

Endogeneity or Anticipation? Evidence from the Effect of Tort Reform on Physician Supply

Anup Malani and Julian Reif*

January 4, 2012

Abstract

While conducting empirical work, researchers sometimes observe changes in outcomes before adoption of a new treatment program. The conventional diagnosis is that treatment is endogenous. Observing changes in outcomes prior to treatment is also consistent, however, with the anticipation effects that arise naturally out of many theoretical models. This paper illustrates the importance of properly distinguishing endogeneity from anticipation effects by showing that incorrectly equating pre-period trends with endogeneity can lead to biased estimates. The paper then provides a framework for comparing the different methods for estimating anticipation effects and proposes a new set of instrumental variables that can address the problem that subjects' expectations are unobservable. Finally, this paper examines a specific set of tort reforms that are unlikely to be endogenous to physician supply but likely to have been anticipated by physicians. Accounting for anticipation effects increases the estimated effect of these reforms by a factor of four compared to a model that equates pre-period changes with endogeneity.

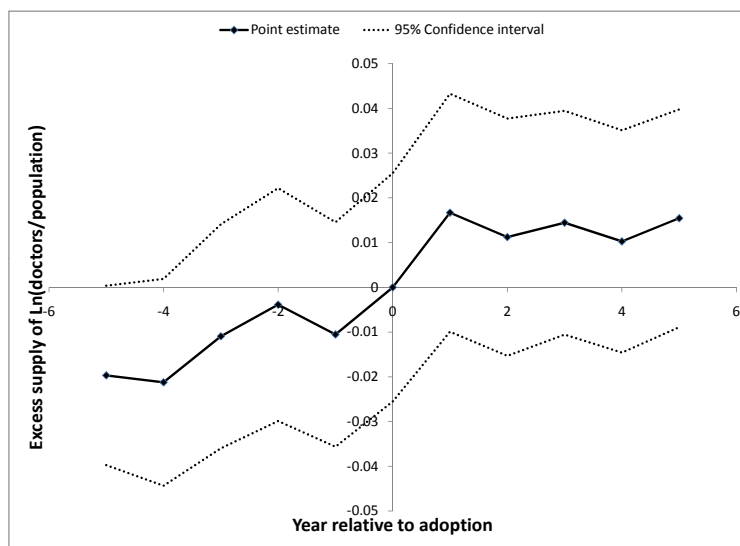
Keywords: Anticipation, Medical Malpractice, Endogeneity, Tort Reform

JEL Classification Numbers: C50, K13, J20

*University of Chicago, RFF and NBER; University of Chicago. We thank Dan Black, Amitabh Chandra, Tatyana Deryugina, Steve Levitt, Jens Ludwig, Derek Neal, Seth Seabury, Heidi Williams, and participants at workshops and conferences at the Searle Center at Northwestern University, Harvard University, New York University, and the University of Chicago for helpful comments. Anup Malani thanks the Samuel J. Kersten Faculty Fund at the University of Chicago for funding. Julian Reif thanks the National Science Foundation for financial support.

While conducting empirical work, researchers sometimes observe changes in outcomes before adoption of a new treatment program or policy. Figure 1 provides an example from the medical malpractice liability context. It shows that equilibrium physician labor supply increased well before states adopted caps on punitive damages to lower physician liability.¹ The conventional diagnosis researchers make upon observing such a pattern in the data is that the treatment was endogenous: it was adopted in response to changes in pre-period outcomes.²

Figure 1: Excess physician supply before and after punitive damage caps: annual leads and lags from 5 years before to 5 years after adoption



Note: This figure plots the normalized coefficients λ_j from the following regression: $y_{ist} = \sum_{j=-5}^5 \lambda_j D_{st+j} + \gamma X_{ist} + u_{ist}$, where y_{ist} is the log of the physician count for specialty i in state s in year t , D_{st+j} is an indicator for whether punitive damage caps was first adopted in period $t + j$, and X_{ist} includes state-specialty and specialty-year fixed effects.

Observing changes in outcomes prior to treatment is also consistent, however, with anticipation effects. Perhaps individuals began changing their behavior in response to an expectation that they would be treated in the future. Anticipation is a reasonable diagnosis if individuals are forward looking, have access to information on future treatment, and there is a benefit to acting before treatment is adopted.

¹It might be surprising that physician supply responds at all to punitive damages since such damages constitute only 1-4% of total malpractice awards (Cohen 2004). We show in Section 4.3.1, however, that this statistic underestimates the impact of punitive damages on physician behavior.

²More generally, endogeneity can be thought of as any correlation between treatment and the error term.

It is unlikely, for example, that the treatment in Figure 1 was endogenous. Punitive damage caps were targeted at all lawsuits, not just medical malpractice suits, and were adopted in states with a wide range of physician supply levels. Moreover, it is quite plausible that physicians simply anticipated the reform. Newspapers and medical malpractice insurance companies signaled there would be reform years prior to actual adoption with news stories and changes in premiums, respectively. Finally, physicians have a large financial incentive to change behavior prior to adoption: assets accumulated prior to the reform are placed at risk by medical errors that occur after the reform.³

To be clear, we do not argue that researchers should always interpret pre-period trends as evidence of anticipation effects. Those trends could be evidence of either endogeneity or anticipation effects, or both. Instead our narrow claim is that researchers should not always default into equating pre-period trends with endogeneity. The proper interpretation of such trends depends on the particular model and application in question. Moreover, the interpretation chosen matters to identification of treatment effects. A researcher who does not account for an endogenous trend may over- or underestimate the true treatment effect, but a researcher who does not account for anticipation usually underestimates the treatment effect. The reason is that the typical before-after comparison attributes anticipatory treatment effects to the pre-period. As a result, it not only ignores, but deducts, anticipatory treatment effects from the overall treatment effect.⁴ To illustrate this point, we consider an application – depicted in Figure 1 – where pre-period trends are more likely to be due to anticipation effects than to endogeneity. We demonstrate that interpreting pre-period trends as anticipation effects produces dramatically different estimated treatment effects than interpreting those trends as evidence of endogeneity.

With this objective in mind, we organize the paper around three contributions.

³In some cases, states make tort reforms explicitly retroactive, in which case the reform applies directly to errors committed before the reform is adopted (Avraham 2007). Even when physicians are insured, they may have incentives to anticipate reforms. First, certain types of damages, such as punitive damages, which we study, are not covered by malpractice insurance. Second, insurance companies have incentive to anticipate reforms, because it takes time to build reserves, and insurance companies may pass on those costs through higher premiums in earlier years. We show this was the case in Section 4.3.3.

⁴This is not always the case. For example, property owners who anticipated the Endangered Species Act deforested land with endangered species before the law went into effect so that the Act would not restrict development on their land (Lueck and Michael 2003). Anticipation of the statute thus increased habitat destruction before adoption and reduced it after adoption.

First, we provide a framework for rigorously comparing the different models that may be employed to estimate anticipation effects. In particular, the framework reveals the assumptions embedded in different empirical models of anticipation effects, describes how those models change as those assumptions are modified, and discusses the relative merits of different assumptions and models. Second, we examine how to address the problem that agents' expectations are unobservable and propose a new set of instrumental variables that can be employed to overcome it. Finally, we estimate the effect of certain tort reforms – selected because they are unlikely to be endogenous – on physician supply and show that accounting for anticipation effects increases the estimated effect of tort reform by a factor of four compared to a model that equates pre-period changes with endogeneity.

Our anticipation effects framework begins with a forward-looking regression of the form

$$y_t = \lambda_0 d_t + \sum_{j=1}^{\infty} \lambda_j E_t [d_{t+j}] + e_t \quad (1)$$

where y_t is some outcome, $\{d_{t+j}\}$ are a sequence of future treatment states, and E_t indicates expectation taken with respect to an agent's information set at time t .⁵ This forward-looking regression model has a wide array of applications, including investments in human capital (e.g., Ryoo and Rosen 2004), rational addiction (e.g., Becker, Grossman, and Murphy 1994), present value models (e.g., Chow 1989), R&D investment decisions (e.g., Acemoglu and Linn 2004), pricing of durable goods (e.g., Kahn 1986) and real estate (e.g., Poterba 1984).

Two main difficulties arise when estimating this model. First, there are potentially an infinite number of anticipation terms. Second, those anticipation terms are generally unobserved.

Consider the problem of infinite anticipation terms. A common response in the empirical microeconomics literature is to estimate a “quasi-myopic” model that omits anticipation terms more than S periods prior to treatment.⁶ Indeed, this is the sort of model employed to generate pre-post graphs such as Figure 1 that are ubiquitous in

⁵In general, we will use i to index agents. However, except in Section 3.1, we suppress the index i in the forward-looking model (1) to simplify the exposition. We assume throughout that the agent's information set is $\Omega_t = \{y_0, \dots, y_{t-1}, x_0, \dots, x_t, d_0, \dots, d_t\}$, where the x are possible covariates.

⁶Although there are a large number of examples, the following are typical: Acemoglu and Linn (2004), Ayers, Cloyd, and Robinson (2005), Bhattacharya and Vogt (2003), de Figueiredo and Vanden Bergh (2004), Finkelstein (2004), Gruber and Koszegi (2001), Heckman and Robb (1985), Jacobson, LaLonde, and Sullivan (1993), Lemos (2006), Lueck and Michael (2003), and Mertens and Ravn (2011).

this literature.⁷ If agents respond earlier than S periods prior to treatment, however, this model will suffer from omitted variable bias.

An alternative approach common in the finance and macroeconomics literature is to posit outcomes as a function of exponentially discounted expectations about future treatment (e.g., Chow 1989). In this formulation treatment has a constant, contemporaneous treatment effect of β and an anticipation effect j periods prior to treatment of $\beta\theta^j$. This model resembles a present-value asset pricing model. Exponential discounting has the useful feature that suitable differencing can eliminate nearly all anticipation terms. Depending on assumptions made about what agents forecast, the resulting Euler equation may be what macroeconomists call the forward-looking rational expectations model. This equation takes the familiar form

$$y_t = \theta E_t [y_{t+1}] + \beta d_t + \varepsilon_t \quad (2)$$

Our framework advances the literature by highlighting the assumptions required to generate the precise regression models estimated in prior literature as well as alternative regression models that emerge if assumptions are changed. It also provides a common benchmark for both the quasi-myopic and exponential discounting models that for the first time allows a comparison of the merits of each.

The second problem with estimating a model of anticipation effects is that expectations are generally not observed. A common response is to examine shocks that alter expectations about treatment but do not actually administer a treatment. An example is a regulation that is enacted at time t but not implemented until time $t+k$ (e.g., Alpert 2010, Gruber and Koszegi 2001, Lueck and Michael 2003, Blundell, Francesconi, and Van der Klaauw 2010).⁸ We show that unless actual expectations are observed, however, the investigator can only demonstrate that expectations affect outcomes; she can not identify the precise slope of the relationship.

An alternative approach is to assume a model of belief formation, such as rational or adaptive expectations, in order to substitute observable variables for unobservable expectations of a variable. Unless the forecast error is orthogonal to the observable variables, however, the researcher will have to instrument for them. The traditional source for these instruments is a subset of the agent's information set, for instance,

⁷See, e.g., Autor, Donohue, and Schwab (2006) and Finkelstein (2004).

⁸For studies that examine shocks to information and no eventual treatment ($t = \infty$), see Stango (2003) and Karpoff, Lott, and Wehrly (2005).

lags of the observable variable (see McCallum 1976). These lags influence the agent’s unobservable forecast of a variable but do not directly influence the outcome variable. For instance, in equation (2), lags of y_t may be suitable instruments for $E_t[y_{t+1}]$.

We propose an alternative set of instruments: leads of the observable variable. From equation (1), we know that both $E_t[y_{t+1}]$ and leads of y_{t+1} depend on future treatment, and are thus correlated. Moreover, from the Euler condition (2), we know the outcome variable y_t , when conditioned on $E_t[y_{t+1}]$, does not depend on outcomes beyond time $t+1$. Thus, in equation (2), leads of y_{t+1} may also be suitable instruments for $E_t[y_{t+1}]$. Our paper outlines the precise conditions necessary for these instruments to be valid.

In general, leads can complement lags as instruments for expectations in the forward-looking regression. There are situations, however, in which lags or certain leads are invalid. For example, if agents do not update their forecasts each period, we shall show that lags are no longer valid. Conversely, if the Euler equation implied by the exponential discounting model includes lags of the dependent variable, we shall show that leads of the outcome variable are invalid instruments. In either scenario, however, leads (but not lags) of an exogenous treatment variable remain valid instruments.

Finally, we explore the practical implications of the foregoing analysis by examining the effect of certain tort reforms on equilibrium physician supply. The specific tort reforms we examine are selected based on factors that suggest they are unlikely to be endogenous, but may have been anticipated. Notwithstanding this screening, we first interpret pre-period trends in physician supply, such as those in Figure 1, as evidence of endogeneity and account for it by including state-specific trends in our estimation equation, much like what was done in Klick and Stratmann (2007). Then, we interpret pre-period trends as evidence of anticipation effects and estimate the different regression models discussed in our framework.⁹ We show in the main text that caps on punitive damages have a positive treatment effect on physician supply of 1.2 percent if we treat pre-period trends as evidence of endogeneity. Conversely, if we treat pre-trends as anticipation effects, caps have a 2.2 percent effect at the time they are adopted and a 5.0 to 6.4 percent effect in the long run.¹⁰ These findings

⁹Prior literature in this area has ignored anticipation effects entirely (e.g., Kessler, Sage, and Becker 2005, Klick and Stratmann 2007, Matsa 2007).

¹⁰In the Appendix we show that curbs on joint and several liability (an increase in physician liability) and split recovery rules (a reduction in liability) have an effect on physician supply of

suggest that accounting for anticipation effects increases the estimated impact of tort reform by a factor of four compared to estimates that control for endogeneity with state-specific trends.

The following is an outline of the remainder of the paper. Section 1 reviews the parameters of interest in a forward-looking regression. Section 2 elaborates on the various parametric restrictions that may be employed to reduce the number of expectation terms in the forward-looking regression. Section 3 discusses how to estimate the Euler equation (2) for a given model of belief formation. It examines the instruments that can be employed to address endogeneity from forecast errors, including leads of endogenous variables. Section 4 applies the different approaches to estimating the forward-looking model using data on tort liability and physician supply. Finally, Section 5 concludes with suggestions for future research.¹¹

1 Parameters of interest

Before estimating a model of anticipation effects such as

$$y_t = \lambda_0 d_t + \sum_{j=1}^{\infty} \lambda_j E_t [d_{t+j}] + e_t$$

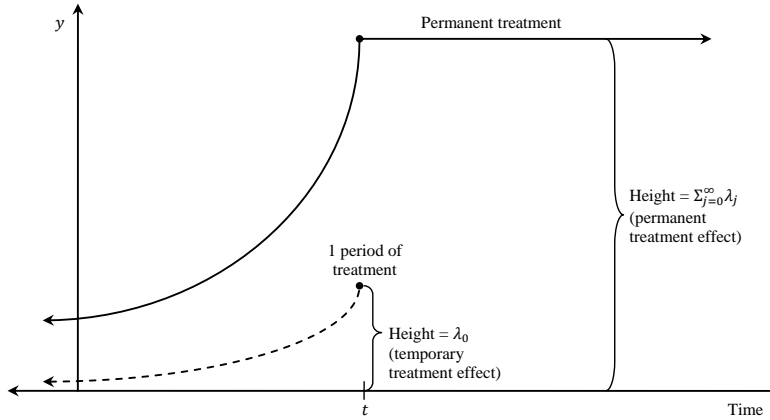
it is useful to define the possible parameters of interest from a policy evaluation framework.¹² The first parameter of interest is λ_0 , which is the effect of one period of treatment at time t on the outcome y_t . This measure ignores anticipation effects prior to treatment. Thus, the baseline for this change is not the outcome at time $t - 1$ but rather outcomes in the infinite past or, more practically, before agents anticipated the adoption of treatment. Following Becker, Grossman, and Murphy (1994) we shall

0.2 percent and 0.7 percent, respectively, when we equate pre-trends with endogeneity. If we treat pre-trends as anticipation effects, these reforms have a -1.3 to -1.5 percent and 1.3 to 1.5 percent effect, respectively, at the time they are adopted and an effect of -4.0 to -6.7 percent and 4.2 to 6.1 percent, respectively, in the long run.

