu. Daniel E. Ho, Assistant Professor of Law, Stanford Law School, 559 Nathan Abbott Way, Stanford, CA 94305; email: dho@law.stanford.edu.

We thank Sascha Becker, Chris Griffin, and Tatiana Neumann for invaluable research assistance. We are indebted to John L. Adams for thoughtful discussion of a previous version of this article, and thank Angela Early, Ted Eisenberg, Eric Helland, Nick Parrillo, Stephen F. Williams, two anonymous reviewers, and participants at the 2006 RAND Institute for Civil Justice/*Journal of Empirical Legal Studies* conference on medical malpractice for helpful comments.

shorter waits for nonurgent surgery) or better overall health outcomes (within the OECD, the United States is in the bottom quartile of population health indicators, such as life expectancy and infant mortality (Hussey et al. 2004)).

The combination of high expense and (seemingly) low quality results has encouraged the search for the presence of substantial avoidable burdens on the heath-care system. An effective lobbying campaign has led many Americans to identify the U.S. tort system as a major component of this problem, and the resulting legislative response in many states has been to cap damages on medical malpractice awards as a means of reducing these perceived unproductive costs. But the simple equation of high malpractice costs and high overall health spending is misleading. While it is true that the direct expenses of malpractice litigation and insurance are higher in the United States than in other countries, these direct expenses are only a small part of overall health spending and thus do not explain the higher American per-capita health costs. Specifically, the total costs to health providers of U.S. malpractice claims—including awards and legal and insurance costs—has been estimated at $6.5 billion in 2001, which was only 0.46 percent of total U.S. health spending (Anderson et al. 2005).[1] Even if medical malpractice caps can lower these overall legal and insurance expenditures, they cannot play a very substantial role in explaining the substantially greater costs of the U.S. health-care system.

A series of papers by Kessler and McClellan (1996, 2002), however, makes the case that the pernicious impact of the tort system goes beyond the relatively small direct expenditures related to medical malpractice. Instead, they argue, efforts by health-care professionals to avoid liability by engaging in defensive medicine impose large costs on the overall U.S. health system. Specifically, Kessler and McClellan (1996) find that the costs of treating Medicare patients hospitalized for acute myocardial infarction or ischemic heart disease was lower (with no sacrifice in outcomes) in states with certain types of tort reform. Extrapolating from these results, one study has found that health-care costs could be reduced by 5–9 percent if national malpractice tort reform were adopted following California's scheme of imposing a $250,000 limit on noneconomic damages in malpractice cases (U.S. Department of Health and Human Services 2003). In other words, according to this

[1] In contrast, the comparable medical malpractice expenses in Canada come to only 0.27 percent of total Canadian heath spending.

HHS study, the lower-bound estimate of the cost of defensive medicine that could be avoided by a fairly stringent damage cap is an order of magnitude greater than the direct costs of medical malpractice litigation.[2]

Of course, we must be careful not to forget that the issue of medical malpractice involves not only the costs of awards, insurance, tort litigation, and any defensive medicine that the liability system encourages, but also the costs of the malpractice itself, which is high. Indeed, it has been estimated that perhaps 100,000 or more individuals die from medical error every year, and many more are seriously injured. The costs of these injuries are difficult to estimate but the 5–9 percent figure that was offered above for the potentially avoidable percentage of the overall U.S. health bill resulting from defensive medicine is roughly in the right ball park and, if anything, may be too low. Thus, even if the costs of malpractice litigation and defensive medicine are reduced by damage caps, one must also consider whether the price for this reduction is more costly malpractice. To gain some purchase on some of these competing considerations, we look at medical malpractice claims using a national database.

II. The Theoretical Ambiguity of Malpractice Caps

Whether medical malpractice caps have any appreciable impact on the number of claims brought against physicians presents an unsolved puzzle. Theory fails to provide strong predictions because legislatively-imposed damage caps may generate conflicting effects on physicians and patients. Curtailing the damages that can be paid by doctors deemed to have committed malpractice (arguably) decreases the level of care taken by physicians, thereby potentially *increasing* medical malpractice claims. At the same time, the presence of such damage caps simultaneously decreases the value of potential claims for an injured patient, thereby *decreasing* the chance that medical malpractice suits are ever brought.

[2] The Kessler and McClellan and HHS studies have not gone unchallenged. The Congressional Budget Office (CBO) did not find similar results of ostensibly defensive medicine for illnesses beyond the two that Kessler and McClellan employed (Beider & Hagen 2004). Dubay et al. (1999) found that tort reform had a much smaller effect on the percentage of births by cesarean section. In general, the CBO was less convinced by the claimed benefits of tort reform and concluded that "savings from defensive medicine would be small" in the wake of tort reform.

This theoretical ambiguity has not yet been resolved by the development of a consistent story in the empirical research. Although careful scrutiny of hospital records has provided solid accounts of the underlying incidence of malpractice (Brennan et al. 1991), such studies have not clarified the impact of actual tort reforms because of minimal variation in liability regimes. Hospital records studies often measure "malpractice pressure" by the number of malpractice claims (e.g., Tussing & Wojtoqycz 1992; Baldwin et al. 1995; Localio et al. 1991; Dhankhar et al. 2005), but without evidence of the impact of tort reform on the number of malpractice claims brought, such correlations provide little information about the direct policies of interest.

The direct study of the full array of "tort reforms," including damage caps, has been challenging because of lack of systematic data across jurisdictions and relatively small variation in liability regimes. Some detailed studies have been possible in the few states, such as Florida and Texas, that have mandated disclosure of all physician reports. For example, Vidmar et al. (2005) study 31,521 Florida malpractice claims closed between 1990 and 2003 and find no difference in the per-capita (per-doctor) claim frequency between the 1990–1993 and the 2000–2003 periods. Studying Texas from 1988–2002, Black et al. (2005) also find that after adjusting for population growth, the number of closed claims were stable between 1990 and 2002. However, when adjusting for the numbers of doctors or the growth in real health-care spending, the number of paid claims and the number of large paid claims actually declined. For example, paid claims declined from 6.4 per 100 doctors per year in 1990–1992 to 4.6 per 100 doctors per year in 2000–2002.

Yoon (2001) exploits the introduction of damage caps in Alabama in 1987 and the subsequent Alabama Supreme Court decision nullifying these caps to perform a difference-in-difference analysis for the years 1987 to 1999 with Arkansas, Mississippi, and Tennessee as the control group. Yoon finds that the average payment for a medical malpractice claim in Alabama was roughly $20,000 lower when the damage caps were in effect.

A number of studies have examined statewide variation over time (e.g., Sharkey 2005; Viscusi & Born 1995; Shepherd & Rubin 2005), with some finding little effect of damage caps and others finding dramatic effects, either on insurance company profits or on lives saved. For example, using three years of data (1992, 1996, and 2001), Sharkey (2005) creates a data set of 557 jury cases in which the plaintiff received compensatory damages. Using regression analysis with an indicator for noneconomic damage caps, Sharkey finds that such caps have no statistically significant effect on the size of overall compensatory damages, as reflected in either jury verdicts or final judgments.

Viscusi and Born (1995) examine the effects of medical malpractice liability reforms in the period 1985–1987 on the overall insurance market. In particular, the study uses annual firm-level data per state for the years 1984–1991 to examine the impact of damage caps on the ratio of losses to premiums (loss ratio). Using an autoregressive OLS model with ln(loss ratio) as the dependent variable, Viscusi and Born find that damage caps decreased the loss ratio by an average of 13–25 percent.

Shepherd and Rubin (2005) reach the provocative conclusion "that caps on noneconomic damages, caps on punitive damages, a higher evidence standard for punitive damages, product liability reform, and prejudgment interest reform lead to fewer accidental deaths, while reforms to the collateral source rule lead to increased deaths. Overall, the tort reforms in the states between 1981–2000 have led to an estimated 14,222 fewer accidental deaths."

Kessler and McClellan (1996), using difference-in-differences on state-year data, find that damage caps reduced expenditures without substantial effects on medical outcomes for Medicare beneficiaries treated for heart disease. Estimating the impact on litigation, Kessler and McClellan (2002) find that tort reforms reduce the probability of a physician facing a malpractice claim by about 1.4 percent. Conversely, Klick and Stratmann (2003) find that malpractice reforms lowered physician levels of care as measured by infant mortality. Studying earlier tort reform efforts from the mid-1970s, Zuckerman et al. (1990) find no impact of caps on the frequency of litigation.

In this article, we investigate whether there is any evidence that caps affect the number of malpractice claims using a new data set of states adopting noneconomic or punitive damage caps from 1991–2004. In particular, we focus on seven "treatment" states that adopt damage caps and 11 "control" states without any caps during our observation period. As a preliminary cut, we use a standard difference-in-differences (DID) estimation strategy that capitalizes on changes between preadoption and postadoption claim rates in the seven adopting states compared to time trends in the 11 control states. Because we seek to explain the number of filed claims, we implement this DID estimation using an overdispersed Poisson (count data) model.

Our findings, however, provide a stark reminder of the fragility of conventional DID approaches. Model sensitivity is acute due to systematic preexisting differences between treatment and control states. Introducing an arbitrary transformation of covariates (e.g., a polynomial term) using the same sample, we can *reverse* a statistically significant effect of caps. Even worse, as a clear diagnostic of the sensitivity of DID, when we *randomly* assign damage caps across states, we observe rejection rates of the null hypothesis of no effect of 50 percent across a range of specifications at a 0.1 level.

