NBER WORKING PAPER SERIES

ACCOUNTING FOR ANTICIPATION EFFECTS: AN APPLICATION TO MEDICAL MALPRACTICE TORT REFORM

Anup Malani
Julian Reif

Working Paper 16593
http://www.nber.org/papers/w16593

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2010

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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Accounting for Anticipation Effects: An Application to Medical Malpractice Tort Reform
Anup Malani and Julian Reif
NBER Working Paper No. 16593
December 2010
JEL No. C50,I18,J20,K13

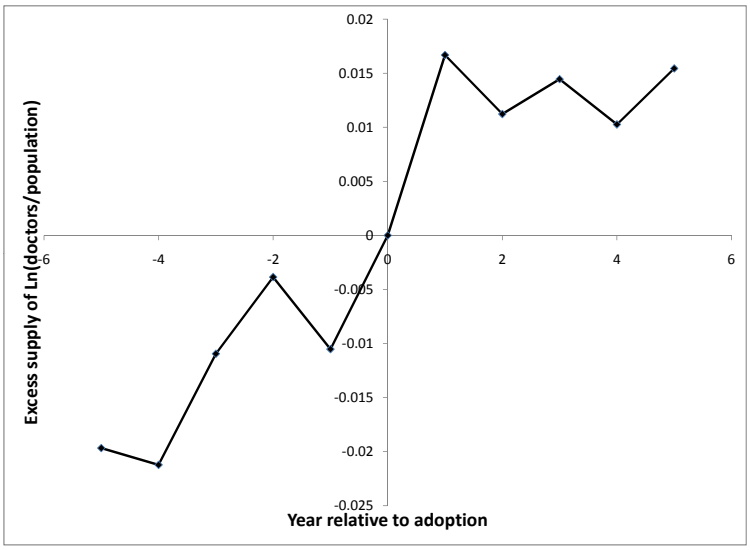
ABSTRACT

While conducting empirical work, researchers sometimes observe changes in behavior before the adoption of a new treatment program or policy. The conventional diagnosis researchers make is that the treatment is endogenous. Observing behavioral changes prior to treatment is also consistent, however, with anticipation effects. In this paper we provide a framework for comparing the different methods for estimating anticipation effects and propose a new set of instrumental variables that can address the problem that subjects' expectations are unobservable. We use our framework to analyze the effect of tort reform on physician supply. We find that accounting for anticipation effects doubles the estimated effect of tort reform.

Anup Malani
University of Chicago Law School
1111 E. 60th Street
Chicago, IL 60637
and NBER
amalani@uchicago.edu

Julian Reif
University of Chicago
5254 S Drexel Ave APT 3C
Chicago IL 60615
jreif@uchicago.edu

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: $\ln y_{ist} = \Sigma_{j=-5}^{5} \lambda_j D_{st+j} + \gamma X_{ist} + u_{ist}$, where y_{ist} is physician supply 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.

While conducting empirical work, researchers sometimes observe changes in behavior 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. 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.[1]

Observing behavioral changes 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

[1] One might object that the graph shows a pre-period trend in treatment states. (The regression generating the pre-post graph controls for time fixed effects.) This trend raises the question of why there is such a trend only in treatment states. The answer suggests that identifying such trends is just another way of suggesting that the treatment is endogenous.

if individuals are forward looking, have access to information on future treatment, and there is a benefit to acting before treatment is adopted.

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. Thus it is more likely 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. Moreover, physicians have a large financial incentive to change behavior prior to adoption: medical errors made prior to tort reforms are subject to lower penalties under the new regime.

Determining whether treatment is endogenous or merely anticipated has important consequences for inference. Whereas endogeneity may cause the researchers to over- or underestimate a treatment effect, anticipation effects usually cause researchers to underestimate them. The reason is that the typical before-after comparison attributes anticipatory treatment effects to the control group. As a result, it not only ignores, but deducts, anticipatory treatment effects from the overall treatment effect.[2]

This paper makes three contributions. First, it provides 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, it examines how to address the problem that agents' expectations are unobservable and proposes a new set of instrumental variables that can be employed to overcome it. Finally, it estimates the effect of tort reform on physician supply accounting for anticipation effects.

Our framework starts with a forward-looking regression of the form

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

[2]This 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). So anticipation of the statute increased habitat destruction before adoption, and reduced it after adoption.