¹¹The Appendix takes up topics that complement the discussion in the main text. Whereas the main text focuses on models with rational expectations, the Appendix takes up models with adaptive expectations. The Appendix also considers problems that arise when treatment variables are binary. Finally, whereas Section 4 presents results for the one reform depicted in Figure 1 (punitive damage caps), the Appendix takes up two other reforms where anticipation effects are likely (joint and several liability reform and split recovery rules).

¹²It is easy to add *ex post* adjustment costs to the forward-looking model. Because we are interesting in *ex ante* changes in behavior, we will without loss of generality ignore all time-varying *ex post* treatment effects.

Figure 2: Potential parameters of interest



call this the **temporary treatment effect**.

The second parameter of interest is $\sum_{j=0}^{\infty} \lambda_j$. One can interpret this as the effect on time- t outcomes of a permanent treatment adopted at time t .¹³ This includes the effect on current outcomes of the current period of treatment plus the anticipation effects on *current outcomes* of the *future periods of treatment*. The baseline again is outcomes before any anticipation effects. Hamilton (1994) (p. 7) calls the second parameter of interest the “long-run effect on y of a permanent change in d .” Following Becker, Grossman, and Murphy (1994), however, we shall refer to the second parameter as simply the **permanent treatment effect**.

The two parameters of interest are illustrated in Figure 2. The dotted line illustrates how the levels of an outcome y change in response to adoption of a temporary, one-period treatment at time t that was perfectly anticipated. The level of y at date $t - j$ is equal to the coefficient λ_j in our forward-looking model. The level of y on date t , when the one-period treatment is actually given, is equal to the first parameter, λ_0 . The solid line illustrates how y would respond if instead a permanent treatment were adopted at time t and that treatment was perfectly anticipated. The anticipation

¹³An alternative interpretation of the second parameter is the effect of one period of treatment at time t on outcomes in time t plus the sum of the effect on outcomes in all prior periods assuming agents have always known treatment would start at time t . This includes both the effect on current outcomes of current treatment and all the anticipation effects on all *past outcomes* of the *current period of treatment*. Hamilton (1994) (p. 7) calls this the “cumulative effect on y of a transitory change in d .” This interpretation corresponds to the area under the dotted line over the interval $(-\infty, t]$ in Figure 2.

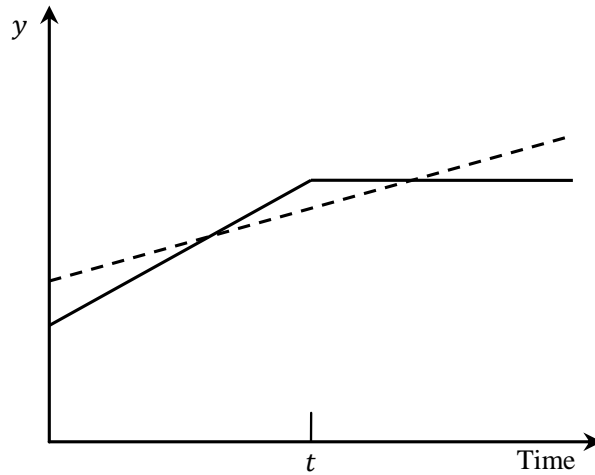
effects are larger with permanent treatment because each pre-period outcome reflects not just the anticipation of treatment in period t , but also anticipation of treatment in period $t + 1$, $t + 2$, etc. Specifically, the level of y at time $t - k$ is equal to $\sum_{j=k}^{\infty} \lambda_j$. An implication is that the level of y in every post-period corresponds to the second interpretation of second parameter of interest.¹⁴

When researchers see outcomes following the pattern depicted in Figure 2, they typically infer that treatment is endogenous and insert a trend to control for endogeneity. If the pattern is instead generated by anticipation effects, as we have shown is possible, the trend will bias estimates of treatment response. We illustrate this in Figure 3, which crudely reproduces anticipation effects from a permanent treatment adopted at time- t . It then plots a dotted trend for treatment groups. It is easy to see that the trend will not only reduce the estimated level treatment effect, but it may also cause the estimated level effect to be negative rather than positive.¹⁵ The use of a pre-period only trend as a control would prevent a sign reversal but will still underestimate anticipation effects because it will credit those effects to the pre-trend. The lesson is that insertion of trends as a control implicitly assumes outcomes are not at all driven by anticipation effects. A better approach is to theoretically justify the presence of endogeneity or anticipation effects. One should exclusively include trends or pre-trends only when endogeneity is present. Likewise, one should exclusively estimate a model such as (1) only when there are anticipation effects but no endogeneity. When a research cannot rule out either endogeneity or anticipation effects, the best that is possible is to get a sense of the range of treatment effects from estimates with trends and estimates from a model such as (1). In our application we attempt to rule out endogeneity, but to be careful we estimate regressions with and without trends in order to quantify the sensitivity of the results to our assumptions.

¹⁴The figure also suggests that permanent treatment adopted at time t not only raises outcomes in each post-period by what we call the permanent effect of treatment, but it also has an effect of $\sum_{k=1}^{\infty} \sum_{j=k}^{\infty} \lambda_j$ across all pre-periods.

¹⁵Wolfers (2006) makes the general point that insertion of trends can lead to biased estimates of treatment response when that response is dynamic.

Figure 3: Inserting trends can bias estimation



The insertion of trends can cause researchers to estimate the wrong sign of the treatment effect in dynamic models.

2 Simplifying the forward-looking model

The primary challenges with estimating a forward-looking model of the form

$$y_t = \lambda_0 d_t + \sum_{j=1}^{\infty} \lambda_j E_t [d_{t+j}] + e_t \quad (3)$$

are the infinite number of expectation terms and their unobservability. Here we discuss three different ways to solve these problems.

First, a researcher might completely ignore anticipation effects. Unfortunately, this approach suffers from the omitted variable bias we describe in Section 2.1. Second, a researcher might estimate a quasi-myopic model that includes only a finite number (S) of anticipation terms. If individuals anticipate a treatment more than S periods ahead, however, the model will also suffer from omitted variable bias. Moreover, the model contains S unobserved expectation terms. We explore these issues in Section 2.2.

Third, a researcher could adopt an exponential discounting model that assumes outcomes are a function of exponentially discounted expectations about treatment. Exponential discounting has the useful feature that suitable differencing can eliminate nearly all anticipation terms. The resulting Euler equation allows the researcher to identify anticipation effects at the cost of only a single degree of freedom. We elaborate on this in Section 2.3.

2.1 Myopic model

The simplest approach to dealing with anticipation effects is to ignore them and estimate a myopic model such as

$$y_t = \beta_0^{myopic} d_t + u_t$$

where y_t is the outcome of interest and d_t is the treatment. The omission of anticipation effects generates omitted variable bias. The specific nature of the bias depends on which parameter of interest (i.e., temporary or permanent treatment effect) from the previous section the researcher seeks to estimate.

If anticipation effects have the same sign as temporary effects, the estimated coefficient $\hat{\beta}_0^{myopic}$ is probably larger (in absolute value) than the temporary effect of treatment (λ_0) in the forward-looking model (3). The reason is that current treatment and expected future treatment are surely positively correlated: $Corr(d_t, E_t[d_{t+j}]) > 0$.¹⁶ In this case,

$$\text{plim } |\hat{\beta}_0^{myopic}| = |\lambda_0| + \sum_{j=1}^{\infty} |\lambda_j| \alpha_j > |\lambda_0|$$

where α_j is the coefficient from a regression of $E_t[d_{t+j}]$ on d_t . Intuitively, the coefficient on current treatment in the myopic model captures some of the effect of future treatment.¹⁷

Conversely, the myopic coefficient estimate is typically smaller (in absolute value) than the permanent effect of treatment in the forward-looking model since $\alpha_j \leq 1$, so that $\sum_{j=0}^{\infty} |\lambda_j| \alpha_j < \sum_{j=0}^{\infty} |\lambda_j|$. Intuitively, the coefficient in the myopic model captures the effect of permanent treatment if the current state of treatment perfectly predicts all future expected states of treatment. This is obviously not the case in periods

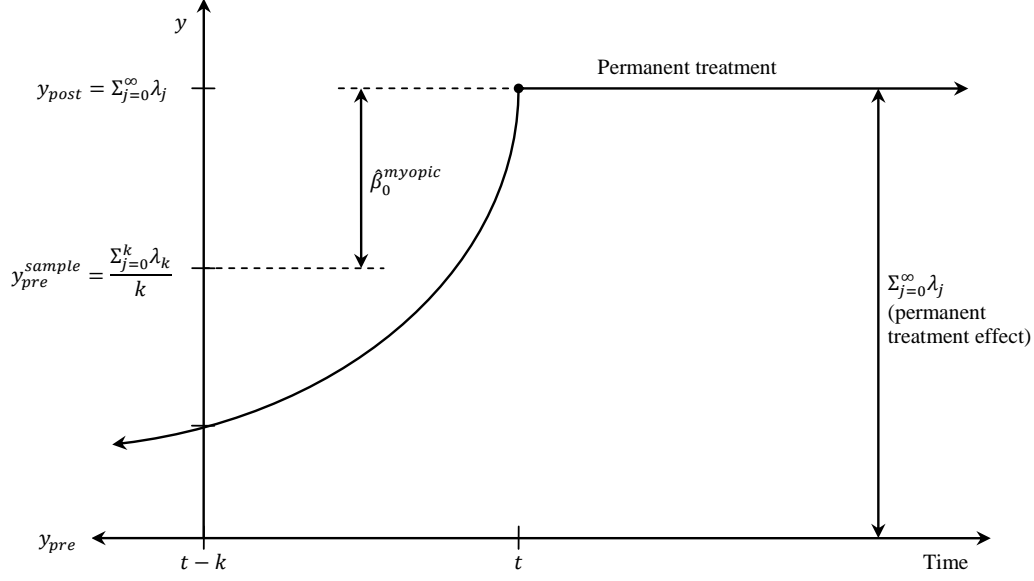
¹⁶Negative correlation between current treatment and expected future treatment implies that subjects frequently alternate between treated and untreated states. It is difficult to come up with examples of such treatments. Zero correlation is possible, but rules out infrequent treatment or treatment that lasts multiple periods.

¹⁷This result is not fully general. If the myopic model is estimated with fixed effects, $|\hat{\beta}_0^{myopic}|$ may dip below $|\lambda_0|$. Fixed effects estimation is equivalent to $y_{it} - \bar{y}_i = \beta_0 (d_{it} - \bar{d}_i) + (u_{it} - \bar{u}_i)$. Thus

$$\text{plim } |\hat{\beta}_0^{myopic}| = |\lambda_0| + \sum_{j=1}^{\infty} |\lambda_j| \frac{Cov[E_t[d_{it+j}] - \bar{d}_i, d_{it} - \bar{d}_i]}{Var[d_{it} - \bar{d}_i]}$$

This may be lower than $|\lambda_0|$ since, e.g., $-Cov[d_{it}, \bar{d}_i] < 0$. The larger is the timespan T of the data, the greater is the probability that $|\hat{\beta}_0^{myopic}| > |\lambda_0|$ since $Cov[d_{it}, \bar{d}_i]$ falls with larger T .

Figure 4: Estimate from a myopic model



before an agent is treated. Thus the estimate from the myopic model underestimates the permanent effect of treatment.

This point is illustrated in Figure 4, which plots the outcome of a forward-looking process after adoption of a permanent treatment at time t . Assume that the true permanent effect of the intervention is to increase outcomes by $\Sigma_{j=0}^{\infty} \lambda_j = y_{post} - y_{pre}$. Estimation of a myopic model, however, yields a treatment effect $\hat{\beta}_0^{myopic}$ that is the difference between the average outcome y_{pre}^{sample} before the law is passed and the average outcome y_{post} after the law is passed. The myopic estimate is less than the true permanent effect because the researcher observes a finite number of pre-treatment periods, say $[t-k, t]$, but expectations may have begun shifting outcomes well before $t-k$. Therefore the average pre-treatment outcome y_{pre}^{sample} in the sample is greater than the true pre-treatment outcome y_{pre} . Another way to put this is that the researcher has assigned some periods that belong in the post-period (because expectations are operating) to the pre-period, and thus overestimated outcomes in the pre-period.

2.2 Quasi-myopic model

To address the shortcomings of the myopic model, a researcher might estimate a quasi-myopic model that assumes agents have anticipation effects, but only for a

finite number of periods S :

$$y_t = \beta_0^{quasi} d_t + \sum_{j=1}^S \beta_j^{quasi} E_t[d_{t+j}] + u_t \quad (4)$$

This addresses the dimensionality problem in the general forward-looking model by ignoring anticipation terms after S periods, perhaps on the theory that agents do not forecast past S periods or that anticipation effects past S years have negligible effects.

Before we derive this equation, we pause to address the sentiment that the quasi-myopic model is preferable to the exponential discounting model because it is in some sense less parametric. Estimation of anticipation effects requires both a mapping from expectation to outcomes such as (3) and, if expectations are not directly observable, a mapping from observables to expectations. The quasi-myopic model is non-parametric on neither dimension. First, it requires assuming that there are no more than S periods of anticipation effects. One cannot determine, *a priori*, whether this parametric restriction or those embodied in the exponentially discounting model yield lower bias. Second, the quasi-myopic model, like the exponential discounting model, requires making parametric assumptions about the relationship between observables and expectations because expectations are generally not directly observable. Otherwise the quasi-myopic model will provide, at best, a non-parametric estimate of the treatment effects of a variable that is measured with error.¹⁸

In practice, the quasi-myopic model has a second shortcoming: the S periods of expectations are unobserved. Researchers frequently address this problem by substituting realizations of d_{t+j} for expectations of those variables, thereby implicitly making a rational expectations assumption. This does not solve the problem so much as transform it: estimation of the quasi-myopic model will be biased if the unobserved forecast error is correlated with treatment. In that case, the quasi-myopic model can still be estimated with instrumental variables to generate consistent estimates of an-

¹⁸It is sometimes possible to study a shock exclusively to expectations. For example, a policy change that is announced before it is implemented may be a good instrument for a change in expectations around the announcement date. (Alpert 2010, Gruber and Koszegi 2001, Lueck and Michael 2003, Blundell, Francesconi, and Van der Klaauw 2010) Such a shock does not, however, permit identification of the treatment effect from a change in expectations because one cannot observe expectations and thus estimate the slope between the shock and expectations. In other words, one can only generate reduced form estimates because one cannot estimate a first stage regression of expectations on the instrument.

anticipation effects. However, it requires at least S instruments for the S periods of anticipation effects the researcher seeks to estimate. As we show in the next section, the exponential discounting model allows one to derive an estimable Euler equation with just one unobserved expectation term. Thus the researcher will need only one instrument.

2.3 Exponential discounting model

The third approach to reducing the dimensionality of the forward-looking model is to assume that treatment has a constant temporary treatment effect of $\lambda_0 = \beta$ and an anticipation effect j periods prior to treatment of $\lambda_j = \beta\theta^j$:

$$y_t = \beta d_t + \beta \sum_{j=1}^{\infty} \theta^j E_t [d_{t+j}] + e_t \quad (5)$$

The permanent effect of treatment is estimated as $\hat{\beta}/(1 - \hat{\theta})$. The central benefit of the assumption that outcomes are a function of exponentially discounted expectations about treatment is that subtracting θy_{t+1} from (5) will enable the researcher to generate an Euler equation with only one expectation term.

Before we derive this equation, we pause to note that one cannot, *a priori*, determine whether the parametric restrictions in the quasi-myopic model or those embodied in exponentially discounted model yield lower bias. If there are more than S periods of anticipation effects, then the quasi-myopic model suffers omitted variable bias. But exponential discounting may also be a poor approximation to the time path of anticipation effects and suffer misspecification bias. It is uncertain which bias is larger.

The precise Euler equation that corresponds to a forward-looking model with exponentially discounted expectations depends on how agents are assumed to update their expectations. The natural response – common among macroeconomic and finance econometricians – is to assume a model of belief formation, such as rational or adaptive expectations, in order to substitute observable variables for unobservable expectations of a variable. Below we derive Euler equations under the assumption that agents have rational expectations, while we take up adaptive expectations in Appendix A.1.

In general, when modeling rational expectations, one must specify exactly what the objects of expectations are and how expectations relate to observed realizations.

There are two possible objects of rational expectations, treatment or outcomes. Moreover, expectations may either depend on realizations ($E[z] = z + v$) or vice versa ($z = E[z] + v$). Economic theory should dictate which path to take. For our application, we have chosen treatment as the object of expectations and have expectations depend on realizations. We consider alternative formulations of the rational expectations assumption in Appendix A.2.