To remedy this, we nest a DID estimator within (nonparametric) randomization inference, demonstrating that model sensitivity is drastically reduced once we incorporate more information about the treatment assignment mechanism. Substantively, we conclude that there is yet no evidence that caps had any effect on claims. Methodologically, we conclude that extensions of randomization inference bear much promise as "reasoned basis of inference" for policy evaluation (Imbens & Rosenbaum 2005; Rosenbaum 2002b; Fisher 1935).

III. Data

In this section, we first discuss measurement of cap adoption and concordant sample selection. To account for unobserved time-invariant heterogeneity between states, we focus on states adopting caps from 1991–2004. We then discuss our source for nationwide malpractice claims and covariate selection.

A. *Treatment and Sample Selection*

A number of articles describe the date of adoption and general nature of state malpractice damage caps (e.g., Weiss et al. 2003), but we found extant sources to be somewhat unreliable. To remedy this, we independently surveyed state malpractice statutes to collect information about whether and when states adopted three types of caps over the period 1991–2004: (1) economic damage caps, which limit the amount of money that plaintiffs can recover as economic compensation primarily for lost wages and medical expenses; (2) noneconomic damages caps, which limit the amount of money that plaintiffs can recover for pain and suffering; and (3) punitive damage caps, which limit the amount that a jury may award to deter and punish egregious conduct.

Over our sample period, no states adopted economic damage caps, but several adopted punitive and/or noneconomic caps. Table 1 summarizes our findings concerning the adoption of caps from 1991–2004. Michigan,

Table 1: Summary of Treatment Classification from 1991–2004

Noneconomic Damage Caps

State	*Adopted (Effective)*	*Amount*
MI	1993 (Apr. 1, 1994)	280,000 (500,000 for severe exceptions)
MS	2003 (Jan. 1)	500,000
MT	1995 (Oct. 1)	250,000
OK	2003 (July 1)	300,000

Punitive Damage Caps

State	*Adopted (Effective)*	*Amount*
AR	2003 (Mar. 25)	Greater of 250,000 or 3 × compensatory damages that are limited to 1,000,000
NJ	1995 (Oct. 27)	Greater of 350,000 or 5 × compensatory damages
NC	1996 (Jan. 1)	Greater of 250,000 or 3 × compensatory damages
OK	1995 (Aug. 25)	Greater of 100,000 or actual damages for recklessness; 500,000 or 2 × actual damages for intentional tort

NOTE: We exclude states that had any caps prior to observation period. Michigan is coded as adopting a cap in 1993 despite a preexisting cap that was not mandatory for all torts. Note that Oklahoma adopts both noneconomic and punitive damage caps in the observation period.

SOURCES: Mich. Comp. Laws Ann. § 600.1483; Miss. Code Ann. § 11-1-60; Mont. Code Ann. § 25-9-411; 63 Okl. St. Ann. § 1-1708.1F; Ar. Code Ann. § 16-55-208; N.J. Stat. Ann. § 2A:15-514; N.C. Gen. Stat. Ann. § 1D-25; 23 Okl. St. Ann. § 9.1.

Mississippi, Montana, and Oklahoma adopted noneconomic caps, limiting recovery from $250,000 to $500,000 each.[3] Arkansas, New Jersey, North Carolina, and Oklahoma adopted punitive damage caps, ranging from $100,000 to some multiple of compensatory damages. We exclude from our treatment (and control) groups three states that experimented with caps for a small number of years, primarily because of the difficultly of coding exactly

[3] Michigan actually enacted a form of a noneconomic cap in 1983. Ten years later, however, Michigan amended the preexisting cap to eliminate a large number of exceptions, applicable, inter alia, to death, intentional tort, injury to the reproductive system, and loss of a vital bodily function. See Mich. Comp. Laws Ann. § 600.1483. The 1993 statute thereby, for the first time, set a manadatory cap of $250,000, with a $500,000 cap applicable to total permanent functional loss of a limb caused by injury to the brain or spinal cord, permanent impairment to cognitive capacity, and permanent loss to a reproductive organ. Consistent with Zeiler (2004) and McCullough et al. (http://www.mcandl.com/michigan.html), we thereby code Michigan as enacting a cap in 1993.

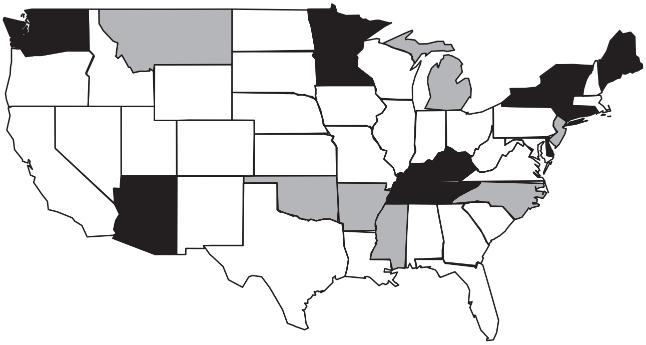
when the treatment was effective.[4] Table 1 also demonstrates that most caps did not take effect until at least several months after the passage of the formal legislation. As a result, we code state-years as "treated" beginning in the year subsequent to adoption. For example, while Michigan adopted a (mandatory) cap in 1993, the act was not to be effective until April 1, 1994; we hence code Michigan's cap as being effective in 1994. The reasonableness of this coding depends in part on when we expect caps to affect physician and patient behavior. We also investigated sensitivity to excluding the year of adoption or effectiveness, and lagging treatment an additional year, with no substantive changes in our results.

Since noneconomic and punitive caps serve, at least in part, different functions (compensatory or deterrent), we might expect that they have differing impacts on claims rates. Yet we found no substantive differences whether we pooled across both types of damage caps or conducted separate analyses for the two types of caps. Consequently, we report pooled results only, thereby defining the treatment to be the enactment of any cap. Of course, model sensitivity may itself arise from the decision of whether to pool or not. One way to address these pooling assumptions is to fit a multilevel model (see generally Gelman & Hill 2007; Raudenbush & Bryk 2002), a worthwhile approach, but one we do not explore further here.

For control states, we found 11 states that had no caps from 1991–2004 (Arizona, Connecticut, Delaware, Kentucky, Maine, Minesota, New York, Rhode Island, Tennesseee, Vermont, and Washington). Figure 1 plots our treatment and control states in grey and black, respectively. The map intuitively suggests a key characteristic that will later become evident in our data set: systematic differences exist between treatment and control states. A large proportion of control states are from the northeast, with the exception of New Jersey, which adopted a punitive damage cap in 1995. Moreover, a large proportion of states (in white) already had enacted damage caps prior to our observation period. The map also underscores that we limit our investigation to in-sample inferences, namely, the effect on the seven states adopting caps from 1991–2004. Estimating the effect of damage caps of all U.S. states may lead to substantial extrapolation—for example, we observe no large states (such as California, Florida, or Pennsylvania) adopting caps from 1991–

[4] Ohio adopted a noneconomic damage cap from 1997–1998, and a punitive damage cap from 1996–1998; Oregon adopted a noneconomic damage cap from 1991–1998; Illinois adopted a noneconomic damage cap from 1995–1997.

Figure 1: Map of states in sample.



NOTE: Grey indicates states adopting caps from 1991–2004; black indicates states with no caps in place; and white indicates states with caps existing before observation period.

2004, thereby potentially making out-of-sample inference unreliable (King & Zeng 2006).

B. Outcomes

Our outcome information comes from the National Practitioner Databank (NPDB), which came into existence with the Health Care Quality Improvement Act of 1986, 42 U.S.C. § 11101, mandating all payments of malpractice claims to be reported to the NPDB. As originally intended, the NPDB data bank would be available to state licensing boards, hospitals, and other professional organizations to enable monitoring of physicians in a system of "peer review." By querying the NPDB, health-care entities would be able to prevent incompetent physicians from moving across jurisdictions without full disclosure of prior adverse medical practice. One of the consequences has been the most complete statewide data set on medical malpractice compiled to date, making it ideal for our study. By 2005, the NPDB contained more than 370,000 records of claims.

We use the public version of the NPDB, containing all disclosable reports from September 1, 1990, to March 31, 2005. Since the treatment of interest (the enactment of a damage cap) is assigned at the state-year level, we aggregate the NPDB at the state-year level to create a data set of counts of claims for each state-year (Moulton 1990).[5] We exclude (1) claims filed in

[5] Moulton has emphasized a key insight often lost in the literature: the unit of assignment drives the power of causal inference. For example, even though NPDB contains claim-level data, the

1990 and 2005 from our analysis, as these years present incomplete state-year observations, and (2) claims not directly resulting in malpractice payments (e.g., licensure, membership, and drug enforcement actions). We aggregate using the physician's work state, and the home state when work state is not reported.[6] We exclude claims for physicians for whom no working or home state information is observed.