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.[3] In this model, treatment at time t increases outcomes by λ_0 in period t and by λ_j in the j-th period prior to treatment. The full effect of treatment at t, including all anticipation effects, is $\lambda_0 + \Sigma_{j=1}^{\infty}\lambda_j$. This forward-looking regression model has a wide array of applications, including investments in human capital (e.g., Ryoo and Rosen 2004), rational addiction models (e.g., Becker, Grossman, and Murphy 1994), present value models (e.g., Chow 1989), R&D investment decisions (e.g., Acemoglu and Linn 2004), and pricing of durable goods (e.g., Kahn 1986) and real estate (e.g., Poterba 1984).

The initial problem with estimating the model above is the potentially infinite number of anticipation terms. A common response in the empirical microeconomics literature is to estimate a quasi-myopic model that omits anticipation terms more than T periods prior to treatment.[4] Indeed, this is the sort of model employed to generate pre-post graphs such as Figure 1 that are ubiquitous in the empirical microeconomics literature (see, e.g., Autor, Donohue, and Schwab 2006, Finkelstein 2004). If agents respond earlier than T 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 (see, 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.[5]

[3] We assume the agent's information set is $\Omega_t = \{y_0, ..., y_{t-1}, x_0, ..., x_t, d_0, ..., d_t\}$, where the x are possible covariates.

[4] Although there are a large number of examples, the following are typical: de Figueiredo and Vanden Bergh (2004), Ayers, Cloyd, and Robinson (2005), Lueck and Michael (2003), Bhattacharya and Vogt (2003), Acemoglu and Linn (2004), Finkelstein (2004), Lemos (2006), Gruber and Koszegi (2001), and Mertens and Ravn (2011).

[5] There is a difference, however, between the approach we employ in this paper and that employed in the macroeconomic literature. Whereas we start with something like a present-value model for outcomes and

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 allows for the first time 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 frequent approach 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).[6] Of course, unless actual expectations are observed, the investigator can only estimate a reduced form model which demonstrates that expectations affect outcomes, but does not identify the precise slope of the relationship.[7]

The natural response 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 usual source for these instruments is a subset of the agent's information set, for instance, lags of the observable variable (see McCallum 1976).[8] We demonstrate, however that the strength of these instruments depends on the omitted portion of the information set, which is unknown.

Therefore, we propose an alternative set of instruments: leads of the observable vari-

derive an Euler equation from that model, the macroeconomics literature starts with an Euler equation, perhaps derived from a utility maximization problem, and derives from that a present value model for asset prices.

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

[7]Isolating shocks that affect expectations but not actual treatment does not address the first problem of reducing the number of coefficients to be estimated in a forward-looking model. While a quasi-myopic model that examines how outcomes change in the periods between enactment at t and implementation at $t + k$ seems natural, it does not capture anticipation effects prior to t. Enactment constitutes a sharp increase in expectations about the probability of treatment at t, but it does not imply expectations about treatment were zero and had no effect on outcome before t.

[8]For a review of how to estimate a forward-looking rational expectations model, see Blake (1991).

able.[9] The idea is inspired by the autoregressive error-components model studied in the dynamic panel literature (e.g., Arellano and Bond 1991, Arellano and Bover 1995, Blundell and Bond 1998). That model resembles the forward-looking rational expectations model, except that the latter looks forward in time. We describe conditions under which leads are valid instruments in anticipation models. Our proposed instruments may be used on their own, or as complements to instruments from the agent's information set.

Finally, we compare the different methods of estimating anticipation effects by examining the effect of tort liability on physician supply. Prior literature has estimated the treatment effect of tort liability ignoring anticipation effects (e.g., Klick and Stratmann 2007, Matsa 2007). Our results suggest that accounting for anticipation effects increases the estimated impact of tort liability by a factor of two. We show in the main text that caps on punitive damages have a positive contemporaneous treatment effect on physician supply of 1.5 to 2.2 percent and a positive full effect of 5.0 to 6.4 percent after accounting for anticipation effects. In 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 a contemporaneous effect on physician supply of -1.3 to -1.5 percent and 1.3 to 1.5 percent and a full effect accounting for all anticipation effects of -4.0 to -6.7 percent and 4.1 to 6.1 percent, respectively.