We formulate the case where agents form rational expectations about treatments and these expectations are a function of actual treatments as follows: $E_t[d_{t+j}] = d_{t+j} + v_{t,t+j}^d$ where $E[d_{t+j}v_{t,t+j}^d] = 0$. The term $v_{t,t+j}^d$ is the forecast error resulting from an agent's time- t forecast of the treatment d_{t+j} . This model is appropriate when treatments are exogenously assigned, as we demonstrate is the case for our application (see Section 4.3.2). In this case we can substitute the rational expectations assumption directly into the basic forward-looking model to obtain

$$y_t = \beta \sum_{j=0}^{\infty} \theta^j d_{t+j} + e_t + \beta \sum_{j=1}^{\infty} \theta^j v_{t,t+j}^d$$

(Since d_t is known at time t , $v_{t,t}^d = 0$.) After performing the same substitution to expand y_{t+1} , subtracting θy_{t+1} from y_t yields

$$y_t = \theta y_{t+1} + \beta d_t + w_t \tag{6}$$

where

$$\begin{aligned} w_t &= e_t - \theta e_{t+1} + \beta \sum_{j=1}^{\infty} \theta^j v_{t,t+j}^d - \beta \sum_{j=2}^{\infty} \theta^j v_{t+1,t+j}^d \\ &= [e_t - \theta e_{t+1}] + \beta \theta v_{t,t+1}^d + \beta \sum_{j=2}^{\infty} \theta^j [v_{t,t+j}^d - v_{t+1,t+j}^d] \end{aligned}$$

The error term w_t has three components. One is the change in model error, $e_t - \theta e_{t+1}$. A second is the error in forecasting time- $t + 1$ treatment at time t . The third component is the change in forecasts about time $t + j$ treatment ($j > 1$) from time t to time $t + 1$. There is, however, only one source of endogeneity between outcome y_{t+1} and the error term: the model error e_{t+1} .¹⁹

¹⁹There is no endogeneity from $v_{t,t+1}^d$ because, although y_{t+1} is a function of d_{t+1} , we have assumed that d_{t+1} is orthogonal to $v_{t,t+1}^d$. Nor is there endogeneity from the change in forecasts ($v_{t,t+j}^d - v_{t+1,t+j}^d, j > 1$) because under rational expectations these forecast updates are orthogonal to prior forecast errors ($v_{t,t+j}^d$) and thus orthogonal to $E_t[d_{t+j}]$ too. Indeed, there is no additional endogeneity even if $\{d_t\}$ are serially correlated because $E[d_{t+j}v_{t,t+j}^d] = 0$ by assumption.

3 Estimation

This section takes up estimation of anticipation effects models. The focus will be on estimating the exponential discounting model, though the section will contain lessons for the quasi-myopic model as well. The Euler equation (6) derived in the previous section greatly reduced the dimensionality of our forward-looking model (1), but consistent estimation still requires the researcher to account for the correlation between y_{t+1} and the model error e_{t+1} contained in the error term w_t . One solution is to find an instrument.

The usual source for these instruments is a subset of the agent's information set, for instance, lags of the endogenous variable (see McCallum 1976). This is typically motivated by modeling the agents' expectations as a linear projection of the variables in the agents' data sets, which include lagged values of the endogenous variable. An alternative motivation is to note that since lags of y_{t+1} and y_{t+1} all depend, according to the forward-looking model (1), on expectations about future treatment, shocks to lags of y_{t+1} can also move y_{t+1} . The exclusion restriction is completed by noting that the current period outcome y_t in equation (1) does not depend on lagged values of the endogenous variable.²⁰

This alternative motivation for using lags as instruments also suggests a new set of instruments we propose here: leads of the endogenous variable. Like y_{t+1} , $\{y_{t+k}\}$ for $k > 1$ depend, according to the forward-looking model (1), on expectations about future treatment. Thus they are correlated with $E_t[y_{t+1}]$ and are plausible instruments. Analogizing to models from the dynamic panel literature, we argue that, although y_{t+1} may be endogenous, leads of y_{t+1} are not (i.e., $E[y_{t+j}w_t] = 0$ for $j > 1$) provided that certain conditions on the correlation structure of w_t are met. The intuition for the exclusion restriction is that the Euler equation demonstrates that y_{t+1} fully captures the influence of future treatments in the forward-looking regression. The current outcome y_t is related to future treatments, and thus future outcome variables, only through y_{t+1} .

One could alternatively instrument for y_{t+1} using leads of d_{t+1} rather than leads of y_{t+1} since, after all, shocks to future expectations of d_{t+1} are what ultimately drive identification. Whether this is more efficient than using leads of y_{t+1} depends on the

²⁰This assumes there is no detrimental serial correlation in the error term. We discuss the precise conditions required for this below.

variance-covariance structure of model error e_t and forecast error v_t , which is unknown *a priori*. However, there is a strong practical reason to prefer leads of y_{t+1} : they embed more information than leads of d_{t+1} for any finite data set. For example, consider a panel with 10 time periods. Instrumenting for y_9 with y_{10} necessarily includes information about $\{d_{11}, d_{12}, \dots\}$ because y_{10} is a function of future treatments. That information is unavailable when using leads of d_9 , however, because the data set only contains 10 time periods. This advantage is reduced if data on the future treatments $\{d_{11}, d_{12}, \dots\}$, but not $\{y_{11}, y_{12}, \dots\}$, are available. There are some scenarios, however, where using leads of d_t is preferable to using leads of y_{t+1} . We discuss them at the end of this section.

The bulk of this section carefully describes the conditions required for consistent estimation of our forward-looking model.²¹ We will work with a panel form of our Euler equation that differs slightly from the Euler equation (6) and the standard autoregressive error components model in the dynamic panel literature. This panel form will introduce controls x_{it} and fixed effects η_i . This is not only more general than equation (6), but also allows us to exploit the similarity between our forward-looking Euler equation ($y_{it} = \theta y_{it+1} + \beta d_t + \eta_i + w_{it}$) and the well-known autoregressive error components model ($y_{it} = \theta y_{it-1} + \beta d_t + \eta_i + e_{it}$). Note, however, that our forward-looking Euler equation is not, strictly speaking, an autoregressive model. (This is easily seen by noting that equation (1), from which our Euler equation was derived, does not include a lagged dependent variable.) One implication of this is that while only lags of the outcome variable are valid instruments in the autoregressive error components model, both lags and leads are valid instruments in our forward-looking model. We elaborate below.

3.1 Instrumenting with lags and leads

We are interested in estimating a model of the form

$$y_{it} = \theta y_{i,t+1} + \alpha_1 x_{it} + \alpha_2 x_{i,t+1} + \beta d_{it} + \eta_i + w_{it} \quad (7)$$

where $i = 1 \dots N$, $t = 1 \dots T$, and $0 < \theta < 1$. We assume that η_i and w_{it} are independently distributed across i with $E[\eta_i] = E[w_{it}] = E[\eta_i w_{it}] = 0$. The number of time

²¹Monte Carlo simulations demonstrating the consistency of our estimator are available upon request.

periods T is fixed and the number of individuals N is large. We assume for simplicity that x_{it} and $x_{i,t+1}$ are strictly exogenous and known in advance.²² Although we shall focus here on the validity of instrumenting with leads, it is easy to adapt our argument to show that lags are also valid.²³

Direct OLS estimation of equation (7) is inconsistent because $E[y_{i,t+1}\eta_i] \neq 0$. Estimating first differences (defined here as $\Delta y_{it} = y_{it} - y_{i,t+1}$) fails because $E[\Delta y_{i,t+1}\Delta w_{it}] \neq 0$. (A within estimator suffers from this same problem, although the bias may disappear as $T \rightarrow \infty$.) The logic in Arellano and Bond (1991), Arellano and Bover (1995) and Blundell and Bond (1998) suggests that one solution is to instrument for $\Delta y_{i,t+1}$ using leads of $y_{i,t+1}$ or to instrument for $y_{i,t+1}$ using leads of $\Delta y_{i,t+1}$. Leads are valid instruments if the following standard assumptions are met:

$$\mathbf{A1:} \quad E[y_{iT}w_{it}] = 0 \quad \forall i, \quad \forall t \leq T - 1$$

$$\mathbf{A2:} \quad E[w_{is}w_{it}] = 0 \quad \forall t \neq s$$

$$\mathbf{A3:} \quad E[\eta_i\Delta w_{i2}] = 0 \quad \forall i$$

Assumption A1 requires y_{iT} to be uncorrelated with past disturbances. Assumption A2 requires these disturbances to be uncorrelated. These two assumptions together imply the following moment conditions:

$$E[y_{i,t+j}\Delta w_{it}] = 0 \quad \forall j \geq 2, \quad \forall t \tag{8}$$

Assumption A3 requires the terminal conditions to be mean stationary. In other words, conditional on the covariates x_{it} , individuals with large random effects η_i must not be systematically closer or farther away from their steady states than individuals with small random effects, so that the terminal conditions are representative of the steady state behavior of the model. If it holds, A3 implies the following additional (non-redundant) moment conditions:

$$E[\Delta y_{i,t+1}w_{it}] = 0 \quad \forall t \tag{9}$$

Equation (7) is overidentified if $T > 3$ but can be estimated using the Generalized Method of Moments (GMM) framework developed by Hansen (1982). “Difference

²²These assumptions can be relaxed.

²³The validity of lags has been shown in earlier papers, e.g., McCallum (1976).

GMM” estimation exploits the moment conditions (8) while “system GMM” estimation exploits both (8) and (9).

Assumption A2 is central to the validity of these estimation procedures. As currently stated, however, it is actually stronger than necessary. Limited serial correlation of order $H > 0$ is acceptable so long as the researcher takes care to omit the affected instruments and enough instruments remain for identification. We therefore loosen A2:

$$\mathbf{A2}': E[w_{it}w_{it+j}] = 0 \quad \forall j > H, H \geq 0$$

This changes our moment conditions (8) and (9) to

$$E[y_{i,t+j}\Delta w_{it}] = 0 \quad \forall j \geq H + 2, \quad \forall t$$

$$E[\Delta y_{i,t+H+1}w_{it}] = 0 \quad \forall t$$

Whether or not these assumptions are satisfied depends on the content of the error term in the Euler equation, which in turn depends on how expectations are specified. Following Section 2.3, we consider the case where expectations are a function of treatments: $E_t[d_{t+j}] = d_{t+j} + v_{t,t+j}^d$. (We now drop the i subscript for the remainder of this section for notational convenience.) The error term for this case, which was derived in equation (6), is

$$w_t = e_t - \theta e_{t+1} + \beta \theta v_{t,t+1}^d + \beta (\sum_{j=2}^{\infty} \theta^j [v_{t,t+j}^d - v_{t+1,t+j}^d])$$

Because w_t follows an $MA(1)$ process, it is clear that adjacent outcomes cannot instrument for each other (e.g., y_{t+2} is not a valid instrument for y_{t+1}). More generally, we are interested in knowing under what conditions Assumption A2' is satisfied. It holds if the following four conditions are satisfied for all periods t and individuals i for some $H \geq 1$:

1. $E[e_t e_{t+j}] = 0 \quad \forall j > H$
2. $E[e_t v_{t+j,t+k}^d] = 0, \quad \forall k > j, \quad \forall j > H$
3. $E[(v_{t,t+k}^d - v_{t+1,t+k}^d)v_{t+j,t+j+1}^d] = 0 \quad \forall k > 1, \quad \forall j > H$
4. $E[(v_{t,t+k}^d - v_{t+1,t+k}^d)(v_{t+j,t+m}^d - v_{t+j+1,t+m}^d)] = 0 \quad \forall k > 1, \quad m > j + 1, \quad j > H$

In words, condition 1 means autocorrelation in e_t cannot be higher than order H . Condition 2 means the model error is orthogonal to the H -step-ahead-and-beyond forecast error. Condition 3 means the *change* in a forecast from period t to period $t + 1$ is uncorrelated with the *level* of a forecast in period $t + j$. Condition 4 means independent information is used to update the forecast each period.

Conditions 1, 2 and 4 are plausible in many scenarios, but condition 3 may be an unrealistic assumption. It holds in the cases of perfect serial correlation (so the change in forecast $(v_{t,t+k}^d - v_{t+1,t+k}^d) = 0$) or no serial correlation (so $E[v_{t+j,t+j+1}^d v_{t+l,t+k}^d] = 0 \forall j, k, l$). These two extremes are not satisfied in most applications. Rational expectations, however, offers some hope. It implies that the (perhaps nonzero) expectation in Condition 3 is not a function of t . In other words, an agent's forecast error might depend on whether she is predicting an event three time periods in the future versus four time periods in the future, but it does not depend on the particular time period she is forecasting from.

This means we can rewrite our moment conditions (8) and (9) as

$$\begin{aligned} E[y_{t+j}\Delta e_t - k_1(j; \beta, \theta)] &= 0 \\ E[\Delta y_{t+j}e_t - k_2(j; \beta, \theta)] &= 0 \end{aligned}$$

where $k_1(\cdot)$ and $k_2(\cdot)$ are constants that do not depend on t or our data (x, y) . They will thus be absorbed into our constant term, but the researcher can still identify the parameters of interest, β and θ .²⁴ Unfortunately, this means we cannot include multiple instruments: because the moment conditions are a function of j , the non-zero moment condition differs for each instrument. The optimal solution is for the researcher to specify each instrument as a separate GMM equation and then estimate the entire system simultaneously.

We have shown theoretically why leads and lags of the endogenous variable can be used as instruments, but this does not guarantee that these instruments are strong. Roodman (2009b) documents the danger of using weak instruments, especially when

²⁴It may appear surprising that we can effectively ignore $k_1(\cdot)$ and $k_2(\cdot)$ even though they are functions of our parameters. We are able to do this because $k_1(\cdot)$ and $k_2(\cdot)$ are constants and thus merely represent level shifts of the GMM minimization problem. Consider the analogous problem for OLS: $\min_{\beta_0, \beta_1} (y - \beta_1 x_1 - \beta_0 - k(\beta_1))^2$ where $k(\beta_1)$ is some constant that is a function of β_1 and independent of x_1 . Identification of β_0 is impossible but OLS can still identify β_1 even without knowing the functional form of $k(\beta_1)$.

they are numerous. The researcher should take care to perform Hansen, Difference-in-Sargan, and Arellano-Bond autocorrelation tests to validate her model’s assumptions.

Finally, we note that estimation is further complicated when the treatment variable is both binary and serially correlated, as will often be the case. We discuss how to handle this in Appendix A.3.

3.2 Comparing lag and lead instruments

The discussion above showed that lags and leads of the outcome variable are both valid instruments under our standard forward-looking specification when agents have rational expectations. We also noted that the researcher could alternatively use leads (but not lags) of the treatment variable as instruments. Here we show that the possibility of instrumenting with leads of the treatment variable provides a flexibility that makes leads generally better instruments than lags. We demonstrate this by considering two common situations that depart from our standard specification.

For instance, if agents do not continuously update their forecasts of future variables with new, orthogonal information, lags of the outcome variable are no longer valid instruments. To illustrate why, we examine the case where agents have rational (i.e., unbiased) forecasts of treatment but never update these forecasts.²⁵ This implies the forecast error no longer depends on the date the forecast was made, so that $E_t[d_{t+j}] = d_{t+j} + v_{t,t+j}^d = d_{t+j} + v_{t+j}^d$. The exponential discounting model may now be written as

$$y_t = \beta \sum_{j=0}^{\infty} \theta^j (d_{t+j} + v_{t+j}^d) + e_t$$

Subtracting y_{t+1} yields the Euler equation

$$y_t = \theta y_{t+1} + \beta d_t + \beta \theta v_{t+1}^d + e_t - \theta e_{t+1}$$

Lags are necessarily invalid instruments in this specification because y_{t-j} for any $j > 1$ is correlated with v_{t+1}^d . Although lags are no longer orthogonal to the error term in the Euler equation, leads remain valid instruments.

Conversely, if the researcher derives an Euler condition that includes lagged de-

²⁵Carroll (2003) provides evidence that household expectations are not rational, but are based on professional forecasts, which may be rational. Importantly, he finds that households only occasionally update their expectations and are therefore “sticky” in the aggregate. Anderson, Kellogg, and Sallee (2011) find that consumer forecasts of gasoline prices are indistinguishable from a no-change forecast.

pendent variables, e.g.,

$$y_t = \theta y_{t+1} + \gamma y_{t-1} + \beta d_t + w_t$$

she cannot use leads of y_{t+1} as instruments for y_{t+1} .²⁶ Because of serial correlation in y_t , y_{t+j} for any $j > 1$ is correlated with y_t and thus w_t . Leads of the outcome variable are thus no longer orthogonal to the error term. The researcher could, however, use leads of the treatment variable, d_t , so long as treatment remains exogenous to the error term. Alternatively, lags of the outcome variable are also valid instruments.