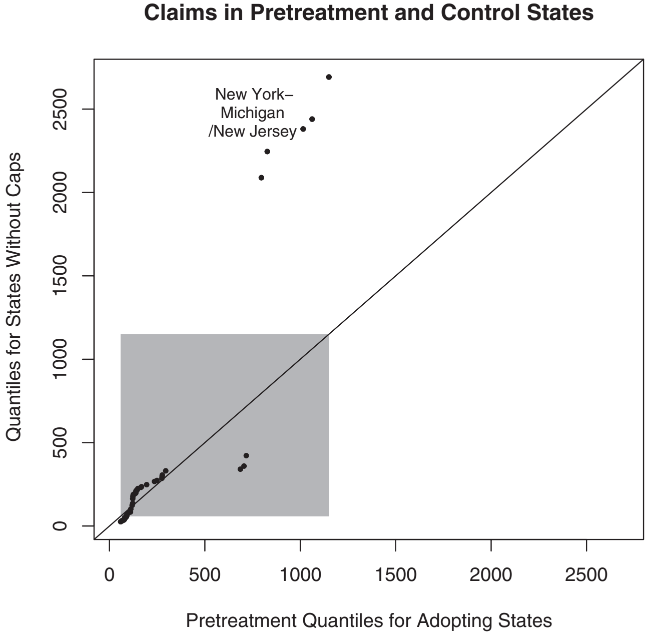
The NPDB is designed to provide comprehensive coverage, but it also has limitations that create challenges for our effort to identify the applicable liability regime governing particular claims. Specifically, for roughly 27 percent of NPDB observations, the year of the act or omission is missing or contains certain obvious coding errors (e.g., the year 3999). Fortunately, the NPDB contains more complete information about the year that a record is entered into the database because a malpractice report is mandated by law to be filed within 30 days of any payment. As a result, this field serves as "a reasonable substitute for the year of [j]udgment or [s]ettlement,"[7] which were optional fields in the early years of NPDB data collection. Excluding years clearly outside the coverage of NPDB (e.g., years after 2006), the correlation between the year of the act or omission and the year the data enter the database is also quite high, with a correlation coefficient of 0.80. While an alternative, albeit more complicated, approach would be to build a model for imputing missing data (Little & Rubin 2002), we instead simply use the fully observed "year of entry" as a reasonable substitute. To the degree that a claim is coded as being subject to a cap when it actually was not due to lag time between the alleged malpractice and ensuing payment, our inferences are likely to be biased toward zero.

relevant unit of randomization when it comes to tort reform is at the state-year; ignoring this is to risk serious Type 1 error.

[6] NPDB Public Use Data File at 9 (Mar. 31, 2005). Because the NPDB also collects information about claim allegation, we also investigated aggregating by subspecialty. On the assumption that some specialties are disparately affected by damage caps, data by state-year specialty may enable researchers to identify effects. However, the claim allegation groups are quite coarse, and do not necessarily represent distinct physician specialties. The categories include, for example, whether a claim is related to diagnosis, anesthesia, surgery, medication, IV & blood products, or obstetrics. Model sensitivity appeared to drive results in models accounting for subspecialty, similar to the results we present below.

[7] NPDB Public Use Data File at 6 (Mar. 31, 2005).

Figure 2: Quantile-quantile plot of pretreatment claims in seven states adopting a cap between 1991–2004 and all claims in 11 states without any cap.



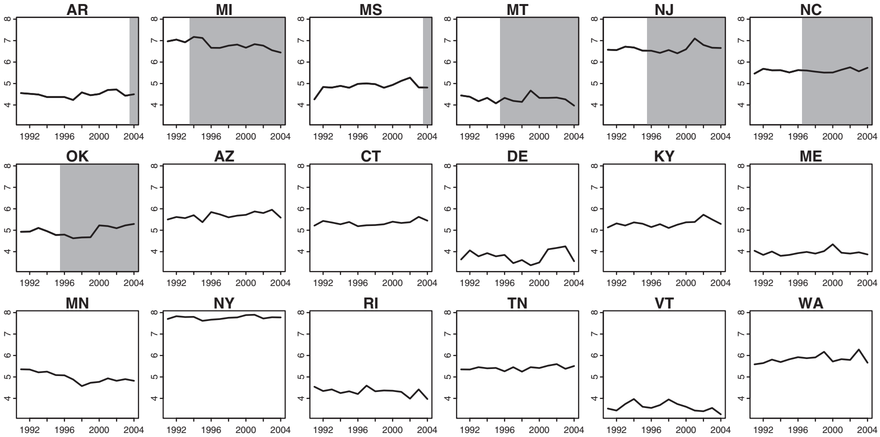
NOTE: The grey box denotes the range of claims in adopting states during pretreatment periods. The 45-degree line indicates perfect balance. Points labeled "New York-Michigan/New Jersey" represent the highest quantiles of claims in adopting and nonadopting states, namely, treatment states of Michigan/New Jersey and the control state of New York, respectively. Because observations are at the state-year level, each state is represented by multiple points.

Other weaknesses of the NPDB are well known (see Helland et al. 2005; Chandra et al. 2005). Duplicate records from multiple payments for a single incident of malpractice exist, payments from corporate entities are excluded from NPDB's coverage, and covariate information on each claim is sparse and/or inconsistent across records. Each of these issues presents potential measurement error, but may well be independent from the adoption of the cap.

Figure 2 presents a quantile-quantile (Q-Q) plot for pretreatment claims in adopting states and all claims for control states.[8] The grey box

[8] Because there is not much of a sample-wide time trend, plotting quantiles of *pretreatment* claims in adopting states and *all* claims in control states does not make much of a difference compared to plotting year-specific QQ plots.

Figure 3: Claims over time.



Note: Claims are logged for visibility across states. The grey shading indicates that a state has adopted a cap.

denotes the range of pretreatment claims for adopting states. The figure illustrates substantial imbalance in pretreatment claims: most noticeably, New York's claim rates, ranging from 2,032 in 1995 to 2,692 in 2001, are far above the highest claims rates of adopting states. Of the adopting states, only Michigan comes close, with some 1,150 claims in 1993. Half of all observations fall within the range of 70 to 300 claims. This foreshadows the degree of model sensitivity we detect later (even when excluding New York): because one of the major determinants of the number of claims is state population size, estimates will prove to be highly sensitive to how one controls for population.

Figure 3 plots claim rates for all states in the sample. Claims are logged to facilitate comparison across states. The first seven panels present adopting states, with the grey bands indicating postadoption periods. Michigan's time series provides suggestive evidence that the cap (effective in 1994) is associated with an overall drop in claims: before 1995, Michigan had more than 1,000 claims in every year (6.9 on the log scale), ranging from 1,015 in 1993 to 1,303 in 1994, but claims dropped to below 1,000 for every subsequent year after 1995, ranging from 630 in 2004 to 929 in 2001. Other adopting states, however, do not exhibit claim rate changes nearly as sharp as Michigan's. The panels also exhibit no clear time trend, which is confirmed

Table 2: Covariate Balance for States Adopting Caps and States Without Caps

			All 18 States			*Excluding NY*		
	Mean Treated	*SE*	*Mean Controls*	*SE*	*p Value*	*Mean Controls*	*SE*	*p Value*
Unemployment rate	5.51	1.25	5.25	1.31	0.11	5.15	1.29	0.03
Income (2004 $10,000)	2.73	0.56	3.06	0.51	0.00	3.01	0.51	0.00
Population (millions)	5.01	3.20	4.54	4.81	0.35	3.14	1.96	0.00
Density	227.37	363.50	269.33	299.35	0.34	257.09	311.39	0.51
Male	0.49	0.01	0.49	0.01	0.36	0.49	0.01	0.07

NOTE: SE indicates standard error. p value reported from a simple t test on the raw marginal distribution of covariate. $N = 252$.

by the lack of a systematic pattern when controlling for year fixed effects or simply the year in an overdispersed Poisson model with state fixed effects. The NPDB claims data exhibit relatively sharp fluctuations in claim rates (e.g., New Jersey in 2001 and Montana in 1999), which may represent measurement error in claim rates.

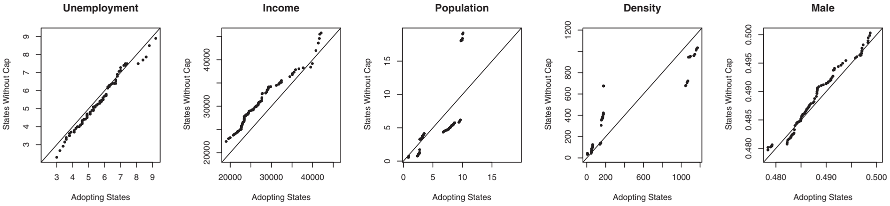
C. Pretreatment Covariates and Balance

For pretreatment covariates, we collected basic state-year information from the U.S. Statistical Abstract (unemployment rate, income converted to 2004 dollars total, population per square mile of land (density), and percent male) and the U.S. Census (population). Each of these covariates varies by year. For illustrative purposes, we limit ourselves to these five pretreatment covariates.[9]

Table 2 presents balance statistics of these covariates, both for the full sample and a sample excluding New York entirely. We compare adopting and nonadopting states not treated and control state-years. An alternative would be to test balance by the same randomization test that we outline below, thereby testing both cross-sectional and time-series balance.

[9] In Table 3, we additionally investigate sensitivity to physicians per capita.

Figure 4: Covariate Q-Q plots.

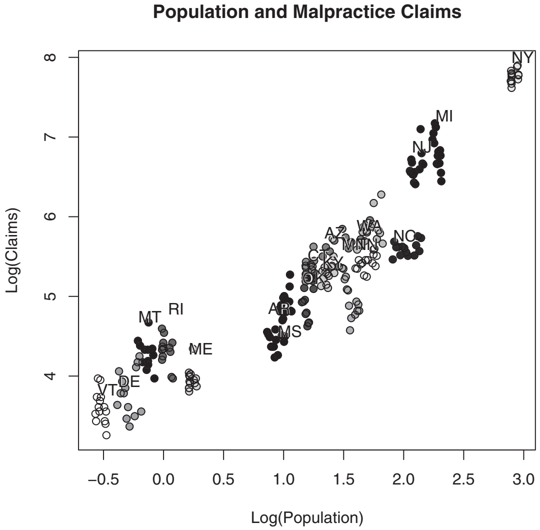


Note: Forty-five-degree line indicates perfect balance.