The following is an outline of the remainder of the paper. Section 1 reviews the parameters of interest in the forward-looking model. Section 2 elaborates on the various parametric restrictions that may be employed to reduce the number of expectation terms in the forward-looking model, including the quasi-myopic model and the exponential discounting model. Section 3 takes up the problem that expectations are not observed. It examines the instruments that can be employed to address endogeneity from forecast errors, including the use of leads of endogenous variables as instruments. 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.[10]

[9] These instruments are valid for both rational expectations and, as we demonstrate in the appendix, adaptive expectations.

[10] The Appendix takes up topics that complement the discussion in the main text. Whereas section 3 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

1 Parameters of interest

Before estimating a model of anticipation effects such as

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

it is useful to define the possible parameters of interest.[11] The first parameter of interest is λ_0, which is the effect of one period of treatment at time t on time-t outcomes. This measure ignores anticipation effects prior to treatment. For convenience, we shall call this a one-period or contemporaneous treatment effect. In our medical malpractice application, the corresponding parameter is the additional physician labor supply in year t due to a temporary decrease in liability in t that lasts just one year. The baseline for this change is not $t-1$ but the inifinite past or, more practically, before agents anticipated the adoption of treatment.

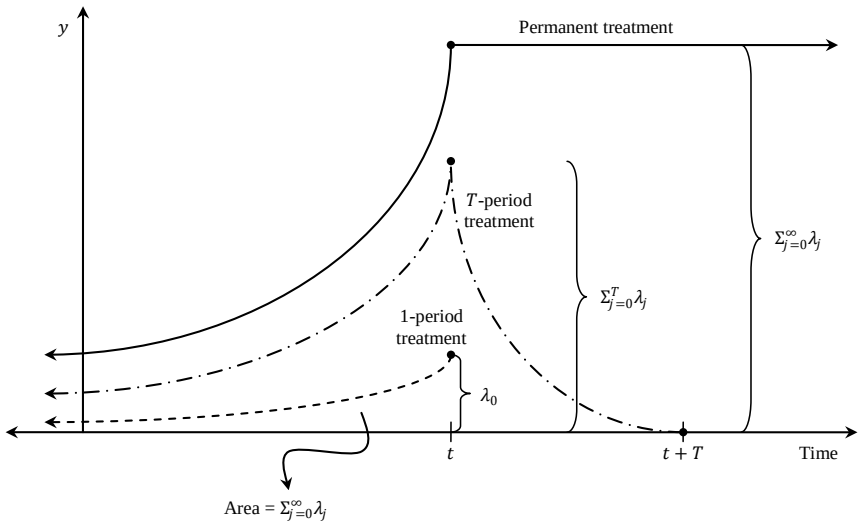
The second parameter of interest is $\Sigma_{j=0}^{\infty} \lambda_j$. This has two interpretations. One is the effect on time-t outcomes of a permanent treatment adopted at time t. This includes the effect on current outcomes of the current period of treatment plus the anticipation effects on *current outcomes* of *future periods of treatment*. In our application, this interpretation corresponds to the effect of a permanent increase in liability starting at time t on physician labor supply in year t. The other interpretation of the 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

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).

[11] Instead of parametrizing anticipation effects using a series of dummies $\{d_{t+j}\}$ that turn on and stay on while the agent is in treatment, one could model those effects using dummies that turn on only for defined periods prior to and after the start or end of treatment, e.g., $D_{t+j} = 1$ if treatment begins in period $t+j$, for $j = \{-\infty, \infty\}$. This will affect how we define parameters of interest, but not our empirical approach. When we estimate the quasi-myopic model, we will use D_t rather d_t dummies because it simplifies interpretation of coefficients.

It 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



effects on all *past outcomes* of the *current period of treatment*. In our application, this interpretation corresponds to the effect of a temporary increase in liability at time t on the total response of physician labor supply during both year t and, due to anticipation effects, in all prior years.