4 Application: effect of tort reform on physician supply

In this section we estimate the effect of tort reform on physician supply using the different methods of estimating anticipation effects we have described. Section 4.1 begins by providing some background on medical malpractice liability. It then presents a model that explains why current physician supply depends on expectations about future tort liability rules. The model explains what the coefficients on future tort rules actually measure. Section 4.2 describes the data we employ. Section 4.3 explains why one particular tort reform – caps on punitive damages – is a good candidate for studying anticipation effects. (It is not the only reform that is a reasonable candidate; we also examine joint and several liability and split recovery rules in the Appendix.) Specifically, we argue that the change in physician supply prior to adoption of these caps, i.e., a pre-period trend, is better interpreted as anticipation effects rather than as endogeneity. Finally, Sections 4.4 and 4.5 explain our empirical models and report our results. Ultimately, we shall compare estimates of the equilibrium supply effect of punitive damage caps from a myopic model, a quasi-myopic model, and an exponential discounting model using both leads and lags of the outcome variables as instruments.

4.1 Background on tort liability and theory

Tort liability is akin to a legally-mandated alteration of the implicit labor contract between a patient and her physician. In most cases, it requires the physician to pro-

²⁶Note that y_t here is a function of y_{t-1} but *not* a function of y_{t+1} . This is because w_t implicitly includes a $-\theta e_{t+1}$ term but not a $-\theta e_{t-1}$ term.

vide the quality of care that a “reasonable physician” would provide and compensates patients who suffer injuries due to inadequate care. Compensation may include economic damages for lost wages and the cost of additional medical care; non-economic damages for pain and suffering from the injury; and punitive damages intended to punish the doctor for outrageous misconduct.

Tort reform refers to various changes to these mandatory contract terms. Table 2 provides a description of the most common reforms. Most of them, such as caps on punitive damages, lower the liability of doctors. Others, such as reform of joint and several liability, which reduces the extent to which hospitals share the liability of doctors, increase the overall liability of doctors (see Currie and MacLeod 2008).

Policymakers are concerned that tort liability is driving away doctors and thus reducing patients’ access to care. This claim has received substantial attention from scholars and the media.²⁷ Here we present a simple model of equilibrium physician supply that captures the economic intuition underpinning this claim and provides one motivation for employing the forward-looking regression model (1) in this market.²⁸

Consider a consumer choosing consumption to maximize the present value of her utility

$$E_t \left[\sum_{j=0}^{\infty} \theta_j U(c_{t+j}) \right] \quad (10)$$

where we allow the discount factors θ_j to vary arbitrarily over time. The consumer

²⁷See Born, Viscusi, and Baker (2006), Currie and MacLeod (2008), Helland, Klick, and Tabarrok (2005), Kessler, Sage, and Becker (2005), Klick and Stratmann (2007), Matsa (2007), and The Economist (2005).

²⁸We do not contend that economic theory implies tort liability must reduce physician supply. For example, transaction costs may prevent a patient and physician from writing a complete contract that covers all contingencies, including instances of malpractice. In that case, mandatory terms imposed by tort liability that improve the contract will increase the value of physician care. We can capture this by modifying (10) to reflect the surplus value v to consumers from physician liability:

$$E_t \sum_{j=0}^{\infty} \theta_j U \left(c_{t+j} + v \sum_{k=0}^j d_{t-k} y_{t-k} \right)$$

Leaving the remainder of the model unchanged, physician supply would still be a linear function of future tort liability. For example, assuming linear utility means equation (11) would take the form

$$y_t = s_0 + s_1 \sum_{j=0}^{\infty} \theta_j r_j + s_1 (v - \tau) \sum_{j=0}^{\infty} \theta_j E_t [d_{t+j}]$$

Prices and thus wages would rise as consumers would be willing to pay extra for the value provided to consumers. They would fall, as the analysis in the main text suggests, to the extent that physicians bore the cost of tort liability. The coefficient on expected future liability might be positive if the former effect exceeds the latter, i.e., $v > \tau$.

can spend her time- t assets A_t on either consumption c_t or health care y_t :

$$c_t + p_t y_t \leq A_t$$

The unit price of physician care is p_t . Health care consumption at time t has two effects on future asset accumulation. First, it functions like an investment that yields a return r_j (net of adverse events) in period $t + j$. For example, health care may increase productivity. Second, if treatment causes an adverse event that is discovered in some future period $t + j$, the tort liability system requires the physician to pay damages d_{t+j} that period to the consumer.²⁹ Thus assets at time t are

$$A_t = A_0 + \sum_{j=0}^t r_j y_{t-j} + d_t \sum_{j=0}^t y_{t-j}$$

The first order condition for the consumer's problem implies

$$p_t = \sum_{j=0}^{\infty} \theta_j \frac{U'(c_{t+j})}{U'(c_t)} r_j + \sum_{j=0}^{\infty} \theta_j \frac{U'(c_{t+j})}{U'(c_t)} E_t [d_{t+j}]$$

where we have assumed $\{r_j\}$ are known at time t , but future tort liability is not. If consumers exhibit either log utility with the ability to fully smooth consumption or linear utility,³⁰ then the price of physician care satisfies

$$p_t = \sum_{j=0}^{\infty} \theta_j r_j + \sum_{j=0}^{\infty} \theta_j E_t [d_{t+j}]$$

This price is not the same as the physician's wage because damages are a monetary transfer from physician to consumer. Damages do not necessarily affect the physician's wage rate, however, because they may be incorporated into the price. This is also true with medical malpractice insurance, which converts stochastic future tort payment into a steady stream of premiums (Baicker and Chandra 2006). Following Currie and MacLeod (2008) (p. 5), however, we assume there is a portion τ of tort liability that cannot be charged to consumers (e.g., if lawsuits create large overhead and psychic costs for physicians, or cause physicians to practice uncompensated

²⁹We assume d_{t+j} captures both the probability of detection and the magnitude of the potential tort damages.

³⁰Linear utility may be a reasonable approximation for consumers with health insurance, since they rarely have large out-of-pocket expenses.

“defensive medicine”). Thus wages will reflect tort liability:

$$w_t = p_t - (1 + \tau) \sum_{j=0}^{\infty} \theta_j d_{t+j} = \sum_{j=0}^{\infty} \theta_j r_j - \tau \sum_{j=0}^{\infty} \theta_j E_t [d_{t+j}]$$

To see how tort liability affects physician supply y_t , we specify a linear (or log-linear) physician supply curve:

$$y_t = s_0 + s_1 w_t = s_0 + s_1 \sum_{j=0}^{\infty} \theta_j r_j - s_1 \tau \sum_{j=0}^{\infty} \theta_j E_t [d_{t+j}] \quad (11)$$

This is essentially our forward-looking regression model (1). The coefficient on the expected liability $E_t [d_{t+j}]$ at time $t + j$ reflects the slope s_1 of the physician supply curve, the fraction τ of liability that is borne by physicians, and the discount factor θ_j .³¹

Several recent studies employ a myopic model instead of (11) to analyze the impact of tort liability on physician supply. Kessler, Sage, and Becker (2005) perform a difference-in-differences analysis and find evidence that reforms directly affecting how much a defendant has to pay increase physician supply by 3%. Matsa (2007) examines the effect of damage caps on physician supply and finds it increases the supply of physicians by about 10%, but only in rural counties. Klick and Stratmann (2007) employ a triple-differences model and estimate that caps on non-economic damages are associated with a 6% increase in physician supply for high-risk specialties.

4.2 Data

Our analysis uses annual physician count data from the American Medical Association’s Physician Masterfile.³² These counts include private practitioners, hospital staff, residents, locum tenens, but not military doctors.³³ Physicians are categorized into one of 20 possible specialties and have state identifiers. The data span the period

³¹One might object that physicians face large relocation costs that block their exit. However, the large inflow of new residents and the large potential outflow of retirees may lead to a relatively quick adjustment on the extensive margin despite high relocation costs (Kessler, Sage, and Becker 2005). In 1996, approximately five percent of the physicians in our sample were new residents (AMA 1997). Extrapolating this trend implies that more than one half of all practicing physicians entered the profession within the past 14 years.

³²This data is also analyzed in Klick and Stratmann (2007). We are grateful to the authors for sharing their code with us.

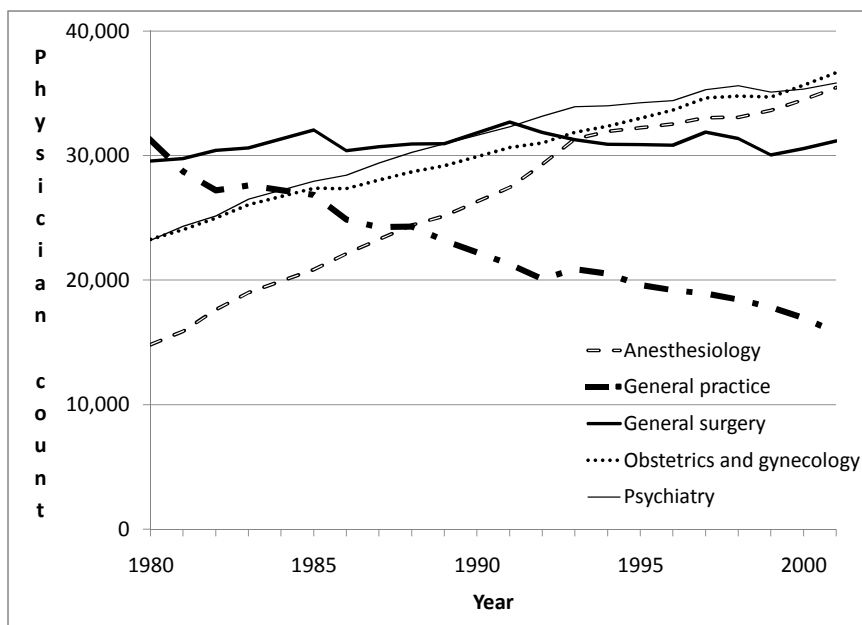
³³Locum tenens are temporary, substitute doctors employed by states when there is a shortfall of doctors.

1980-2001, with gaps in 1984 and 1990.

Klick and Stratmann (2007) note that some physician specialties are sued more often than others and correspondingly group them into four equally-sized risk tiers, displayed in Table 1. We use their definitions to limit our data and analysis to the two riskiest tiers (tiers 1 and 2) because we expect these to be more affected by tort liability than the other two tiers.

Figure 5 graphs the total counts over time of the five most populated specialties in our data set. The supply of general practitioners is declining over time, the supply of general surgeons is stagnant, and the rest are rising.

Figure 5: Physician supply from 1980 to 2001



Note: data for 1984 and 1990 are interpolated.

Our tort reform data come from Avraham (2010).³⁴ These data indicate, for the same time period as our physician supply data, whether ten different tort reforms are in effect at the state-year level. These reforms are defined in Table 2 and coded as 0-1 indicator variables.

³⁴Klick and Stratmann (2007) and Matsa (2007), by contrast, use tort reform data from the American Tort Reform Association (ATRA) to estimate the effect of tort reform on physician supply. Avraham (2010) corrects errors in the ATRA data set and includes data on three additional tort reforms: split recovery, punitive damage evidence, and caps on punitive damages.

4.3 Evidence against endogeneity and for anticipation effects

Among the set of tort reforms that may affect physician supply, we focus on punitive damage caps. This reform either imposes a specific dollar upper bound such as \$250,000 on punitive awards or requires that punitive awards be no more than a fraction or multiple of economic damages. We chose punitive damage caps because it is a good candidate for our model of anticipation effects. Specifically, it meets the following three criteria: (1) physician supply changes prior to enactment of the reform, i.e., there is a pre-period trend; (2) the reform is exogenous to physician supply; and (3) there is evidence that physicians could directly or indirectly anticipate the reform years prior to its enactment. We provide evidence for these three criteria below. In the Appendix, we take up two other tort reforms – joint and several liability reform and split recovery rules – that also meet these criteria.

Before we begin, we note it might surprise some readers that physician supply responds at all to punitive damages. It has been reported that punitive damages are awarded in only 1-4% of all medical malpractice trials (Cohen 2004, Cohen and Harbacek 2011). However, this figure underestimates the impact of punitive damages. First, according to Table 2, seventeen states have adopted caps on punitive damages. Since the 1-4% figure is a national figure, it reflects both states that allow punitive awards and states that cap or prohibit those awards. Punitive damages will play a larger role in states that do not cap those damages. Viscusi and Born (2005) estimate that medical malpractice insurers incur 6-7% lower losses in states with caps on punitive damages and 15% lower in states that ban any punitive awards. Second, even if punitive damages are a small percentage of awards, they may be the one aspect of tort damages that cannot be insured by physicians. Nearly half the states do not allow malpractice liability insurance to cover punitive damages (Wilson, Elser, Moskowitz, Edelman, and Dicker LLP 2008, McCullough, Campbell, and Lane LLP 2004). Moreover, even in states that allow liability insurance to cover punitive damages, many insurers refuse to do so. Viscusi and Born (2005) also estimate that medical malpractice insurers incur 6-7% lower losses in states that prohibit insurance coverage of punitive awards. Given that 98.8% of total malpractice awards are covered by liability insurance (Zeiler, Silver, Black, Hyman, and Sage 2007), this implies that punitive damages are a disproportionate source of risk to doctors, perhaps more than 83% ($= 6\% / (6\% + 1.2\%)$) of all financial risk they bear from medical malpractice liability. It is not surprising, therefore, that a number of papers that examine med-

ical malpractice tort reform have found a significant effect of punitive damages on physician behavior (e.g., Avraham 2007, Avraham, Dafny, and Schanzenbach 2010, and Currie and MacLeod 2008).

4.3.1 Physician supply changed prior to enactment of punitive damage caps

Figure 6 shows that six reforms in our data – including caps on punitive damage – exhibit a supply change prior to enactment of the reform, even after controlling for state-specialty and specialty-year fixed effects.³⁵ Some of these trends continue after the law was adopted, suggesting there may have been *ex post* adjustment, but a good portion of the change occurs before the reform is adopted. Importantly, the change in supply prior to adoption of caps on punitive damages is positive, which is consistent with physicians anticipating a liability-reducing reform.

4.3.2 Punitive damage caps are exogenous to physician supply

Although a change in outcomes prior to treatment is consistent with anticipation effects, it does not rule out the possibility of endogeneity. For example, changes in physician supply may induce changes in tort laws (reverse causality) or an unobserved set of factors may cause both changes in physician supply and tort reforms (omitted variable bias). We now present evidence ruling out these two possible explanations.