Despite the sharp claim rate imbalance depicted in Figure 2, simple balance statistics for the full sample (paradoxically) appear reasonable. Adopting states had an average unemployment rate of roughly 5.5 percent, compared to roughly 5.3 percent in control states (p value = 0.11). Even population differences appear close enough to render the two groups of states comparable—adopting states have populations of roughly 5 million and control states have populations of roughly 4.5 million, with standard errors of 3.2 and 4.8, respectively. With an asymptotic means test we detect systematic differences only for income, estimating that control states have an income over $3,300 higher than adopting states. The three last columns of Table 2 show, however, that once we exclude New York, differences appear more substantial. One telling sign is that the standard errors for population are reduced by over 50 percent, leading to a p value less than 0.00.

To visualize the imbalance more clearly, especially compared to the simple t test, Figure 4 plots Q-Q plots for each of the five covariates on the full sample. Perfect balance would lead to alignment with the 45-degree line in each panel. The middle panel illustrates the imbalance in population: similar to the claim rate Q-Q plot, New York is the clear outlier, with a population of roughly 18–19 million. At most quantiles, nonadopting states have higher incomes and lower unemployment rates. These sample imbalances, especially areas of nonoverlapping support, indicate that estimates may be sensitive to distributional assumptions. For example, Figure 5 plots logged population on the x-axis against logged claims on the y-axis. This figure illustrates that while the cross-state correlation between population and claims is clearly positive, (1) the functional form may not be linear, (2) there is substantial interpolation, and (3) intra-state correlation may vary substantially. This suggests that models may be sensitive to adjustments for population.

Figure 5: Population and claims across states.



Note: The shading of each dot indicates a unique state.

IV. Analysis

In this section, we present our analysis of the medical malpractice data. First, we assess the key assumption of exogeneity. Second, we discuss the DID approach that is commonly used in the literature of assessing the effects of legal institutions. We show that conventional approaches are subject to large model sensitivity. Lastly, we nest DID within nonparametric randomization inference, thereby reducing the role of unwarranted distributional assumptions, incorporating more information about treatment assignment, and reducing model dependence.

A. *Exogeneity*

The key assumption of a DID approach is that the enactment of a cap is conditionally exogenous (i.e., that caps are not correlated to unobserved factors that affect malpractice). To make exogeneity more credible, a DID approach can account for state-specific unobserved heterogeneity (e.g., the fact that Arkansas may be different from Arizona in ways we cannot observe) with state fixed effects and year-specific unobserved heterogeneity (e.g., sample-wide "shocks" to the medical system) with year fixed effects. We also control for basic demographic and economic covariates, as described in Section III.C and summarized in Table 2.

One immediate concern is whether these variables are proper pretreatment covariates. If the enactment of caps affects the total population by affecting mortality, or generates general economic effects by inducing or causing deviations from optimal precautions (Shepherd & Rubin 2005), then controlling for population or income may induce posttreatment bias (see, e.g., Ho 2005). Bias is likely to be the greatest for covariates directly affected by the imposition of a medical malpractice cap. Some scholars might attempt to assess sensitivity of results to including and excluding such possibly tainted covariates, which can illustrate the degree of model sensitivity.[10]

Exogeneity is violated if states adopt caps for reasons that are unobserved (and unaccounted for by DID), and those reasons independently affect the number of malpractice claims. For example, suppose there exists cyclicality in claims, such that a cap is only adopted when claims exceed a certain threshold, which could represent, for example, sufficient political clout due to an upsurge in malpractice claims. The number of claims might reduce postadoption solely because claims hit a peak in this natural cycle, and a researcher would falsely attribute the decrease to the enactment of a cap. While such a pathological case of omitted variables is unlikely to be present in our data, the credibility of our inferences still depends on the absence of it. In any study—observational or experimental—certain assumptions must be made, and we seek to minimize the role of unwarranted ones.

B. *Difference-in-Differences*

To characterize our DID approach more formally, let Y_{it} indicate the number of claims in state $i = \{0, \ldots, 18\}$, and year $t = \{1991, \ldots, 2004\}$. The treatment indicator T_{it} equals 1 for an adopting state i when t is greater than the year or cap enactment. Let X_{it} represent the vector of pretreatment covariates. Since our claims data are count data, we use a Poisson model with overdispersion parameter ω, which can be written as:

[10] Yet if a covariate is simultaneously correlated to an unobserved pretreatment factor and affected by the treatment, the true treatment effect is not necessarily even bounded by these estimates (Rosenbaum 1984).

$$\begin{aligned} &Y_{it} \sim \text{Poisson}(\lambda_{it}, \omega) \\ &\lambda_{it} = \exp\{\gamma T_{it} + \beta X_{it} + \delta_i + \tau_t\}, \end{aligned} \tag{1}$$

where γ is the treatment coefficient of interest, δ_i and τ_t are state and year fixed effects, and ω accounts for the fact that there is overdispersion in the data (i.e., the variance is higher than the mean) (see, e.g., Zheng et al. 2006).[11] Implicit in the model is the assumption that state-years are conditionally independent. This assumption could be ciolated if there are state-specific time trends in claims, or if one state's liability regime affect claims brought in another state. (e.g., if plaintiffs engage in forum shopping to avail themselves of higher or nonexistent caps).

A key distributional assumption is that covariates enter *linearly* into Equation (1). Yet sample imbalance implies that estimates may be sensitive to the specification of X_{it}. For example, Table 3 illustrates consequences of controlling for population. The top half excludes, and the bottom half includes, doctors per capita as a covariate. Examining the top half that excludes doctors per capita as a covariate, the first column includes linear and quadratic terms (population and population^2), yielding an *insignificant* treatment effect of a decrease of 6 percent in claims ($1 - \exp(-0.06) \approx 0.06$). The second and third columns show that by controlling for other covariates one might arrive at a statistically significantly negative effect of roughly 10–11 percent. Columns 4 through 6 show that excluding New York—a seemingly reasonable choice given the fact that it is the clear population outlier—can further impact findings, weakening or even reversing full sample estimates. The fourth column controls for linear

[11] Another alternative specification would be to let n_{it}, the population in any state-year, enter as an *offset*, changing the second expression of Equation (1) to:

$$\lambda_{it} = n_{it} \times \exp\{\gamma T_{it} + \beta X_{it} + \delta_i + \tau_t\}, \tag{2}$$

where X_{it} no longer contains population as a covariate. Letting n_{it} enter as an offset is equivalent to controlling for log n_{it} as a covariate and setting its coefficient to 1. We do not do so here to avoid unnecessarily constraining the effect of population. Moreover, for certain specifications we can reject the hypothesis that the coefficient equals 1—for example, controlling for covariates and excluding New York, we reject the hypothesis that the coefficient on log n_{it} equals 1 at the 0.1 level. Specification issues, such as whether to use log n_{it} as opposed to n_{it} and whether other transformations of population (or other offsets) should be employed are, of course, associated with further model dependence, which we aim to reduce in Section IV.C via randomization inference.

Table 3: Treatment Effect Estimates from Overdispersed DID Poisson Model

	Excluding Doctors per Capita as Covariate					
	Full Sample			*Excluding NY*		
Coefficient	–0.06	–0.10**	–0.11**	0.11**	–0.08*	0.07
	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)
Population	Y**	Y**	Y**	Y**	Y**	Y**
Population^2	Y*	N	Y	Y**	N	Y**
Covariates	N	Y	Y	N	Y	Y
N	252	252	252	238	238	238
	Controlling for Doctors per Capita as Covariate					
Coefficient	–0.03	–0.08*	–0.09*	0.13**	–0.05	0.09*
	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Population	Y**	Y*	Y**	Y**	Y**	Y**
Population^2	Y	N	Y	Y**	N	Y**
Covariates	N	Y	Y	N	Y	Y
N	234	234	234	221	221	221

NOTE: All models control for state and year fixed effects. Covariates indicates whether the DID regression controls for unemployment, income, population, population density, and gender. **(*) indicates statistical significance at the 0.05 (0.1) level.

and quadratic population terms in this subset that excludes New York, estimating that damage caps cause a statistically significant *increase* of 11 percent in claims. Adding covariates in Column 5, but excluding the quadratic term, yields a statistically significant decrease of 8 percent, which accords more with the full sample results. Yet when we include all the covariates including the quadratic population term, as in the sixth column, the estimated effect is insignificiant.

The bottom half yields similar model instability across the same specifications when additionally controlling for physicians per capita.[12] (Because information on physicians per capita was not available for 1991, the sample size decreases in these models.) While two of three specifications that include New York might suggest negative effects of a cap (Columns 2 and 3), two of three specifications excluding New York suggest positive effects

[12] Specifically, this covariate measures physicians per 100,000 persons, combined from the Statistical Abstract and yearly volumes published by the American Medical Association (2006).

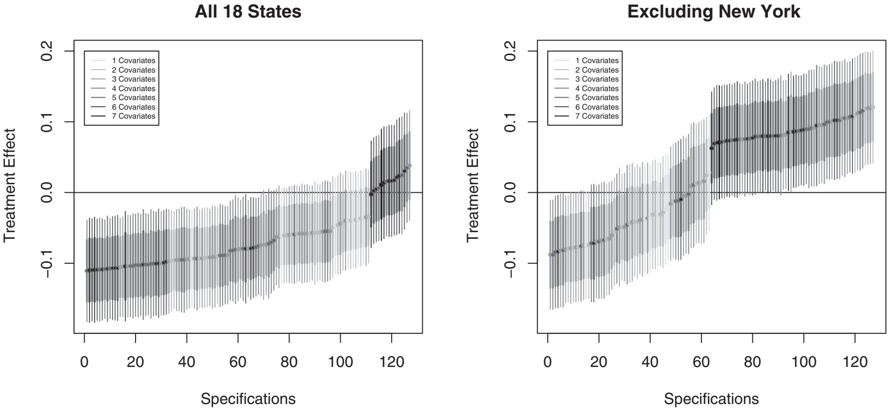
of a cap (Columns 4 and 6). Of course, controlling for physicians per capita may induce posttreatment bias (see, e.g., Kessler & McClellan 1996: 365–66), and so there are good reasons to exclude this covariate. Moreover, because model sensitivity results are similar and physician data were not available for 1991, we do not control for it in our subsequent analysis.