The two interpretations of the second parameter 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 area under the dotted line in the interval $(-\infty, t]$ is equal to the second parameter of interest. 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 effects are larger 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 $\Sigma_{j=k}^{\infty}\lambda_j$. An implication is that the level of y in every post-period is the second parameter of interest.[12]

Given the alternative interpretations of the second parameter, one might also be inter-

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

ested in a third parameter that examines the effect of intermediate duration treatments. It measures the effect of T periods of treatment starting at time t on outcomes in time t. We call this parameter, $\Sigma_{j=0}^{T}\lambda_j$, the current effect of finite treatment. An example from our medical malpractice application is the effect on time-t physician labor supply of a perfectly anticipated tort reform passed in time t that is fully expected to be overturned or repealed after T years. This quantity is illustrated by the dashed line in Figure 2. The line falls between adoption date t and expiration date $t+T$ because anticipation effects decline as agents anticipate the forthcoming end of treatment. This is recognized as an increasingly important parameter given that government policy is non-permanent and often subject to explicit sunsets, as discussed in Gersen and Posner (2007) and Maltzman and Shipan (2008). Estimating this parameter of interest requires knowledge of T, a topic taken up in Maltzman and Shipan (2008).

A fourth parameter of interest is the effect of a permanent treatment given expectations prior to adoption. In this case, the corrected full effect of treatment is

$$\Sigma_{j=1}^{\infty}\lambda_j - \Sigma_{j=0}^{\infty}\lambda_j E_t\left(d_{t+j}\right)$$

If subjects already expect that, with some positive probability, they will be treated, this may mitigate the effect of actually giving treatment. Estimating this corrected treatment effect requires knowledge of pre-adoption expectations, which are difficult to measure.

For the remainder of the paper, we will focus on estimating the contemporaneous effect of treatment and the full effect of treatment. The other parameters of interest require additional knowledge that does not hinge on how the forward-looking model is estimated. It is estimation of that model that is primary objective of this paper.

2 Simplifying the forward-looking model

The primary challenges with estimating a forward-looking model are the infinite number of expectation terms (the dimensionality problem) and their unobservability (the unobserved expectations problem). Here we explore the first challenge.

Researchers can reduce the number of expectation terms in three basic ways. 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 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. As we shall show in Section 2.2, this requires more instruments than the exponential discounting model, which demands only one.

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 depends on what the agent is able to forecast and how she forms those forecasts. 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

$$y_t = \beta_0 d_t + u_t$$

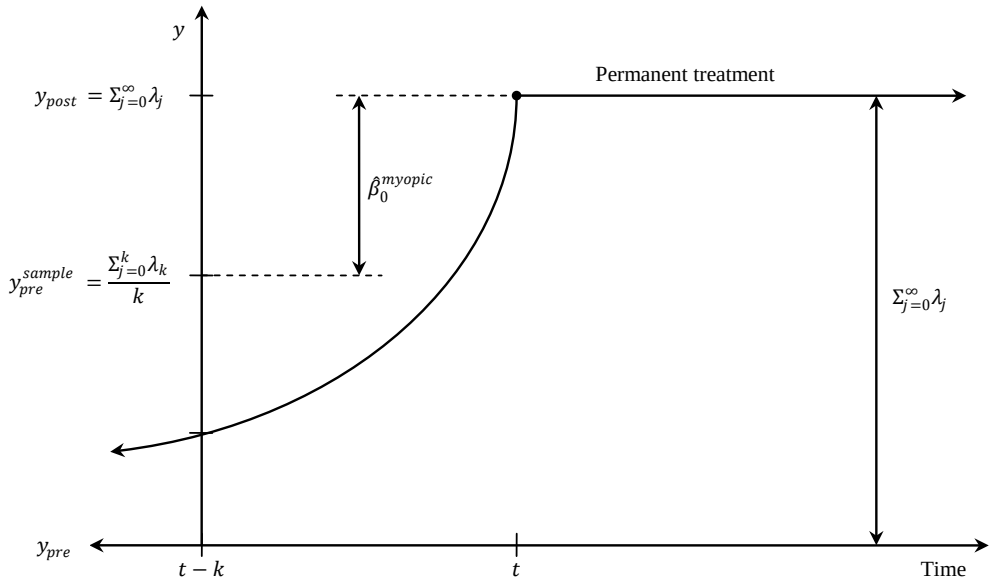
The omission of anticipation effects generates omitted variable bias. The specific nature of the bias depends on which parameter of interest from the previous section the researcher seeks to estimate.