First, unlike other reforms, punitive damage caps are generally targeted at all tort suits, not just medical malpractice suits.³⁶ This is verified in Table 3, which lists the specific states that adopted different reforms. States that adopted reforms that were restricted in application to medical malpractice suits are listed in bold. Out of 17 states that adopted caps on punitive damages, only five states restricted the reform to medical malpractice cases.³⁷

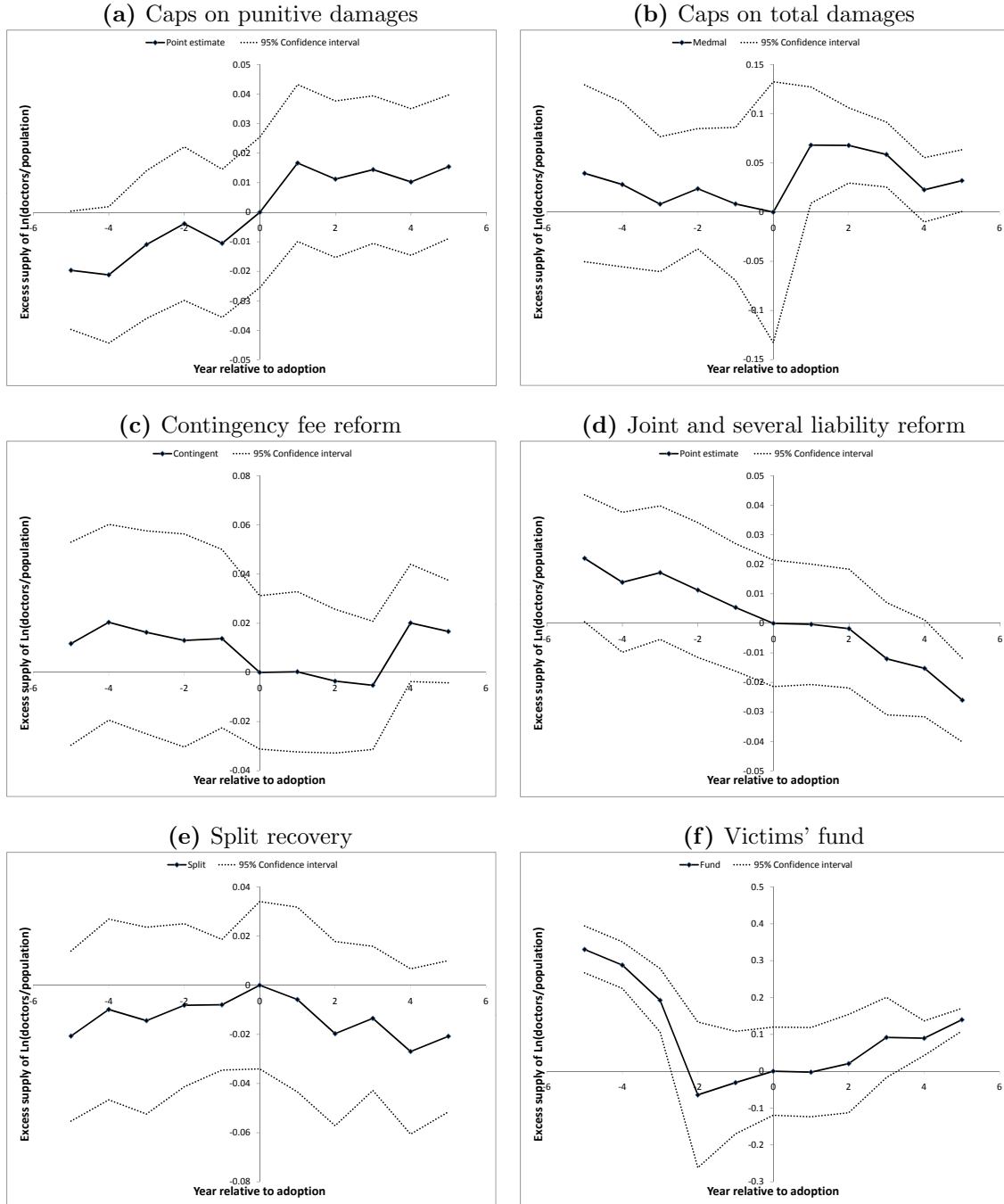
Furthermore, we can rule out specific channels by which physician supply might be thought to effect adoption of punitive damage caps. For example, one might suppose that state legislatures are public-spirited and decrease liability only when physician supply falls. Figure 6b demonstrates this phenomenon for states that adopted caps on total damages: a steady decline in physician supply is followed by a large increase

³⁵Figure 6a replicates Figure 1 from the introduction.

³⁶Currie and MacLeod (2008) makes a similar argument for the exogeneity of this reform.

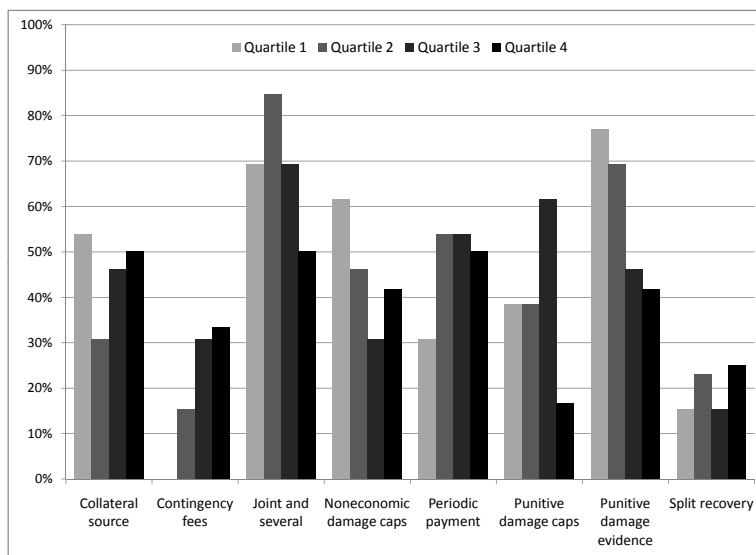
³⁷We exclude these five states from our main specification.

Figure 6: Tort reforms exhibiting pre-period changes in behavior



Note: These figures plot the normalized coefficients λ_j from the following regression: $y_{ist} = \sum_{j=-5}^5 \lambda_j D_{st+j} + \gamma X_{ist} + u_{ist}$, where y_{ist} is the log of the physician count for specialty i in state s in year t , D_{st+j} is an indicator for whether reform was first adopted in period $t + j$, and X_{ist} includes state-specialty and specialty-year fixed effects.

Figure 7: Fraction of states adopting reform from 1980-2001 by quartile of physician supply



Note: states are assigned to quartiles based on the total number of per capita physicians in a state in 1980. Total damage caps and victims' fund reforms are excluded due to insufficient number of adoptions.

once this liability-reducing reform is adopted. By contrast, Figure 6a shows that the exact opposite occurred for caps on punitive damages: supply *rose* prior to adoption.

Another potential channel for endogeneity is that legislatures could be captured by doctors so that, when the supply of doctors is high, legislatures reduce liability. Yet there does not appear to be a connection between states with high physician supply and states with punitive damage caps. This can be verified in Figure 7, which plots the fraction of states that ever adopt different tort reforms by quartile of physician supply in 1980. There does not appear to be a correlation between supply and adoption for punitive damage caps. By contrast there are clear patterns suggesting a public-spirited model for contingency fee reforms and a legislative capture model for punitive damage evidence reform.

An alternative concern is that tort reform is correlated with an unobserved variable that affects physician supply. For example, business-friendly states may attract physicians because they make it easy to set up a new practice, and coincidentally these states may also be more likely to pass tort reform. If this were the case, we would expect to observe a trend in physician supply both before and after the enactment of punitive damage caps. But, Figure 6a shows a pre-trend only. Moreover,

the unobserved variable ought to cause trends for all sorts of tort reforms to behave similarly. As the remaining graphs in Figure 6 show, that is not the case. Finally, our regression results do not change when we additionally control for state per capita income and per capita government medical benefits.

Even if we could not rule out endogeneity in this manner, one must still justify why, among the two possible explanations for pre-period trends – endogeneity and anticipation effects – one should default to endogeneity. If the correct explanation is actually anticipation effects, filtering out pre-trends with state-specific trends to address endogeneity will cause one not only to underestimate permanent treatment effects, but also perhaps to estimate the wrong sign on treatment effects, as discussed in Section 1.

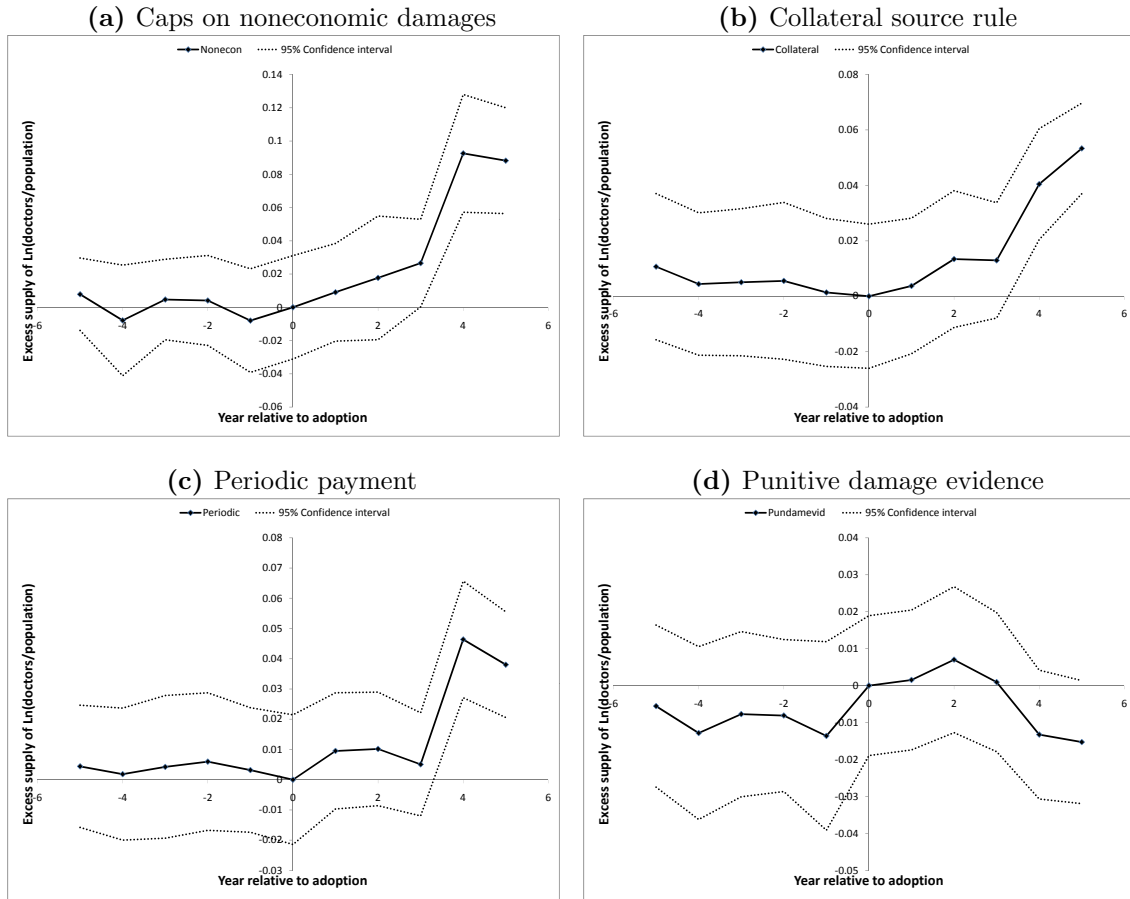
4.3.3 Punitive damage caps were anticipated

Ruling out endogeneity is a necessary condition for estimating anticipation effects but it may not be sufficient. We must show that physicians had both motive and capacity to anticipate the reforms we examine, our third criteria. Physicians have a large incentive to care about tort reform: variations in liability regimes across states have large impacts on their income. For example, neurosurgeons in St. Clair county, Illinois, paid an average premium of \$228,396 in 2004, but their colleagues in neighboring Wisconsin paid less than one-fifth of that (The Economist 2005). Moreover, they can be alerted to forthcoming reform through at least two possible channels: newspapers and insurance premiums.

Newspaper articles discussing upcoming legislation can directly inform physicians about potential future reforms. To verify this, we searched for newspaper stories about punitive damage caps prior to adoption of that reform. For example, in Pennsylvania, a large adopter, we found over 80 articles during the two years prior to adoption of reform in 1997. One article published about two years prior to enactment wrote that “the key goals of the [state] administration...have been to place a cap on punitive-damage awards” (Siegel 1995). We describe these findings in greater detail in Appendix A.4.

One might argue that physicians are not sophisticated enough to understand the impact of particular reforms on their liability exposure. Medical malpractice insurance companies, however, are certainly well informed about these reforms and it is possible that they indirectly signal forthcoming reform to physicians by decreasing

Figure 8: Tort reforms exhibiting no change in pre-period behavior

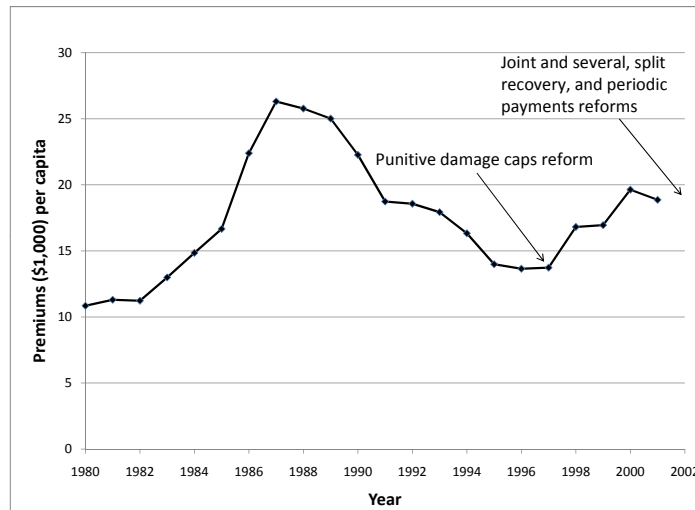


Note: These figures plot the normalized coefficients λ_j from the following regression: $y_{ist} = \sum_{j=-5}^5 \lambda_j D_{st+j} + \gamma X_{ist} + u_{ist}$, where y_{ist} is the log of the physician count for specialty i in state s in year t , D_{st+j} is an indicator for whether reform was first adopted in period $t+j$, and X_{ist} includes state-specialty and specialty-year fixed effects.

premiums when expected future physician liability decreases.³⁸ Figure 9 displays the premiums per capita for Pennsylvania during the period 1980-2001. Pennsylvania enacted punitive damage caps in 1997. This reform decreases liability and, indeed, we observe a decrease in premiums prior to this year. In 2002 Pennsylvania enacted another reform which raised liability (joint and several) and two reforms which decrease liability (split recovery and periodic payment). The rise in premiums prior to 2002 again suggests that joint and several reform was also anticipated by insurance companies.

Of course, one should not make strong inferences from this figure since it represents only one state and does not include any controls. In Figure 10, however, we plot medical malpractice premiums in the period leading up to enactment of punitive damage caps for all 50 states. This plot, which controls for state and year fixed effects, shows a fall in premiums prior to adoption. This is consistent with the increase in physician supply shown in Figure 1.³⁹

Figure 9: Per capita insurance premiums for Pennsylvania

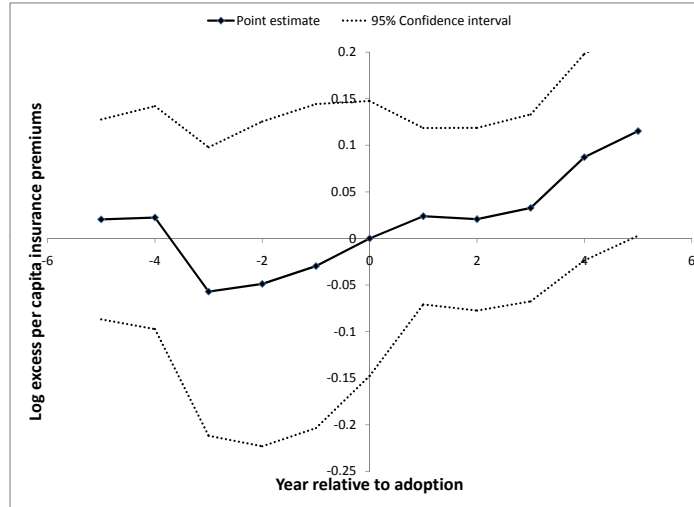


Note: Premium data are from AM Best. These plots display the direct premiums earned in a given calendar year divided by the number of high-risk physicians in that year. Physician data for 1984 and 1990 are interpolated. Amounts are in 1984 dollars.

³⁸This is not inconsistent with the fact that some states prohibit insurance coverage for punitive damages. Most states do not, and in those states premiums may signal forthcoming punitive damage caps. In Pennsylvania, for example, punitive damages assessed via vicarious liability, e.g., through physician groups, are insurable.

³⁹Appendix A.5 shows corresponding pre-trends in insurance premiums also exist for joint and several and split recovery reforms.

Figure 10: Excess amount of premiums before and after reform: annual leads and lags from 5 years before to 5 years after adoption



Note: Premium data are from AM Best. This plot displays the normalized coefficients λ_j from the OLS regression $y_{st} = \sum_{j=-5}^5 \lambda_j D_{st+j} + \gamma_s + \gamma_t + u_{st}$, where y_{st} is the log of the total amount of direct premiums earned in state s in time t divided by the number of high-risk physicians in state s in time t , D_{st} is a dummy variable that takes on the value of 1 only in the year that a state adopts reform, and γ_s and γ_t are state and year fixed effects. Standard errors are clustered by state.