Classical approaches provide some guidance on covariate selection. Predictive criteria typically employed in classical approaches, however, do not necessarily accord with the goal of causal inference. Limitations to classical approaches of covariate selection (Ho et al. forthcoming; Strnad 2006) stem in large part from the difficulty of searching across multidimensional covariate space. Embedded in Table 3, for example, are choices about the functional form of population, the inclusion of physicians, and the inclusion of New York, and dozens of other choices about other covariates. While researchers may be able to winnow down the set of reasonable models using qualitative knowledge, model sensitivity may still plague the remaining set. Here, we do not purport to establish that one of the models presented is the "best," but, given that there are often not very strong theoretical reasons to choose one model over the other, we systematically investigate model sensitivity across specifications.

To do so, we examine treatment effect estimates across all subsets of seven covariates (population, population^2, population^3, unemployment, income, population density, and gender), both for the full sample and the sample excluding New York. In all specifications, we control for state and year fixed effects and allow for overdispersion as in Equation (1). In total, we estimate 127 treatment effects $\left(127 = \Sigma_i^7 \binom{7}{i}\right)$ for each sample.

Figure 6 plots the resulting point estimates from this exercise (ordered by magnitude) with associated standard deviations and 90 percent intervals. The left panel, which depicts estimates for the full sample, suggests that the effect of a cap is to decrease claims—more than 55 percent of the specifications yield statistically significantly negative treatment effects and none yield statistically significantly positive treatment effects. The right panel shows that by excluding New York, the clear outlier (and high leverage point), one can reverse the findings; 34 percent of the specifications now suggest a statistically significantly positive effect, while 7 percent indicate a negative effect. The shading of intervals in Figure 6 also indicates how many covariates are controlled for in each specification—exogeneity certainly becomes more plausible, and estimates thereby more credible (barring posttreatment bias), when controlling for more covariates. The shading in the right panel, for

Figure 6: Model dependence of treatment effect.

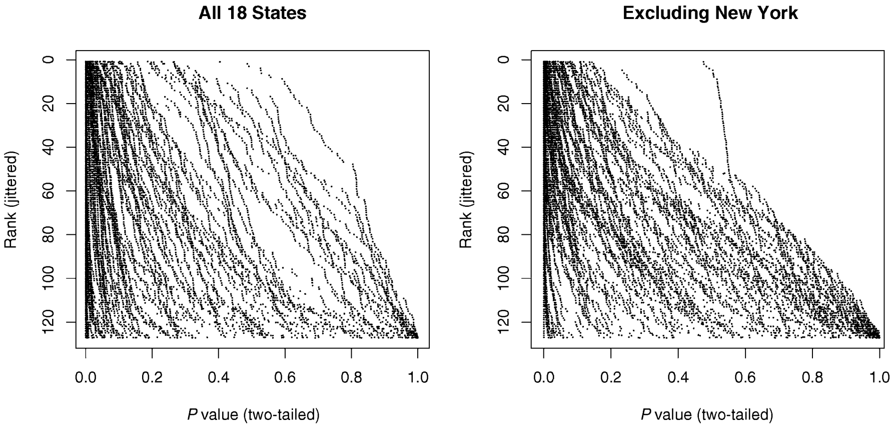


Note: This figure plots point estimates, standard deviations (thick lines), and 90 percent intervals (thin lines) for all 127 DID estimates with any combination of seven pretreatment covariates $(127 = \Sigma_i^7 \binom{7}{i})$. Shading indicates the number of control covariates, with darker shading indicating more control covariates. All specifications include year and state fixed effects.

example, would seem to indicate that by controlling for more covariates (thereby increasing model fit), we have evidence that caps cause an increase in malpractice claims. Although extant research examines samples of different states in different periods and with different specifications, Figure 6 may shed light on the key disagreement in the literature about whether caps have any effect, with some studies finding a negative effect and others finding none: some of the disagreement may be attributable to varying specifications.

Yet another related reason for the disagreement may lie in the fact that conventional estimators assume independent state-year assignment, whereas in our sample states only adopt caps once. As a result, conventional confidence intervals may be too short, leading to large Type I error (Bertrand et al. 2004). To investigate this possibility, Figure 7 plots 127 p values from each of 100 simulated placebo treatments (i.e., 12,700 points). For each simulation, we randomly activated the treatment at any given year for a randomly selected seven states, and estimated Equation (1). Within each simulation, the p values are sorted by magnitude on the y-axis. Given that the treatments are random, we would expect the p values to be roughly uniformly distributed over the unit interval. Yet all values are shifted substantially toward the origin, indicating a sharp trend toward rejecting the null

Figure 7: Asymptotic p values of 127 specifications for 100 placebo treatments.



NOTE: P values are ordered in magnitude within each simulation. This figure illustrates high rejection rates across specifications.

hypothesis that malpractice caps have no effect. In fact, the null is rejected for a substantial plurality of specifications and placebo treatments, even for the highest of all 127 p values for certain given (placebo) treatment vectors, as can be seen by the dark shading in the lower-left corner of the left panel. Rejection rates decrease slightly for the sample excluding New York in the right panel but are still quite high.

Table 4 summarizes rejection rates at conventional significance levels from 1,000 placebo treatments. As illustrated in Figure 7, the rejection rates are staggeringly high. Averaging across simulations and specifications, we reject the null hypothesis of no effects at the 0.01 level 32 percent of the time for the full sample and 27 percent of the time for the sample excluding New York. Rejection rates improve after conditioning on more covariates, but still exhibit high Type I error. Even controlling for all covariates in the full sample, we reject the null hypothesis 40 percent of the time at the 0.1 level. For the sample excluding New York, the rejection rates are slightly less acute, but still far higher than we would expect given the random placebo assignment: we still reject the null hypothesis for 28 percent of all placebo treatments even when controlling for all covariates. Nonetheless, the fact that rejection rates decrease monotonically in the number of covariates for the sample excluding New York suggests that gathering more covariates may be invaluable in validating policy effects. But even with a larger covariate set, the

Table 4: Rejection Rates for DID Regressions for 1,000 Placebo Treatments

All 18 States

		Number of Covariates						
Level	*All*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1%	0.32	0.31	0.32	0.33	0.33	0.31	0.26	0.17
5%	0.47	0.47	0.48	0.48	0.48	0.46	0.41	0.31
10%	0.56	0.55	0.56	0.57	0.56	0.54	0.49	0.40

Excluding NY

		Number of Covariates						
Level	*All*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1%	0.27	0.39	0.37	0.31	0.24	0.17	0.12	0.09
5%	0.41	0.54	0.52	0.46	0.38	0.29	0.23	0.21
10%	0.49	0.62	0.60	0.54	0.46	0.38	0.31	0.28

NOTE: Level indicates the significance level. Top panel presents results for the full sample; bottom panel presents results for the sample excluding New York. In all instances, the rejection rate is substantially higher than expected.

resulting high level of Type I error casts substantial doubt on conventional DID approaches. We therefore offer next an alternative that reduces model dependence, incorporates more information about the assignment mechanism, and reduces the role of unwarranted distributional assumptions.

C. Randomization Inference with DID as the Test Statistic

A considerable number of alternatives to reduce model dependence exist, such as preprocessing the sample by matching on pretreatment covariates (Rubin 1979; Ho et al. forthcoming), model averaging (Hoeting et al. 1999; Strnad 2006), forms of robust regression (Huber 1981; Rousseeuw & Leroy 1987), or bounds analysis (Manski 1990). If the primary concern is with serial correlation, an alternative may also be to extend the model using generalized estimating equations (Liang & Zeger 1986), which, however, additionally require some specification of the correlation structure. We investigate an approach that capitalizes on the fact that parametric estimators can be nested entirely within nonparametric randomization inference (Ho & Imai 2006; Greevy et al. 2004; Rosenbaum 2002b; Fisher 1935), thereby requiring no additional specification of the correlation structure. The intuition of randomization inference is that all the data are fixed, except for the treatment. Under the (sharp) null hypothesis of no treatment effects, we can impute each of the unobserved missing potential outcomes.[13] The only random variable is therefore the treatment. With knowledge of the treatment assignment mechanism, we can then calculate the exact randomization distribution of the treatment coefficient. If our observed treatment coefficient deviates substantially from the randomization distribution of treatment coefficients, this provides evidence to reject the null hypothesis. If outcomes are dichotomous and the test statistic is a simple difference-in-means, the test reduces to the well-known version of Fisher's exact test of the equivalence of two proportions (Fisher 1935).