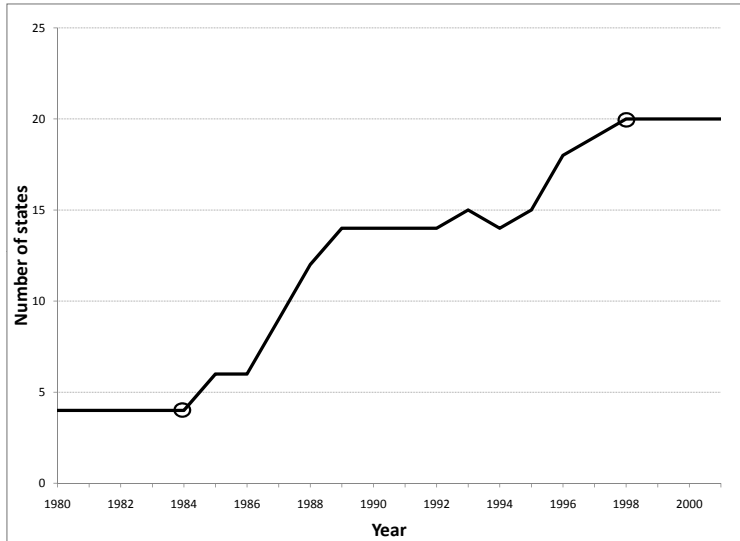
4.4 Empirical model

We estimate the effect of tort reform on the log of physician supply using a difference-in-differences strategy. Treatment effects are identified by comparing within-state changes in high-risk physician supply (tiers 1 and 2 in Table 1) in states that adopt reform in a year to within-state changes in supply among states that do not adopt in that year.⁴⁰ It would be sufficient to include state and year fixed effects to implement our difference-in-differences estimator. However, we go further and employ state-specialty and specialty-year fixed effects. The former control for specialty-level unobservables within each state. The latter allow time paths for physician supply to vary across specialty, as Figure 5 suggests may be appropriate.

We must select a pre and a post period in order to implement our difference-in-differences design. We could use the entire 1980-2001 panel to calculate these

⁴⁰We also separately estimated our models including all four risk tiers instead of only the two high-risk ones. This reduced the significance for joint and several and split recovery reforms, as expected. Our estimate of the effect of punitive damage caps, however, was *more* significant, suggesting that this reform has a broad impact across all specialties.

Figure 11: Cumulative number of states adopting punitive damage caps



contrasts but this is unappealing: observations from states that adopted reform early (late) would receive less weight in the pre (post) period than states that adopted reform later (earlier). Figure 11 shows that all caps on punitive awards were adopted in the period 1984 to 1998 (the circled points). Given the 1980 beginning and 2001 end of our sample, we implement the widest window that ensures full pre and post coverage for each treated state: a 9-year pre-post moving window that includes the 5 years preceding adoption of punitive caps and the 4 years after adoption.

We first estimate a myopic model to serve as a baseline:

$$y_{ist} = \beta_0^{myopic} d_{st} + \gamma_{is} + \gamma_{it} + u_{ist} \quad (12)$$

The outcome y_{ist} is the log of the number of physicians per capita practicing specialty i in state s in period t , d_{st} is an indicator for reform in state s in period t , and γ_{is} and γ_{it} are state-specialty and specialty-year fixed effects, respectively. We estimate this model with the addition of state-specific trends to obtain estimates under the assumption that pre-trends reflect endogeneity. Then we estimate it without state-specific trends to see estimates under the assumption that there is no endogeneity (and no anticipation effects).

We then estimate quasi-myopic models to account for anticipation effects. Specifically we estimate four quasi-myopic models that include up to four leading indicators

for whether a law was passed:

$$y_{ist} = \beta_0^{quasi} d_{st} + \sum_{j=1}^S \beta_j^{quasi} D_{s,t+j} + \gamma_{is} + \gamma_{it} + u_{ist} \quad (13)$$

$S = 1 \dots 4$ is the number of leading indicators in the regression and $D_{s,t+j}$ is a dummy variable equal to 1 if a reform was adopted in time period $t+j$. For example, if a reform is adopted in period 5, then $D_{s,t+1} = 1$ in period 4 and 0 otherwise. We parameterize the quasi-myopic model using treatment adoption dummies $D_{s,t+j}$ rather than merely concurrent treatment dummies d_{st} so that regression coefficients directly identify the parameters of interest: The permanent treatment effect is estimated by $\hat{\beta}_0^{quasi}$ and the temporary effect is estimated by $\hat{\beta}_0^{quasi} - \hat{\beta}_1^{quasi}$.

Finally we estimate the model we derived in Section 4.1, where physician supply is modeled as a function of exponentially discounted expectations of tort reforms:

$$y_{ist} = \beta d_{st} + \beta \sum_{j=1}^{\infty} \theta^j E_t [d_{s,t+j}] + \gamma_{is} + \gamma_{it} + \varepsilon_{ist}$$

We assume agents have rational expectations of future tort reforms and tort reform is exogenous.⁴¹

$$E_t [d_{s,t+j}] = d_{s,t+j} + v_{t,t+j}^d$$

This yields the estimable Euler equation

$$y_{ist} = \theta y_{is,t+1} + \beta d_{st} + \gamma_{is} + \gamma_{it} + w_{ist}$$

where $w_{ist} = \varepsilon_{it} - \theta \varepsilon_{i,t+1} + \beta \theta v_{t,t+1}^d + \beta (\sum_{j=2}^{\infty} \theta^j [v_{t,t+j}^d - v_{t+1,t+j}^d])$.

As discussed in the Appendix, binary treatment variables cause d_{st} to be endogenous if it is serially correlated over time. Adding a lead of the treatment variable to the estimation equation is sufficient to address this problem because our reform-generating process follows an $AR(1)$ process.⁴² Thus, our estimable Euler equation

⁴¹We also estimated an exponential discounting model, described in Appendix A.1, where agents have adaptive expectations. That analysis, available upon request, yielded estimates of the discount factor θ outside of the $[0, 1]$ interval. Because this is non-sensical, we conclude that adaptive expectations is not a good assumption for our model of physician supply.

⁴²To confirm this, we estimated the following state-level regression for *each* tort reform: $d_{st} = \alpha_1 x_{st} + \alpha_2 d_{st-1} + \alpha_3 d_{st-2} + e_{st}$, where x_{st} is a vector of controls that includes all other tort reforms. Our results (not reported) show that, at a 5% level of significance, α_2 is significant for all ten tort reforms while α_3 is insignificant for nine of them, which provides good support for our $AR(1)$ assumption. These regressions clustered standard errors at the state level and were unweighted.

becomes

$$y_{ist} = \theta y_{is,t+1} + \beta d_{st} + \delta d_{s,t+1} + \gamma_{is} + \gamma_{it} + w_{ist} \quad (14)$$

We estimate equation (14) first using OLS, then using our proposed leads of $y_{is,t+1}$ as instruments, and finally using lags of $y_{is,t+1}$ as instruments. We employ all available instruments in each category; restricting the number of instruments does not substantively affect our results.

We weight all our estimations by state population because y_{ist} is a per capita measure. Following the recommendations of Bertrand, Duflo, and Mullainathan (2004), we allow for arbitrary serial correlation in the error term as well as arbitrary cross-sectional correlation within states when computing standard errors. We employ one-step GMM estimation when estimating the exponential discounting models to alleviate concerns about finite sample problems associated with two-step GMM estimation as discussed in Judson and Owen (1999) and Doran and Schmidt (2006). We transform our data using forward orthogonal deviations instead of the usual first differences when estimating the exponential discounting models because this preserves sample size in panels with gaps.⁴³ The GMM standard error estimates and Arellano and Bond’s autocorrelation test assume error terms are uncorrelated across panels. The specialty-year fixed effects we include in our estimations increase the likelihood that this assumption holds.

Recall that punitive damage caps are most likely to be exogenous in states where the reform is not targeted solely at medical malpractice cases (see Table 3). We therefore exclude potentially endogenous states when performing our estimations.⁴⁴ Furthermore, our main specification excludes other tort reforms as controls because their endogeneity could contaminate our estimates. To allow for the possibility that we have been too conservative in enforcing exogeneity, however, we also estimate an alternative specification that includes all states and controls for all other tort reforms.⁴⁵

⁴³Recall that we do not have data on physician counts for 1984 or 1990. See Arellano and Bover (1995) and Roodman (2009a) for descriptions of the orthogonal deviations transform.

⁴⁴Specifically, we exclude CO, IL, OR, PA, and WI from the punitive damage caps analysis.

⁴⁵We also estimated a specification that controlled for state per capita income and per capita government medical benefits. Including these additional controls does not substantively affect our results.

4.5 Results

Table 4 reports estimates from the myopic model (0 leads) and versions of the quasi-myopic model (1–4 leads). The coefficient estimates on the time- t treatment variable identify the permanent effect of reform, including anticipation effects, and can be interpreted as relative changes in physician supply. Column 1 includes state-specific trends and estimates that punitive damage caps reduced physician supply by 1.2%. As predicted, this is the smallest effect reported in this table. (Recall that trends bias the estimated treatment effect downward, as illustrated in Figure 3.) Column 2 removes the state-specific trends, which increases the estimated effect to 3.9%. Moving across the first row reveals that the estimated permanent effect monotonically increases from 3.9% to 5.0% as we add leads. All estimates are significant.

Next we turn to estimates for our exponential discounting model. Under rational expectations, the regression model is given by the Euler equation (14). Column 1 of Table 5 reports OLS estimates of this equation. These estimates, although statistically significant, are inconsistent because OLS estimation of equation (14) does not account for the correlation between the error term and $y_{is,t+1}$. Column 2 estimates the Euler equation using leads of $y_{is,t+1}$ as instruments for $y_{is,t+1}$. The estimated temporary and permanent effects of punitive damage caps are significant at 2.2% and 5.7%, respectively. Finally, column 3 reports results when we use lags of y_{ist} rather than leads to instrument for $y_{is,t+1}$. The estimated effects are similar to those estimated using leads as instruments.

Table 5 also displays results for the Hansen test of overidentifying restrictions and Arellano and Bond’s autocorrelation test. Both tests support the validity of the instruments we use for our exponential discounting model. A remaining concern is that our instrument collection is overfitting our model (and also possibly weakening the power of our Hansen test, as discussed in Roodman 2009b). We address this by alternatively estimating our models using only two lags or two leads as instruments. Those results are similar in both magnitude and significance to what we report in Table 5.

Specification 1 in Table 6 summarizes the results from Tables 4 and 5 for the different models of anticipation effects. All estimates are strongly significant. In Section 2.1, we explained that imperfect correlation between time- t reform status and future reform status means that the estimated treatment effect in a myopic model

likely underestimates the permanent effect of reform.⁴⁶ Including leads reduces this bias by reducing omitted variable bias. Table 6 shows that the estimated effects from a quasi-myopic model are larger than the corresponding estimates from the myopic model, as predicted. The fact that these estimates increase as we keep adding leads to the quasi-myopic model suggests that each additional lead moves us closer to an unbiased estimate of the permanent effect. Combining this with our result that the exponential discounting model yields even larger estimates of the permanent effect of tort reform strongly suggests that anticipation effects matter and for perhaps longer than our data permit in the quasi-myopic model.

Specification 2 in Table 6 reports results when we include the nine other, potentially endogenous, tort reforms as controls and exclude no states. Compared to Specification 1, the magnitude of the estimated effects are larger for the quasi-myopic model and insignificant for the exponential discounting model. We still observe an increase in the estimated permanent effect for this reform as we add leads to the quasi-myopic model.

Like other prior studies on this topic, we do not account for general equilibrium effects. A physician fleeing one state necessarily enters another, magnifying the relative supply differences between the two states. Kessler, Sage, and Becker (2005) have previously demonstrated, however, that most of the equilibrium adjustment comes from newly graduated residents deciding where to practice and retirees leaving practice. Furthermore, we are primarily interested in the *relative* differences between our model estimates, for it is these relative differences that reveal the importance of anticipation effects.

Our results are consistent with Currie and MacLeod (2008), which finds that joint and several liability (examined in our Appendix) and punitive damage caps have a significant effect on birth outcomes. They find only limited evidence of anticipation, but this is probably due to their inclusion of state \times time fixed effects. This can bias estimation of anticipation effects, as shown in Figure 3.

⁴⁶This assumes temporary treatment effects and anticipation effects have the same sign and that current treatment is positively correlated with future treatment, on average.

5 Conclusion

There is a wide array of applied economics topics in which a researcher may be confronted with forward-looking agents whose responses anticipate future treatment. Economic theory suggests, for example, that individuals are forward looking when purchasing durable goods such as cars or houses or making human capital investments, and that firms are forward looking when investing in physical capital or entering new markets. In all these cases, anticipation causes a change in outcomes before the treatment occurs. Incorrectly assuming that this change is due to endogeneity rather than anticipation will produce biased estimation. While not all economic decisions are made with an eye towards the future and not all shocks are anticipated, enough are that empirical work should consider how to define and estimate treatment effects in the context of anticipation effects.

This paper develops a framework that addresses the two basic problems with estimating forward-looking models: the researcher does not know to what extent agents are forward looking and cannot observe their expectations. The framework itself posits that outcomes are additively separable in each period's expectations. We discuss two sets of parametric restrictions on expectations terms: one that caps the number of terms the researcher has to consider (the quasi-myopic model) and another that restricts their influence in a manner that allows differencing to eliminate all but one expectation term (the exponential discounting model). We also discuss two ways of relating unobserved expectations to observables: rational expectations in the text and adaptive expectations in the Appendix. For each we discuss some instruments that can be employed to address measurement errors that arise when using variables as proxies for unobservable expectations. Our application illustrated the potential importance of accounting for anticipation effects. Both the quasi-myopic and exponential discounting model suggest that the permanent effect of the tort reforms we study are double that suggested by a myopic model.

The framework has a number of limitations. Foremost, we offer no clean test that distinguishes between the presence of anticipation effects versus endogeneity. We merely provide informal evidence in favor of anticipation and thus our framework. Further, within our framework, we offer no formal way to discriminate between the different sets of parametric restrictions (i.e., the quasi-myopic and exponential discounting models) we discuss. There may be other restrictions a researcher might

employ or estimation strategies that do not require any restrictions at all. For example, if two agents were both treated but one found out about the treatment earlier than the other, one could estimate anticipation effects with a difference-in-differences estimator that would eliminate many expectations terms. Likewise, there may be alternative models of updating or belief formation that can be employed. Ideally the researcher would directly survey agents about their expectations or at least survey a subsample to empirically estimate the relationship between expectations and observables. Even where this is not possible, there are gains to specifying a more general model of forecasting than rational or adaptive expectations, even one that includes both future realizations of the forecasted variable as well as past forecasting errors. Each of these limitations is a useful topic for future research.

Tables

Table 1: Physician specialties by risk tier

Tier 1	Tier 2	Tier 3	Tier 4
Emergency medicine	Anesthesiology	Allergy & immunology	Diabetes
General practice	General surgery	Dermatology	Medical oncology
Neurological surgery	Orthopedic surgery	Nephrology	Neoplastic diseases
Obstetrics & gynecology	Plastic surgery	Physical medicine & rehabilitation	Psychiatry
Thoracic surgery	Radiology	Rheumatology	Public health & general preventive medicine

Source: Klick and Stratmann (2007). Specialties in tier 1 exhibit the highest average medical malpractice award per doctor and specialties in tier 4 exhibit the lowest average.

Table 2: Tort reform descriptions

Tort reform	Description
Collateral source	Allows damages to be reduced by the value of compensatory payments already made to the plaintiff
Contingency fees	Places limits on attorney contingency fees
Joint and several	Limits damages recoverable from parties only partially responsible for the plaintiff's harm
Noneconomic damage caps	Limits awards for noneconomic damages in malpractice cases
Periodic payment	Requires part or all of damages to be paid in the form of an annuity
Punitive damage caps	Prohibits or limits recovery of punitive damages from physicians
Punitive evidence	Requires plaintiff to show by clear and convincing evidence that a defendant acted recklessly
Split recovery	Requires some of the punitive damages to go to a state fund for uncompensated tort victims
Total damage caps	Limits awards for total damages
Victims' fund	Establishes a no-fault compensation fund for medical malpractice victims

Source: Avraham (2010).

Table 3: Summary of tort reform laws enacted during 1980-2001

Tort reform	States enacting tort reform
Collateral source	AL (87), CO (87) , CT (85), HI (87) , ID (90), IN (87), KY (89), MA (87), ME (90) , MI (86), MN (85), MT (88), ND (88), NJ (88), NY (85) , OR (88), UT (87), WI (95)
Contingency fees	CT (87), FL (86) , HI (87), IL (85), MA (87) , ME (89), MI (85) , NH (87), UT (86)
Joint and several	AK (86), AZ (87), CA (86), CO (87), CT (87), FL (86) , GA (88), HI (87), IA (84), ID (88), LA (81), MI (87) , MN (89), MO (86) , MS (90), MT (88), ND (88), NE (92), NH (90), NJ (88), NM (82), NY (87), TN (92), TX (86), UT (86), WA (86), WI (94), WV (86) , WY (86)
Noneconomic damage caps	AL (87), CO (87), HI (87), KS (87) , MD (87), MN (86), MO (86), MT (96), ND (96) , OR (88), UT (88), WI (95)
Periodic payment	AZ (89), CO (89) , CT (88), FL (87), IA (88), ID (88), IL (86), IN (85), LA (85) , MD (87), ME (87), MI (86), MN (89), MO (86), MT (87), NY (86), OH (88), RI (88), SD (88), UT (86), WA (86)
Punitive damage caps	AK (98), AL (87), CO (87) , GA (88), IL (85) , IN (95), KS (88), NC (96), ND (93), NH (87), NJ (96), NV (89), OK (96), OR (88), PA (97) , VA (89), WI (85)
Punitive evidence	AK (86), AL (87), AZ (87), CA (88), DC (96), FL (00) , GA (88), IA (87), ID (88), IN (84), KS (88), KY (89), MD (92), ME (85), MO (86), MS (94), MT (85), NC (96), ND (87), NJ (96), NV (89), OH (88), OK (87), OR (88), SC (88), TN (92), TX (88), UT (90), WI (95)
Split recovery	AK (98), CO (87), FL (87) , IA (87), IN (96), OR (88), UT (90)
Total damage caps	CO (89), SD (86)
Victims' fund	ND (83)

Source: Avraham (2010). Year of enactment given in parentheses. Bold face indicates the reform applies to medical malpractice torts only.

Table 4: Myopic and quasi-myopic (QM) OLS estimates for punitive damage caps

Tort reform	Number of leads					
	0	0	1	2	3	4
Punitive damage caps	0.012*	0.039**	0.042**	0.045*	0.050**	0.050*
	(0.006)	(0.018)	(0.020)	(0.022)	(0.025)	(0.026)
Lead (t+1)			0.012	0.015	0.020	0.020
			(0.010)	(0.014)	(0.016)	(0.017)
Lead (t+2)				0.008	0.014	0.014
				(0.011)	(0.013)	(0.014)
Lead (t+3)					0.015*	0.015
					(0.008)	(0.010)
Lead (t+4)						-0.000
						(0.006)
State trends	Yes	No	No	No	No	No
Model	Myopic	Myopic	QM	QM	QM	QM
Observations	6,363	6,363	6,363	6,363	6,363	6,363
R^2	0.992	0.989	0.989	0.989	0.989	0.989

Dependent variable is log of count of high-risk physicians per 100,000 population. Reported treatment effects compare within-state changes in adopting versus non-adopting states. Standard errors, given in parentheses, are clustered by state. A */** next to the coefficient indicates significance at the 10/5% level.

Table 5: Exponential discounting model estimates for punitive damage caps

Tort reform	(1)	(2)	(3)
Punitive damage caps (permanent effect)	0.046** (0.018)	0.057** (0.017)	0.064** (0.027)
Punitive damage caps (temporary effect)	0.015** (0.005)	0.022** (0.007)	0.022** (0.007)
Discount rate ($\hat{\theta}$)	0.665** (0.060)	0.622** (0.098)	0.654** (0.100)
Estimation method	OLS	GMM	GMM
IV	None	Leads	Lags
Observations	5,389	4,448	5,089
R^2	0.994		
Hansen test (p-value)		1	1
AR(3) test (p-value)		0.686	0.763

Dependent variable is log of count of high-risk physicians per 100,000 population. Reported treatment effects compare within-state changes in adopting versus non-adopting states. Standard errors, given in parentheses, are clustered by state. A */** next to the coefficient indicates significance at the 10/5% level. Permanent effect is defined as the temporary effect divided by one minus the discount rate. Column (2) instruments for the endogenous variable y_{ist+1} with all available leads of y_{ist+1} . Column (3) alternatively instruments with all available lags of y_{ist+1} . The AR(3) test checks for order-3 serial correlation in the residuals.

Table 6: Summary of estimated permanent effects for punitive damage caps

Model	IV	Specification	
		(1)	(2)
Myopic	None	0.039**	0.045**
Quasi-myopic (1 lead)	None	0.042**	0.046**
Quasi-myopic (2 leads)	None	0.045*	0.050**
Quasi-myopic (3 leads)	None	0.050**	0.056**
Quasi-myopic (4 leads)	None	0.050*	0.059**
Exponential discounting	Lags	0.064**	0.049
Exponential discounting	Leads	0.057**	0.054

Dependent variable is log of count of high-risk physicians per 100,000 population. Reported treatment effects compare within-state changes in adopting versus non-adopting states. A */** next to the coefficient indicates significance at the 10/5% level. Specification 1 summarizes the previous results from Tables 4 and 5. Specification 2 includes all 50 states in the analysis and adds the nine other (potentially endogenous) tort reforms in our data set as controls.

References

- Acemoglu, D. and J. Linn (2004, August). Market size in innovation: Theory and evidence from the pharmaceutical industry. *The Quarterly Journal of Economics* 119(3), 1049–1090.
- Alpert, A. (2010). The anticipatory effects of medicare part d on drug utilization. Working paper, University of Maryland.
- AMA (1997, September). Appendix ii: Graduate medical education. *Journal of the American Medical Association* 278(9), 775–776.
- Anderson, S. T., R. Kellogg, and J. M. Sallee (2011, July). What do consumers believe about future gasoline prices?
- Arellano, M. and S. Bond (1991, April). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *Review of Economic Studies* 58(2), 277–97.
- Arellano, M. and O. Bover (1995, July). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* 68(1), 29–51.
- Autor, D. H., J. J. Donohue, and S. J. Schwab (2006, 08). The costs of wrongful-discharge laws. *The Review of Economics and Statistics* 88(2), 211–231.
- Avraham, R. (2007). An empirical study of the impact of tort reforms on medical malpractice settlement payments. *Journal of Legal Studies* 36(S2), S183–S229.
- Avraham, R. (2010). Database of State Tort Law Reforms (DSTLR 3rd). *SSRN eLibrary*.
- Avraham, R., L. Dafny, and M. Schanzenbach (2010). The impact of tort reform on employer-sponsored health insurance premiums. *Journal of Law, Economics, and Organization*.
- Ayers, B. C., C. B. Cloyd, and J. R. Robinson (2005, April). “read my lips . . .”: Does the tax rhetoric of presidential candidates affect security prices? *Journal of Law & Economics* 48(1), 125–48.
- Baicker, K. and A. Chandra (2006, July). The labor market effects of rising health insurance premiums. *Journal of Labor Economics* 24(3), 609–634.
- Becker, G. S., M. Grossman, and K. M. Murphy (1994, June). An empirical analysis of cigarette addiction. *American Economic Review* 84(3), 396–418.

- Bertrand, M., E. Duflo, and S. Mullainathan (2004, February). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics* 119(1), 249–275.
- Bhattacharya, J. and W. B. Vogt (2003, October). A simple model of pharmaceutical price dynamics. *Journal of Law & Economics* 46(2), 599–626.
- Blundell, R. and S. Bond (1998, August). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87(1), 115–143.
- Blundell, R., M. Francesconi, and W. Van der Klaauw (2010). Anatomy of policy reform evaluation: Announcement and implementation effects. Technical report, Working paper.
- Born, P., W. K. Viscusi, and T. Baker (2006, March). The effects of tort reform on medical malpractice insurers’ ultimate losses. Working Paper 12086, National Bureau of Economic Research.
- Carroll, C. D. (2003, February). Macroeconomic expectations of households and professional forecasters. *The Quarterly Journal of Economics* 118(1), 269–298.
- Chow, G. C. (1989, August). Rational versus adaptive expectations in present value models. *The Review of Economics and Statistics* 71(3), 376–84.
- Cohen, T. (2004). *Medical malpractice trials and verdicts in large counties, 2001*. US Dept. of Justice, Office of Justice Programs, Bureau of Justice Statistics.
- Cohen, T. and K. Harbacek (2011). *Punitive Damage Awards in State Courts, 2005*. US Dept. of Justice, Office of Justice Programs, Bureau of Justice Statistics.
- Currie, J. and W. B. MacLeod (2008, 05). First do no harm? tort reform and birth outcomes. *The Quarterly Journal of Economics* 123(2), 795–830.
- de Figueiredo, Rui J P, J. and R. G. Vanden Bergh (2004, October). The political economy of state-level administrative procedure acts. *Journal of Law & Economics* 47(2), 569–88.
- Doran, H. E. and P. Schmidt (2006, July). Gmm estimators with improved finite sample properties using principal components of the weighting matrix, with an application to the dynamic panel data model. *Journal of Econometrics* 133(1), 387–409.

- Finkelstein, A. (2004, May). Static and dynamic effects of health policy: Evidence from the vaccine industry. *The Quarterly Journal of Economics* 119(2), 527–564.
- Gruber, J. and B. Koszegi (2001, November). Is addiction “rational”? theory and evidence. *The Quarterly Journal of Economics* 116(4), 1261–1303.
- Hamilton, J. D. (1994, January). *Time Series Analysis* (1 ed.). Princeton University Press.
- Hansen, L. P. (1982, July). Large sample properties of generalized method of moments estimators. *Econometrica* 50(4), 1029–54.
- Heckman, J. J. and R. J. Robb (1985). Alternative methods for evaluating the impact of interventions. *Longitudinal Analysis of the Labor Market Data*, 156–245.
- Helland, E., J. Klick, and A. Tabarrok (2005, Spring). Data watch: Tort-uring the data. *Journal of Economic Perspectives* 19(2), 207–220.
- Jacobson, L. S., R. J. LaLonde, and D. G. Sullivan (1993, September). Earnings losses of displaced workers. *American Economic Review* 83(4), 685–709.
- Judson, R. A. and A. L. Owen (1999, October). Estimating dynamic panel data models: a guide for macroeconomists. *Economics Letters* 65(1), 9–15.
- Kahn, C. M. (1986, March). The durable goods monopolist and consistency with increasing costs. *Econometrica* 54(2), 275–94.
- Karpoff, J. M., J. Lott, John R., and E. W. Wehrly (2005, October). The reputational penalties for environmental violations: Empirical evidence. *Journal of Law & Economics* 48(2), 653–75.
- Kessler, D. P., W. M. Sage, and D. J. Becker (2005, June). Impact of malpractice reforms on the supply of physician services. *Journal of the American Medical Association* 293, 2618–2625.
- Klick, J. and T. Stratmann (2007, 06). Medical malpractice reform and physicians in high-risk specialties. *Journal of Legal Studies* 36(S2), S121–S142.
- Lemos, S. (2006, February). Anticipated effects of the minimum wage on prices. *Applied Economics* 38(3), 325–337.

- Lueck, D. and J. A. Michael (2003, April). Preemptive habitat destruction under the endangered species act. *Journal of Law & Economics* 46(1), 27–60.
- Matsa, D. A. (2007, 06). Does malpractice liability keep the doctor away? evidence from tort reform damage caps. *Journal of Legal Studies* 36(S2), S143–S182.
- McCallum, B. T. (1976, January). Rational expectations and the natural rate hypothesis: Some consistent estimates. *Econometrica* 44(1), 43–52.
- McCullough, Campbell, and Lane LLP (2004). The insurability of punitive damages.
- Mertens, K. and M. O. Ravn (2011). Understanding the aggregate effects of anticipated and unanticipated tax policy shocks. *Review of Economic Dynamics*.
- Poterba, J. M. (1984, November). Tax subsidies to owner-occupied housing: An asset-market approach. *The Quarterly Journal of Economics* 99(4), 729–52.
- Roodman, D. (2009a). How to do xtabond2: An introduction to difference and system gmm in stata. *Stata Journal* 9(1), 86–136.
- Roodman, D. (2009b, 02). A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics* 71(1), 135–158.
- Ryoo, J. and S. Rosen (2004, February). The engineering labor market. *Journal of Political Economy* 112(S1), S110–S140.
- Siegel, R. (1995, February 8). Aarp, madd oppose tort-reform proposals they’re against punitive damage caps, changes in liability. whitman favors the controversial reforms. *The Philadelphia Inquirer*.
- Stango, V. (2003, October). Strategic responses to regulatory threat in the credit card market. *Journal of Law & Economics* 46(2), 427–52.
- Stark, K. (1999, September 2). City tipped scale for malpractice pa.’s cat fund payouts reached a record this year. / more than half the money went to cases in phila. *The Philadelphia Inquirer*.
- The Economist (2005, December 17). Scalpel, scissors, lawyer. *The Economist*.
- Viscusi, W. and P. Born (2005). Damages caps, insurability, and the performance of medical malpractice insurance. *Journal of Risk and Insurance* 72(1), 23–43.
- Wilson, Elser, Moskowitz, Edelman, and Dicker LLP (2008). *Punitive Damages Review*.

- Witt, J. (1984, March 4). Why soak the taxpayers? *The San Diego Union-Tribune*.
- Wolfers, J. (2006, December). Did unilateral divorce laws raise divorce rates? a reconciliation and new results. *American Economic Review* 96(5), 1802–1820.
- Zeiler, K., C. Silver, B. Black, D. Hyman, and W. Sage (2007). Physicians' insurance limits and malpractice payments: Evidence from texas closed claims, 1990-2003. *Journal of Legal Studies* 36(2), 9.

A Appendix

A.1 Adaptive expectations

In this section we derive an estimable Euler equation under the assumption that agents have adaptive expectations and show how to estimate it. One can show that the object of the agent's expectations (outcomes or treatment) does not affect identification of β or θ . For ease of exposition we assume agents have adaptive expectations about outcomes:

$$E_t [y_{t+1}] = E_t [y_{t+j}] = \phi y_t + (1 - \phi) E_{t-1} [y_t]$$

Plugging these equations into equation (17) and simplifying yields

$$y_t = \theta \phi y_t + \theta (1 - \phi) E_{t-1} [y_t] + \beta d_t + \varepsilon_t \tag{15}$$

The one-step back version of equation (17) is:

$$y_{t-1} = \theta E_{t-1} [y_t] + \beta d_{t-1} + \varepsilon_{t-1}$$

Solve this for $E_{t-1} [y_t]$ and plug the result into (15). Simplifying then produces the estimable Euler equation

$$y_t = \gamma (1 - \phi) y_{t-1} + \gamma \beta d_t - \gamma \beta (1 - \phi) d_{t-1} + \gamma \varepsilon_t - \gamma (1 - \phi) \varepsilon_{t-1}$$

where $\gamma \equiv 1 / (1 - \theta \phi)$. Time- t outcomes are now a function of past rather than future outcomes. The reason is that adaptive expectations is a backward-looking model of learning. The coefficient on current treatment no longer directly identifies β , though that parameter can be identified. Finally, the only source of endogeneity is previous period model error: $E [y_{t-1} \varepsilon_{t-1}] \neq 0$. Estimation is therefore straight-forward: use lags of order three or deeper and/or leads of order one or higher as instruments.

A.2 Rational expectations about outcomes

Consider the case where the agent is able to form rational expectations about outcomes.⁴⁷ Initially we assume realizations are a function of expectations: $y_{t+j} = E_t[y_{t+j}] + v_{t,t+j}^y$, where $v_{t,t+j}^y$ indicates the error given time t expectations about outcomes in time $t+j$ and $E[E_t[y_{t+j}]v_{t,t+j}^y] = 0$. This model is appropriate, for example, where outcomes are stock prices since realizations of stock prices are a composite of expectations (e.g., Chow 1989). Expectations at time t about θy_{t+1} are

$$\theta E_t[y_{t+1}] = E_t[\theta \beta \sum_{j=0}^{\infty} \theta^j E_{t+1}[d_{t+1+i}]] \quad (16)$$

since $E_t[e_{t+1}] = 0$. Subtracting (16) from (5) yields the Euler equation

$$y_t = \theta E_t[y_{t+1}] + \beta d_t + e_t \quad (17)$$

Plugging in our rational expectations assumption produces the estimation equation

$$y_t = \theta y_{t+1} + \beta d_t + w_t \quad (18)$$

where $w_t = e_t - \theta v_{t,t+1}^y$.

The error term has two components, model error (e_t) and unexpected, mean-zero forecast error ($v_{t,t+1}^y$), that cause outcomes to deviate from forecasts. Thus rational expectations introduces measurement error. The result is endogeneity between next period's outcome y_{t+1} and $v_{t,t+1}^y$. Furthermore, if $\{d_t\}$ are serially correlated, then d_t would be correlated with y_{t+1} and thus $v_{t,t+1}^y$ through d_{t+1} .