In more detail, the basic steps of randomization inference are:

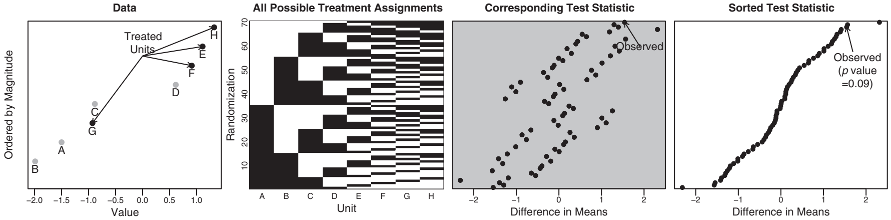
1. Assume the (sharp) null hypothesis that claims are independent of the cap: $Y(T_{it} = 1) = Y(T_{it} = 0)$, where $Y()$ denotes the potential outcomes under treatment and control.
2. Formulate the test statistic, γ (e.g., γ from Equation (1)).
3. For M simulations, permute the treatment according to the *known assignment mechanism*, yielding $m = 1, \ldots, M$ treatment vectors.
4. For each of M treatment vectors, estimate γ_m under the null hypothesis.
5. Calculate the (estimated) p value as: $\frac{\Sigma_m^M I(\gamma_{obs} \leq \gamma_m)}{M}$, where γ_{obs} is the observed test statistic and $I()$ is an indicator function.[14]

Figure 8 presents the intuition with a small data set, where γ is just the difference in means [= mean$(Y(1))$ − mean$(Y(0))$]. The first panel presents the hypothetical data of eight units, labeled A–G, four of which (E–H) are treated and depicted in black. The x-axis represents the outcome measurement, and we can see that three of the treated units (F, E, H) are also the highest in outcomes. The inferential goal is to calculate the probability of these outcomes under the null hypothesis that there is no treatment effect. After imputing missing potential outcomes (e.g., $Y_A(1) = Y_A(0)$), we derive

[13] The sharp null hypothesis differs from the null hypothesis of an *average* treatment effect of zero (see Rosenbaum 2002a).

[14] When the reference distribution is symmetric, one could alternatively compute $\Sigma_m^M I(|\gamma_{obs}| \leq |\gamma_i|)/M$ for a two-tailed test. Because reference distributions may not be symmetric here, we report one-tailed p values.

Figure 8: Hypothetical data set and randomization distribution.

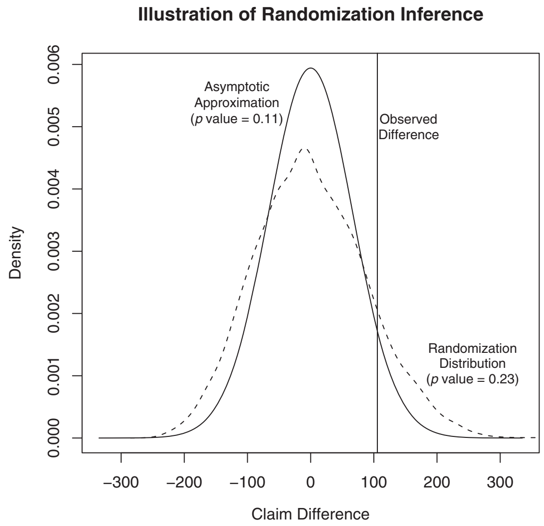


Note: The first panel depicts a hypothetical data set of eight units labeled A–H, where black indicates treated unit. The second panel depicts all possible $\binom{8}{4} = 70$ treatment assignments, with black indicating treated status—the first row corresponds to observed treatment assignment. The third panel calculates the test statistic (difference in means) corresponding to the placebo treatment assignment under the null hypothesis. The fourth panel sorts the test statistics, which is akin to a CDF of the randomization distribution. Comparing the observed test statistic to the reference distribution yields a two-tailed p value of 0.09.

all possible $\binom{8}{4} = 70$ combinations of treatment assignments, as depicted in the second panel, where black indicates treated status and white is control status. The first line hence corresponds to the *observed* treatment status. For each of these potential randomizations, we can then calculate γ_m, as plotted in the third panel. In the first line, for example, as marked by the arrow $\gamma_m = \gamma_{obs} = 1.54$. The last panel sorts the test statistics: γ_{obs} is the third largest test statistic, which yields a two-tailed p value of 0.09. Given the small sample size, the lowest possible (two-tailed) p value is $\frac{2}{70} = 0.03$. We would reject the null under $\alpha = 0.1$.

To illustrate randomization inference using our medical malpractice data, suppose we continue to use a simple difference-in-means as the test statistic. The mean number of claims is 443 claims for treated state-years and 337 for control state-years, yielding $\gamma_{obs} = 106$. For simplicity, assume that the treatment is assigned independently for each state-year, yielding an exact randomization distribution of $\binom{252}{48} \approx 1.2 \times 10^{52}$ treatment vectors and concordant test statistics, where 252 is the total number of observations and 48 is the number of treated state-years. Because calculation of the exact distribution is infeasible, we simulate the distribution with $M = 10{,}000$ draws. Under the null hypothesis, we calculate the difference-in-means for every mth treatment vector. Finally, we calculate the (one-tailed) p value to be 0.23. That is, under the null hypothesis that caps have no impact on claim rates, the probability that we would observe the difference of 106 claims (or more) would be 0.23, *assuming that treatment is independently assigned in each state-year.* Figure 9 plots the randomization distribution for this example in the dashed curve, with the observed test statistic as the vertical line. For

Figure 9: Illustration of randomization inference for difference-in-means as test statistic.



NOTE: This figure overlays the randomization distribution of the difference in means, assuming independent state-year assignment, comparing the distribution to the asymptotic *t* distribution. When assuming independent state-year assignment, the reference distributions are quite similar.

comparison, Figure 9 also plots the conventional approximation of reference distribution, namely, the scaled *t* distribution. The two are quite similar, showing that conventional tests approximate the true randomization distribution, although an asymptotic means test difference yields a lower *p* value of 0.11.

The key insight for our application, elaborated on most lucidly by Rosenbaum (2002b), is that any test statistic that is a function of the treatment can be used to conduct randomization inference. The primary benefits of this approach are twofold. First, randomization inference relaxes unfounded distributional assumptions (e.g., a reference *t* distribution). Second, and most important for our application, randomization inference permits us to directly incorporate information about the *assignment mechanism*, modeling the fact that caps are not adopted independently for each state-year. Quite to the contrary, caps are typically adopted during one year and not repealed, which we incorporate directly. The consequences of ignoring the assignment mechanism have been recognized, for example, by

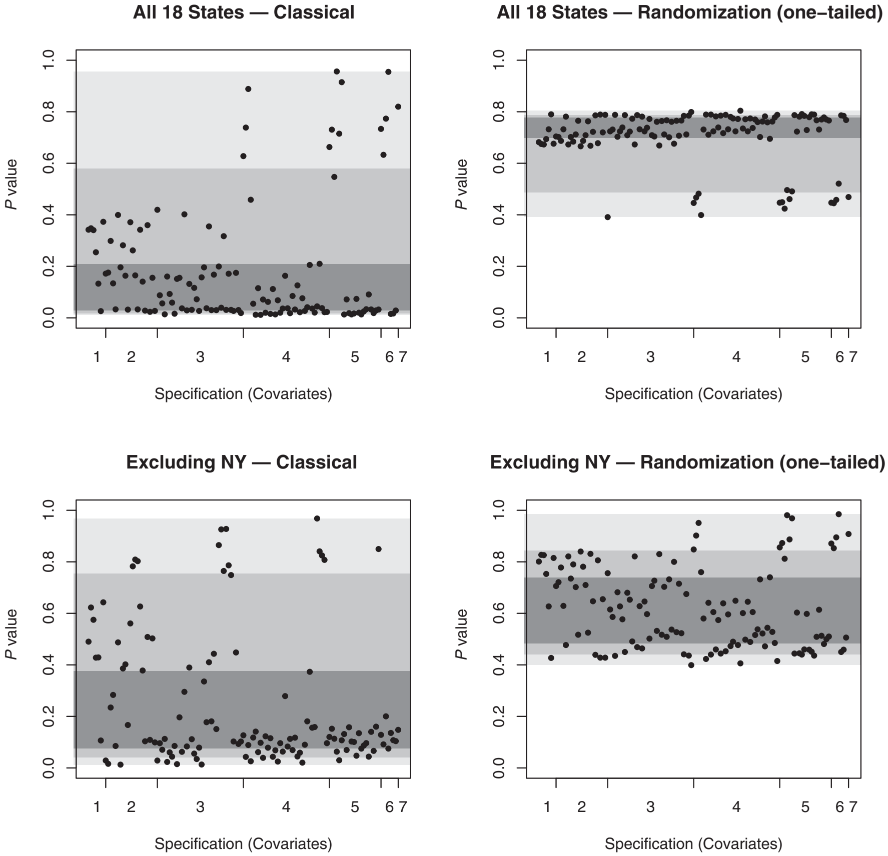
Bertrand et al. (2004), who emphasize the inconsistency of standard errors when treatment and outcomes are serially correlated, and Ho and Imai (2006), who note inadequate coverage of standard estimators when treatment is assigned by systematic rotation with a random start. While Bertrand et al. (2004) propose permutation-based tests as a direct solution, we are aware of few DID studies implementing randomization inference, likely because the approach is quite computationally intensive (but see Donohue & Ho 2005; Rosenbaum 2002b).

To understand why randomization inference is computationally intensive, note that the exact randomization distribution for our application, assuming conditionally exogenous years and states of adoption, consists of $2 \times 10^{12} \left(= (\#\,\text{Years} - 1)^{\#\,\text{Treated States}} \times \binom{\#\text{States}}{\#\,\text{Treated States}}\right)$ test statistics. Because exact calculation is infeasible, we approximate the distribution with 1,000 simulations, randomly activating the treatment at any year (most importantly preserving the assignment mechanism of *one* cap adoption in the observation period), for any (randomly sampled) seven treated states, and calculating the test statistic for each simulated sample.[15] We use γ from Equation (1) as our test statistic. To investigate model sensitivity, we conduct the same process for both the full sample and the sample excluding New York, and for each of 127 model specifications.

Figure 10 plots p values from classical DID estimates in the left panels and randomization-based p values in the right panels. The top panels present results for the full sample; the bottom panels present results from the sample excluding New York. The x-axis presents the 127 specifications, ordered by the number of covariates included in the specification (in addition to state and year fixed-effects and the treatment vector). The grey bands indicate the range and 80 percent and 50 percent quantiles of the p values across specifications. The range of p values (in light grey) for classical estimates spans the entire unit interval, whereas it covers a much narrower range of (0.39, 0.80) for randomization-based p values. In addition, randomization-based p values all fall outside any conventional significance levels, while classical estimates

[15] A closely related alternative would be the block bootstrap, which samples observed treatment vectors for each state with replacement. Compared to randomization inference, the block bootstrap reduces the total number of combinations considerably, but since we simulate from the exact distribution in both instances, the benefits are (1) the (relatively small) reduction in variability of estimates due to sampling from a smaller total number of combinations, and (2) more direct computation of standard errors and confidence intervals (see Appendix B).

Figure 10: Comparison of *p* values from conventional DID estimates and randomization inference for all 127 specifications.



lead to high rejection rates. These results are substantially the same for the full sample and the sample excluding New York. These findings demonstrate that model sensitivity is reduced by randomization inference in our application. Note that randomization inference requires no change in extant identification approaches; it simply relaxes the role of unwarranted distributional assumptions and incorporates more information about treatment assignment. These results thereby suggest that scholars may be more likely to arrive at the truth when using randomization inference.

That said, randomization inference does have several practical downsides. First, as already mentioned, one of the major downsides is its computational intensity (Tamura et al. 1994:771): randomization inference for the full and trimmed sample with 1,000 simulations for 127 specifications requires 254,000 DID regressions ($= 2 \times 127 \times 1{,}000$).[16] Second, the test of the null hypothesis itself does not provide directly interpretable substantive estimates of the effect. In certain circumstances, one can invert the test to calculate the p value associated with each null hypothesis of γ (Imbens & Rosenbaum 2005; Ho & Imai 2006). But inverting the test for DID estimation is not straightforward, as we outline in Appendix B. Third, randomization inference tests the *sharp* null of no effects at the unit-level (but see Rosenbaum 2002a:322–24), which may not be reasonable under certain conditions (Rosenbaum 2003; Ho & Imai 2006). Pragmatic costs will, of course, only decrease with advances in computational technology.

V. Conclusion

Our findings are, on the one hand, sobering and, on the other, encouraging: sobering because of the fragility that we find in existing DID approaches, and encouraging because a more robust alternative exists. By employing randomization inference, we are led to conclude that medical malpractice caps have not altered the number of medical malpractice awards and settlements.

Substantively, it remains a puzzle whether the imposition of caps on medical malpractice awards affects physicians. Null findings on claim rates are consistent both with (1) no effects on physicians and no effects on litigants, or (2) cross-cutting effects of equal magnitude (large or small) on both. Isolating physician and litigant effects in light of these dual effects remains a major challenge in the medical malpractice literature.

Methodologically, our study yields useful insights for applied scholars. First, gathering more covariates appears to reduce (apparently) false null hypothesis rejection rates with DID approaches. Second, balance checks may provide telltale warning signs of model sensitivity. Third, and most importantly, randomization inference permits scholars to model the assignment mechanism directly and to reduce unwarranted distributional assumptions. In so doing, randomization inference sets forth a "reasoned basis for inference," as Fisher coined, for policy evaluation.

[16] The process took roughly four hours on a Pentium M 1.4 Ghz machine. To calculate this using the exact distribution would take over 900,000 years ($\approx 2 \times 10^{12} \times 4/(1000 \times 24 \times 365)$), meaning the process would need to have started sometime in the Pleistocene Epoch.

References

Adams, J., & V. J. Hotz (2000) "Evaluating Welfare Reform in an Era of Transition," in *The Statistical Power of National Data to Evaluate Welfare Reform*, pp. 209–19. Washington, DC: National Academy Press.

American Medical Association (2006) *Physician Characteristics and Distribution in the U.S.* Chicago, IL: AMA Press.

Anderson, G. F., P. S. Hussey, B. K. Frogner, & H. R. Waters (2005) "Health Spending in the United States and the Rest of the Industrialized World," 24 *Health Affairs* 903.

Baldwin, L.-M., L. G. Hart, M. Lloyd, M. Fordyce, & R. A. Rosenblatt (1995) "Defensive Medicine and Obstetrics," 274 *J. of the American Medical Association* 1606.

Beider, P., & S. Hagen (2004) *Limiting Tort Liability for Medical Malpractice.* Washington, DC: U.S. Congressional Budget Office.

Bertrand, M., E. Duflo, & S. Mullainathan (2004) "How Much Should We Trust Differences-in-Differences Estimates?" 119 *Q. J. of Economics* 249.

Black, B., C. Silver, D. A. Hyman, & W. M. Sage (2005) "Stability, not Crisis: Medical Malpractice Claim Outcomes in Texas, 1988–2002," 2 *J. of Empirical Legal Studies* 207.

Brennan, T., L. Leape, N. Laird, L. Hebert, A. Localio, A. Lawthers, J. Newhouse, P. Weiler, & H. Hiatt (1991) "Incidence of Adverse Events and Negligence in Hospitalized Patients: Results of the Harvard Medical Malpractice Study I," 324 *New England J. of Medicine* 370.

Chandra, A., S. Nundy, & S. A. Seabury (2005) "The Growth of Physician Medical Malpractice Payments: Evidence from the National Practitioner Data Bank," *Health Affairs* 240.

Dhankhar, P., M. M. Khan, & I. M. S. Alam (2005) *Threat of Malpractice Lawsuit, Physician Behavior and Health Outcomes: Testing the Presence of Defensive Medicine.* Unpublished Manuscript.

Donohue III, J. J., & D. E. Ho (2005) *Does Terrorism Increase Crime? A Cautionary Tale.* Unpublished Manuscript.

Dubay, L., R. Kaestner, & T. Waidmann (1999) "The Impact of Malpractice Fears on Cesarean Section Rates," 18 *J. of Health Economics* 518.

Fisher, R. A. (1935) *The Design of Experiments.* London: Oliver and Boyd.

Gelman, A., & J. Hill (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models.* New York: Cambridge University Press.

Greevy, R., J. H. Silber, A. Cnaan, & P. R. Rosenbaum (2004) "Randomization Inference with Imperfect Compliance in the Ace-Inhibitor After Anthracycline Randomized Trial," 99 *J. of the American Statistical Association* 7.

Helland, E., J. Klick, & A. Tabarrok (2005) "Data Watch: Tort-Uring the Data," 19 *J. of Economic Perspectives* 207.

Ho, D. E. (2005) "Why Affirmative Action Does Not Cause Black Students to Fail the Bar," 114 Y*ale Law J.* 1997.

Ho, D. E., & K. Imai (2006) "Randomization Inference with Natural Experiments: An Analysis of Ballot Effects in the 2003 California Recall Election," 101 *J. of the*

American Statistical Association 888. Available at http://imai.princeton.edu/research/fisher.html.

Ho, D. E., K. Imai, G. King, & E. A. Stuart (forthcoming) "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis*. Available at http://gking.harvard.edu/files/matchp.pdf.

Hoeting, J. A., D. Madigan, A. E. Raftery, & C. T. Volinsky (1999) "Bayesian Model Averaging: A Tutorial," 14 *Statistical Science* 382.

Huber, P. J. (1981) *Robust Statistics*. New York: Wiley.

Hussey, P. S., G. F. Anderson, R. Osborn, C. Feek, V. McLaughlin, J. Millar, & A. Epstein (2004) "How Does the Quality of Care Compare in Five Countries?" 23 *Health Affairs* 89.

Imbens, G. W., & P. R. Rosenbaum (2005) "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education," 168 *J. of the Royal Statistical Society, Series A* 109.

Kessler, D. & M. McClellan (1996) "Do Doctors Practice Defensive Medicine?" 112 *Q. J. of Economics* 353.

—— (2002) "How Liability Law Affects Medical Productivity," 21 *J. of Health Economics* 931.

King, G., & L. Zeng (2006) "The Dangers of Extreme Counterfactuals," 14 *Political Analysis* 131.

Klick, J., & T. Stratmann (2003) *Does Medical Malpractice Reform Help States Retain Physicians and Does it Matter?* Unpublished Manuscript.

Liang, K.-Y., & S. L. Zeger (1986) "Longitudinal Data Analysis Using Generalized Linear Models," 73(1) *Biometrika* 13.

Little, R. J., & D. B. Rubin (2002) *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley & Sons.

Localio, A., A. Lawthers, T. Brennan, N. Laird, L. Hebert, L. Peterson, J. Newhouse, P. Weiler, & H. Hiatt (1991) "Relation Between Malpractice Claims and Adverse Events Due to Negligence: Results of the Harvard Medical Malpractice Study III," 325 *New England J. of Medicine* 245.

Manski, C. F. (1990) "Non-Parametric Bounds on Treatment Effects," 80 *American Economic Rev., Papers & Proceedings* 319.

Moulton, B. R. (1990) "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unit," 72(2) *Rev. of Economics & Statistics* 334.

Raudenbush, S. W., & A. S. Bryk (2002) *Hierarchical Linear Models*. London, UK: Sage.

Rosenbaum, P. R. (1984) "The Consequences of Adjusting for a Concomitant Variable that Has Been Affected by the Treatment," 147(5) *J. of the Royal Statistical Society, Series A* 656.

—— (2002a) "Covariance Adjustment in Randomized Experiments and Observational Studies (with Discussion)," 17 *Statistical Science* 286.