If we had instead assumed expectations about outcomes were a function of actual outcomes, i.e., $E_t[y_{t+j}] = y_{t+j} + v_{t,t+j}^y$ where $E[y_{t+j}v_{t,t+j}^y] = 0$, then there would be no endogeneity. The Euler equation would look the same, but $w_t = e_t + \theta v_{t,t+1}^y$. By assumption y_{t+1} is exogenous, and even with serial correlation in $\{d_t\}$ we would have $E[d_t v_{t,t+1}^y] \neq 0$.

Estimating this model is straightforward. Suppose that expectations are a function of outcomes and $E_t[y_{t+1}] = y_{t+1} + v_{t+1}^y$.⁴⁸ Our error term is

$$w_t = e_t + \theta v_{t+1}^y$$

⁴⁷We ignore the role of covariates x_t to simplify the exposition. However, it is straightforward to incorporate covariates into the analysis.

⁴⁸If outcomes are a function of expectations ($y_{t+1} = E_t[y_{t+1}] + v_{t+1}^y$), then y_{t+1} is exogenous and no instruments are required.

The analogue to Assumption A2' is that, for some constant H ,

$$E[w_t w_{t+j}] = E[(e_t - \theta v_{t+1}^y)(e_{t+j} - \theta v_{t+j+1}^y)] = 0 \quad \forall j > H$$

If e_t and v_t^y are serially and mutually uncorrelated then the usual difference and system GMM estimators can be used so long as Assumptions A1 and A3 hold. Limited correlation in the error term can be accommodated by using higher order leads.

A.3 Binary treatment variables

In many applications the treatment variable, d_t , is binary. The forecast error corresponding to rational expectations of a binary variable is necessarily mean reverting, which induces a negative correlation between d_{t+j} and $v_{t,t+j}^d$.⁴⁹ If treatment states are correlated over time then endogeneity occurs because $E[d_t v_{t,t+j}^d] \neq 0$ and the error term will be serially correlated, violating Assumption A2' from Section 3.1. This problem can be resolved if agent forecasts follow a Markov process because then future treatment states can be used to absorb the endogeneity. More specifically, suppose that

$$\begin{aligned} Cov[d_t, v_{t,t+j}^d | d_{t+1}, d_{t+2}, \dots, d_{t+K}] &= Cov[d_t, v_{t+1,t+1+j}^d | d_{t+1}, d_{t+2}, \dots, d_{t+K}] \\ &= 0 \quad \forall t, \forall j > 0, K \geq 1 \end{aligned}$$

Note that this assumption can be tested by running simple OLS regressions. If the assumption holds, one can then consistently estimate our Euler equation

$$y_t = \theta y_{t+1} + \alpha x_t + \beta d_t + \sum_{k=1}^K \delta_k d_{t+k} + \eta_t + w_t$$

where d_t is binary and d_{t+k} accounts for endogeneity.

A.4 Discussion of reforms in newspapers

In this section we provide evidence that the three tort reforms we examine (punitive damage caps in the main text, and joint and several reform and split recovery in Appendix A.5) were discussed in local newspapers years prior to actual passage of these reforms. We first determine, for each reform, the largest state that adopted it. We then search the online archives of the two largest newspapers in that state for

⁴⁹Recall that $v_{t,t+j}^d$ is defined as the forecast error from the time- t forecast of d_{t+j}

articles pertaining to the reform in question.⁵⁰ Some states do not have searchable databases of articles from local newspapers that span the period before adoption of reform. In those cases we search the archives of local papers for the next-largest state that adopted the reform.

California reformed its joint and several liability rules on June 3, 1986. Two large local newspapers, the Los Angeles Times and the San Diego Union-Tribune, have archives going back to January 01, 1985 and December 05, 1983, respectively. We searched the online archives of these two papers from their earliest available point up through June 3, 1986 and found 84 articles mentioning “joint and several”, 20 articles mentioning both “joint and several” and “tort reform”, and 19 articles mentioning “medical malpractice” and “tort reform”. One article published more than two years prior to actual tort reform discusses the need for the California state legislature to carefully re-examine its laws regarding joint and several liability (Witt 1984).

Pennsylvania reformed its punitive damage caps rules on January 25, 1997 and adopted no other tort reforms in that decade. Two local newspapers, The Philadelphia Inquirer and The Pittsburgh Post-Gazette, have archives reaching back to January 1, 1994 and March 1, 1993, respectively. We found 84 articles published between January 1, 1994 and January 25, 1997 that mentioned “tort reform” and 6 that mentioned “punitive damage caps”. One article written about two years prior to enactment said that “the key goals of the [state] administration... have been to place a cap on punitive-damage awards” (Siegel 1995).

Pennsylvania also reformed its split recovery rule for punitive damages on March 20, 2002. We searched all articles published in The Philadelphia Inquirer and The Pittsburgh Post-Gazette between January 1, 1999 and March 20, 2002. We found 627 articles mentioning “punitive damages” and 115 articles mentioning “tort reform”. One article published more than two years prior to passage of split recovery reform mentions that a state senator was advocating a bill “that would limit recovery of punitive damages” (Stark 1999).

A.5 Results for joint and several reform and split recovery

In this section we estimate anticipation effects for two additional tort reforms: joint and several liability reform split recovery reform. The doctrine of joint and several

⁵⁰Data on the circulation size of local newspapers can be obtained from Mondo Newspapers at <http://www.mondonewspapers.com/usa/index.html>

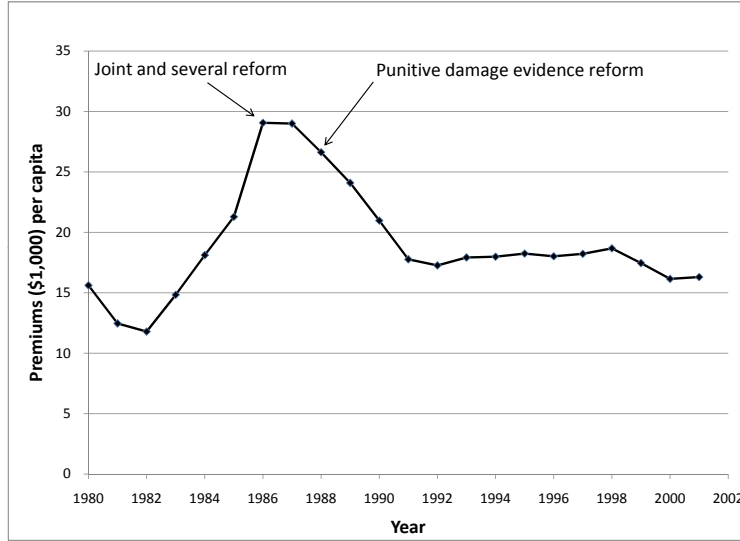
liability allows plaintiffs to recover full damages from a defendant who is only partially at fault. In the context of medical malpractice, this means a plaintiff can sue her hospital rather than her doctor for large sums of money even if the hospital bears little blame for the plaintiff's injury. Reform of joint and several liability limits this by either requiring defendants to be responsible for a large fraction of the blame before have to pay full damages or holding defendants responsible for only their proportionate share of damages based on their comparative fault for the plaintiff's injury. This increases physician liability by holding physicians more accountable for their actions. Split recovery decreases physician liability by stipulating that the state receive a portion of any punitive damages awarded to the plaintiff.

Figures 6d and 6e suggest that supply fell prior to joint and several reform (which *increases* liability) and rose prior to split recovery reform (which *reduces* liability). Furthermore, there does not appear to be a connection between states with low physician supply and states with joint and several reform or between states with high physician supply and states with split recovery reform. This can be verified in Figure 7, which plots the fraction of states that ever adopt different tort reforms by quartile of physician supply in 1980.

Appendix A.4 provided evidence that these two reforms were discussed in newspapers prior to their reforms. Figure 12 displays the log of per capita medical malpractice insurance premiums for California during the period 1980-2001. California enacted reforms to joint and several liability in 1986 and increased the amount of evidence required to justify punitive damage awards in 1988. These two reforms increase and decrease liability, respectively. The rise in premiums prior to 1986 and the subsequent decrease provide evidence that insurance companies anticipated these reforms. Figure 13a plots medical malpractice premiums in the period leading up to joint and several reform for all 50 states. This plot, which controls for state and year fixed effects, shows a rise in premiums prior to adoption. This is consistent with the decrease in physician supply shown in Figure 6d. The analogous plot for split recovery reform, shown in Figure 13b, displays a decrease in premiums, consistent with the increase in physician supply shown in Figure 6e.

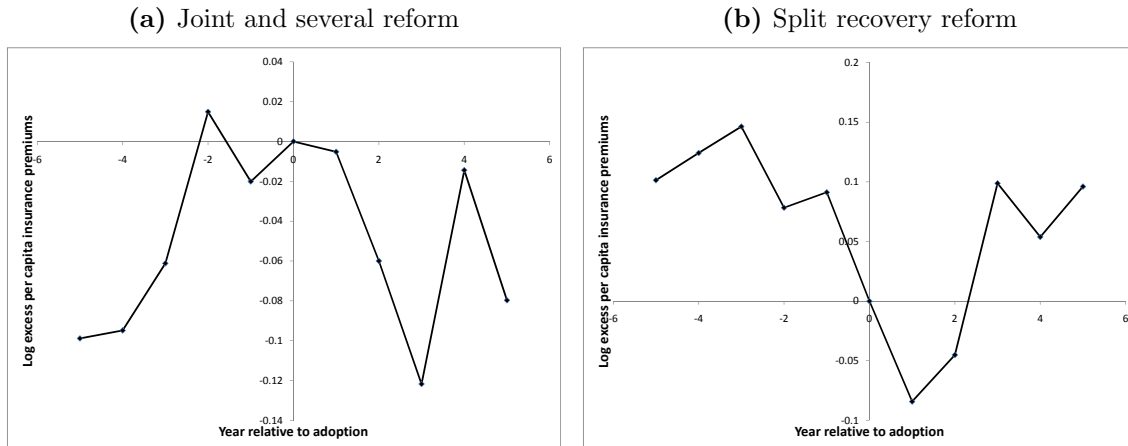
Appendix Tables 7 – 10 present our estimation results and Appendix Table 11 summarizes them. Our estimates for joint and several reform are significant for almost all models while those for split recovery are significant only for the exponential discounting model. Table 11 shows that, just as with punitive damage caps, the es-

Figure 12: Per capita insurance premiums for California



Note: Premium data are from AM Best. These plots display the direct premiums earned in a given calendar year divided by the number of physicians in the state in that year for all 50 states. Physician data for 1984 and 1990 are interpolated. Amounts are in 1984 dollars.

Figure 13: Excess amount of premiums before and after reform: annual leads and lags from 5 years before to 5 years after adoption



Note: premium data are from AM Best. This plot displays the normalized coefficients λ_j from the OLS regression $y_{st} = \sum_{j=-5}^5 \lambda_j D_{st+j} + \gamma_s + \gamma_t + u_{st}$, where y_{st} is the log of the total amount of direct premiums earned in state s in time t divided by the number of physicians in state s in time t , D_{st} is a dummy variable that takes on the value of 1 only in the year that a state adopts reform, and γ_s and γ_t are state and year fixed effects.

timated effects increase as we add leads in the quasi-myopic model and are largest overall for the exponential discounting model. Including state-specific trends reduces estimates to almost precisely zero.

Appendix tables

Table 7: Myopic and quasi-myopic (QM) OLS estimates for joint and several

Tort reform	Number of leads					
	0	0	1	2	3	4
Joint and several	0.002 (0.004)	-0.028 (0.017)	-0.032* (0.018)	-0.037* (0.019)	-0.038* (0.020)	-0.040** (0.020)
Lead (t+1)			-0.014* (0.007)	-0.021** (0.009)	-0.023** (0.010)	-0.024** (0.010)
Lead (t+2)				-0.026** (0.013)	-0.028* (0.014)	-0.029** (0.014)
Lead (t+3)					-0.004 (0.007)	-0.005 (0.008)
Lead (t+4)						-0.004 (0.010)
State trends	Yes	No	No	No	No	No
Model	Myopic	Myopic	QM	QM	QM	QM
Observations	5,473	5,473	5,473	5,473	5,473	5,473
R^2	0.995	0.995	0.995	0.995	0.995	0.995

Dependent variable is log of count of high-risk physicians per 100,000 population. Reported treatment effects compare within-state changes in adopting versus non-adopting states. Standard errors, given in parentheses, are clustered by state. A */** next to the coefficient indicates significance at the 10/5% level.

Table 8: Myopic and quasi-myopic (QM) OLS estimates for split recovery

Tort reform	Number of leads					
	0	0	1	2	3	4
Split recovery	0.007 (0.006)	0.036 (0.029)	0.036 (0.030)	0.035 (0.032)	0.036 (0.032)	0.042 (0.033)
Lead (t+1)			0.003 (0.013)	0.002 (0.015)	0.003 (0.015)	0.008 (0.016)
Lead (t+2)				-0.004 (0.009)	-0.004 (0.009)	0.002 (0.009)
Lead (t+3)					0.004 (0.008)	0.010 (0.010)
Lead (t+4)						0.024** (0.009)
State trends	Yes	No	No	No	No	No
Model	Myopic	Myopic	QM	QM	QM	QM
Observations	8,965	8,965	8,965	8,965	8,965	8,965
R^2	0.993	0.991	0.991	0.991	0.991	0.991

Dependent variable is log of count of high-risk physicians per 100,000 population. Reported treatment effects compare within-state changes in adopting versus non-adopting states. Standard errors, given in parentheses, are clustered by state. A */** next to the coefficient indicates significance at the 10/5% level.

Table 9: Exponential discounting model estimates for joint and several

Tort reform	(1)	(2)	(3)
Joint and several (permanent effect)	-0.036* (0.021)	-0.041* (0.022)	-0.067 (0.044)
Joint and several (temporary effect)	-0.013* (0.007)	-0.015* (0.008)	-0.013** (0.006)
Discount rate ($\hat{\theta}$)	0.641** (0.048)	0.642** (0.087)	0.802** (0.074)
Estimation method	OLS	GMM	GMM
IV	None	Leads	Lags
Observations	4,615	3,746	4,445
R^2	0.997		
Hansen test (p-value)		1	1
AR(3) test (p-value)		0.226	0.604

Dependent variable is log of count of high-risk physicians per 100,000 population. Reported treatment effects compare within-state changes in adopting versus non-adopting states. Standard errors, given in parentheses, are clustered by state. A */** next to the coefficient indicates significance at the 10/5% level. Permanent effect is defined as the temporary effect divided by one minus the discount rate. Column (2) instruments for the endogenous variable y_{ist+1} with all available leads of y_{ist+1} . Column (3) alternatively instruments with all available lags of y_{ist+1} . The AR(3) test checks for order-3 serial correlation in the residuals.

Table 10: Exponential discounting model estimates for split recovery

Tort reform	(1)	(2)	(3)
Split recovery (permanent effect)	0.050* (0.026)	0.041* (0.021)	0.061* (0.037)
Split recovery (temporary effect)	0.014** (0.006)	0.015** (0.008)	0.013** (0.005)
Discount rate ($\hat{\theta}$)	0.723** (0.051)	0.622** (0.098)	0.791** (0.076)
Estimation method	OLS	GMM	GMM
IV	None	Leads	Lags
Observations	7,579	6,216	7,119
R^2	0.995		
Hansen test (p-value)		1	1
AR(3) test (p-value)		0.993	0.982

Dependent variable is log of count of high-risk physicians per 100,000 population. Reported treatment effects compare within-state changes in adopting versus non-adopting states. Standard errors, given in parentheses, are clustered by state. A */** next to the coefficient indicates significance at the 10/5% level. Permanent effect is defined as the temporary effect divided by one minus the discount rate. Column (2) instruments for the endogenous variable y_{ist+1} with all available leads of y_{ist+1} . Column (3) alternatively instruments with all available lags of y_{ist+1} . The AR(3) test checks for order-3 serial correlation in the residuals.

Table 11: Summary of estimated permanent effects for joint and several (JS) and split recovery (SP)

Model	IV	Reform	
		JS	SP
Myopic	None	-0.028	0.036
Quasi-myopic (1 lead)	None	-0.032*	0.036
Quasi-myopic (2 leads)	None	-0.037*	0.035
Quasi-myopic (3 leads)	None	-0.038*	0.036
Quasi-myopic (4 leads)	None	-0.040**	0.042
Exponential discounting	Leads	-0.041*	0.041*
Exponential discounting	Lags	-0.067	0.061*

Dependent variable is log of count of high-risk physicians per 100,000 population. Reported treatment effects compare within-state changes in adopting versus non-adopting states. A */** next to the coefficient indicates significance at the 10/5% level. Standard errors are clustered by state.