—— (2002b) *Observational Studies*, 2nd ed. New York, Springer Verlag.

—— (2003) "Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed Rank Test," 57 *American Statistician* 132.

Rousseeuw, P. J., & A. M. Leroy (1987) *Robust Regression and Outlier Detection.* New York: Wiley.

Rubin, D. B. (1979) "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," 74 *J. of the American Statistical Association* 318.

Sharkey, C. M. (2005) "Unintended Consequences of Medical Malpractice Damages Caps," 80 *New York Univ. Law Rev.* 391.

Shepherd, J. M., & P. H. Rubin (2005) *Tort Reform and Accidental Deaths.* Unpublished Manuscript.

Strnad, J. (2006) *Should Legal Empiricists Go Bayesian?* Unpublished Manuscript.

Tamura, R. N., D. E. Faries, J. S. Andersen, & J. H. Heiligenstein (1994) "A Case Study of an Adaptive Clinical Trial in the Treatment of Out-Patients with Depressive Disorder," 89 *J. of the American Statistical Association* 768.

Tussing, A. D., & M. A. Wojtoqycz (1992) "The Cesarean Decision in New York State, 1986," 30 *Medical Care* 529.

U.S. Department of Health and Human Services (2003) *Addressing the New Health Care Crisis: Reforming the Medical Litigation System to Improve the Quality of Health Care.* Office of the Assistant Secretary for Planning and Evaluation Report.

Vidmar, N., P. Lee, K. MacKillop, K. McCarthy, & G. McGwin (2005) "Uncovering the 'Invisible' Profile of Medical Malpractice Litigation: Insights from Florida," 54 *DePaul Law Rev.* 315.

Viscusi, W. K., & P. Born (1995) "Medical Malpractice Insurance in the Wake of Liability Reform," 24 *J. of Legal Studies* 463.

Weiss, M. D., M. Gannon, & S. Eakins (2003) *Medical Malpractice Caps: The Impact of Non-Economic Damage Caps on Physician Premiums, Claims Payout Levels, & Availability of Coverage.* Weiss Ratings.

Yoon, A. (2001) "Damage Caps and Civil Litigation: An Empirical Study of Medical Malpractice in the South," 3 *American Law & Economics Rev.* 199.

Zeiler, K. (2004) *An Empirical Study of the Effects of State Regulations on Medical Malpractice Litigation Decisions.* Unpublished Manuscript.

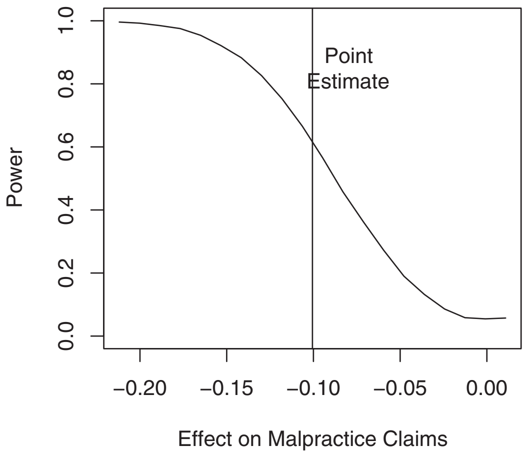
Zheng, T., M. J. Salganik, & A. Gelman (2006) "How Many People Do You Know in Prison?: Using Overdispersion in Count Data to Estimate Social Structure in Networks," 101 *J. of the American Statistical Association* 409.

Zuckerman, S., R. R. Bovbjerg, & F. Sloan (1990) "Effects of Tort Reforms and Other Factors on Medical Malpractice Insurance Premiums," 27 *Inquiry* 167.

Appendix A: Statistical Power

Statistical power—the probability of rejecting the null hypothesis when an effect exists—remains a crucial part of experimental and observational design. We present posthoc power calculations to obtain a sense of the power of the overdispersed poisson DID model with our data set (e.g., Adams & Hotz 2000). Our examination focuses on the model using the full sample, controlling for population, unemployment, income, population density, and gender, with state and year fixed effects (i.e., Column 2 of Table 3). We derive the power curve by simulation: first, for each hypothesized effect (the DID coefficient), we generate 10,000 data sets conditioning on the model (except, of course, the DID point estimate) and covariates; second, for each simulated data set (for each given effect size), we estimate the overdispersed poisson DID regression and conduct a hypothesis test for the DID coefficient at $\alpha = 0.05$; the mean rejection rate for a given effect size represents the estimated power. Figure A1 plots results, showing that statistical power varies substantially with the effect size. At the empirical point estimate the probability of detecting an effect at $\alpha = 0.05$ is just above 0.6. Increasing sample size and gathering more covariates to increase precision would improve statistical power. Disaggregating the NPDB data into finer subsets remains challenging because of relatively sparse information collected for each claim.

Figure A1: Power to detect an effect for overdispersed DID Poisson model.



NOTE: Vertical line indicates the empirical point estimate.

Appendix B: Inverting a Randomization Inference Test with DID/DDD

Inverting a randomization inference test when multiple random variables affect potential outcomes in multiple ways poses certain challenges. To illustrate the issue, take a conventional cross-sectional estimator, using as the test statistic the coefficient on a treatment indicator (β) from:

$$E(Y_i) = \alpha + \beta T_i + \gamma X_i. \tag{B1}$$

To test the sharp null hypothesis that $\beta = 0$, we simply impute missing potential outcomes as:

$$\{Y_i(1)|T_i = 0\} = \{Y_i(0)|T_i = 0\}$$
$$\{Y_i(0)|T_i = 1\} = \{Y_i(1)|T_i = 1\}.$$

Then we permute the treatment, estimate Equation (B1) for each permuted sample, and compare how the observed β_{obs} compares to the distribution of βs. For example, the exact *p* value is calculated as:

$$\frac{\sum_m^M \{I(\beta_{obs} \leq \beta_m)\}}{M},$$

for *M* permutations of the treatment, and accordingly values of β_m, where $I()$ is an indicator function.

To invert the test, we can set up a grid of J β's. For any null hypothesis $\beta = \beta_j$,

$$\{Y_i(1)|T_i = 0\} = \{Y_i(0)|T_i = 0\} + \beta_j \tag{B2}$$

$$\{Y_i(0)|T_i = 1\} = \{Y_i(1)|T_i = 1\} - \beta_j. \tag{B3}$$

For each β_j, we permute the treatment, and calculate the *p* value, thereby obtaining the *p* value plot (analogous to the CDF).

Now suppose we take a basic DID estimator:

$$E(Y_i) = \alpha + \beta_1(D_i * T_i) + \beta_2 D_i + \beta_3 T_i + \gamma X_i, \tag{B4}$$

where D_i is an indicator for the treated group, and T_i is an indicator for before/after treatment, such that β_1 represents the treatment effect of interest.

The problem arises from the fact that the DID estimators technically involve two types of exogeneity assumptions, namely, exogeneity about time (when treatment is adopted) and states (that either state could have adopted the treatment), each of which enter separately in Equation (B4). To conduct randomization inference for the null hypothesis, we permute both D and T to calculate the p value.

But for the inversion, the null hypothesis $\beta = \beta_j$ is not sufficient to impute missing potential outcomes because T_i and D_i also affect β_3 and β_2, respectively. To impute missing potential outcomes for controls, we can use Equation (B4):

$$\begin{aligned}
\{Y_i(1,1)|T_i = 0, D_i = 0\} &= \{Y_i(0,0)|T_i = 0, D_i = 0\} + \beta_{1j} + \beta_2 + \beta_3 \\
\{Y_i(1,1)|T_i = 1, D_i = 0\} &= \{Y_i(1,0)|T_i = 1, D_i = 0\} + \beta_{1j} + \beta_2 \\
\{Y_i(1,1)|T_i = 0, D_i = 1\} &= \{Y_i(0,1)|T_i = 0, D_i = 1\} + \beta_{1j} + \beta_3,
\end{aligned}$$

where $Y_i(T_i, D_i)$ is written as a function of the couple T_i and D_i. For the treated, we similarly have:

$$\begin{aligned}
\{Y_i(0,1)|T_i = 1, D_i = 1\} &= \{Y_i(1,1)|T_i = 1, D_i = 1\} - \beta_{1j} - \beta_3 \\
\{Y_i(1,0)|T_i = 1, D_i = 1\} &= \{Y_i(1,1)|T_i = 1, D_i = 1\} - \beta_{1j} - \beta_2 \\
\{Y_i(0,0)|T_i = 1, D_i = 1\} &= \{Y_i(1,1)|T_i = 1, D_i = 1\} - \beta_{1j} - \beta_2 - \beta_3.
\end{aligned}$$

The inversion hence becomes substantially more complicated, as potential outcomes are a function of two random variables entering in three terms. The problem would be exacerbated for a triple-difference approach, where each unit has eight potential outcomes as a function of three random variables entering in seven terms.

Tentatively, there appear to be three possible approaches to dealing with this problem. First, one could impute missing potential outcomes using only β_1. This would seem to make sense because we are interested only in *one* treatment effect, but by collapsing $Y(1,0) = Y(0,1) = Y(0,0)$, we implicitly assume that β_2 and β_3 equal 0, which is inconsistent with the DID model. Second, we might impute all three missing potential outcomes, holding β_2 and β_3 at fixed values, such as at observed values. Lastly, we might explore three-dimensional parameter space simultaneously, setting up a three-dimensional grid for DIDs (or analogously a seven-dimensional grid for triple differences).

Many papers have recognized the usefulness of randomization inference (see e.g., Bertrand et al. 2004; Imbens & Rosenbaum 2005), but we are not aware of any discussing this particular complication in inverting the test.