If anticipation effects have the same sign as contemporaneous effects, the estimated coefficient $\hat{\beta}_0^{myopic}$ is probably larger (in absolute value) than the contemporaneous effect of treatment (λ_0) in the forward-looking model. The reason is that current treatment and expected future treatment are surely positively correlated: $Corr(d_{it}, E_t[d_{it+j}]) > 0$.[13] In this case,

$$\text{plim } |\hat{\beta}_0^{myopic}| = |\beta_0| + \Sigma_{j=1}^{\infty} |\beta_j| \alpha_j > |\lambda_0|$$

[13]Negative 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.

Figure 3: Estimate from a myopic model



where α_j is the coefficient from a regression of $E_t\left[d_{t+j}\right]$ on d_t. Intuitively, the coefficient on current treatment in the myopic model captures some of the effect of future treatment.[14]

Conversely, the myopic coefficient estimate is typically smaller (in absolute value) than the full effect of treatment in the forward-looking model since $\alpha_j \leq 1$, so that $\Sigma_{j=0}^{\infty}\left|\beta_j\right|\alpha_j < \Sigma_{j=0}^{\infty}\left|\lambda_j\right|$. 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 before an agent is treated. Thus the estimate from the myopic model underestimates the full effect of treatment.

This point is illustrated in Figure 3, which plots the outcome of a forward-looking process after adoption of a permanent treatment at time t. Assume that the full 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

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

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

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

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 treatment group (because expectations are operating) to the control group, and thus overestimated outcomes in the control group.

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:

$$y_t = \beta_0 d_t + \Sigma_{j=1}^{S} \beta_j E_t[d_{t+j}] + u_t \tag{2}$$

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.

The main weakness of this model is that the researcher must know the number of periods in which there are anticipation effects.[15] If the researcher underestimates this number, her coefficient estimates will suffer omitted variable bias just as the myopic coefficient estimate does. Assuming positive correlation in treatment over time, one would expect the quasi-myopic estimate $\hat{\beta}_0^{quasi}$ to overestimate the contemporaneous effect of treatment (λ_0) and the estimate $\Sigma_{j=0}^{S} \hat{\beta}_j^{quasi}$ to underestimate the full effect of treatment ($\Sigma_{j=0}^{\infty} \lambda_j$).

In practice, the quasi-myopic model has a second shortcoming. Researchers frequently address the problem that expectations are not observed by assuming rational expectations and substituting realizations of d_{t+j} for expectations of those variables. This not only

[15] Even if one employs a regulation enacted at time t but not implemented until time $t+k$ as a shock to expectations, one may not know the number of periods of anticipation effects. Unless the regulation was not at all anticipated until enacted, there may be pre-enactment anticipation effects. This implies more than k overall periods of anticipation effects.

reduces the effective sample by S periods, but also raises a more subtle problem. If adoption of treatment is not orthogonal to forecast error (i.e., $E[d_{t+j}, v^d_{t,t+j}] \neq 0$ where $v^d_{t,t+j}$ is the error in forecasting time $t+j$ treatment in time t), the quasi-myopic model suffers bias from measurement error. In that case, the quasi-myopic model can be estimated with instrumental variables to generate consistent estimates of anticipation effects. The quasi-myopic model requires more instruments, however, than the exponential discounting model. 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 contemporaneous treatment effect of $\lambda_0 = \beta$ and an anticipation effect j periods prior to treatment of $\lambda_j = \beta\theta^j$:

$$y_t = \beta d_t + \beta\Sigma_{j=1}^{\infty}\theta^j E_t\left[d_{t+j}\right] + e_t \tag{3}$$

The full effect of treatment is estimated with $\hat{\beta}/(1-\hat{\theta})$. The central benefit of the assumption that outcomes are a function of exponentially discounted expectations about treatment is that subtracting θy_{t+1} or $\theta E_t\left[y_{t+1}\right]$ from (3) 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 constant treatment effects yield lower bias. If there are more than S periods of anticipation effect, 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. In this subsection we derive Euler equations under the assumption that agents

have rational expectations. (We take up adaptive expectations in the appendix.) There are two sets of primitives about which agents may be able to form expectations, outcomes $\{y_t\}$ or the treatment $\{d_t\}$. In each case, moreover, there are two formulations of rational expectations, one where the realization of a variable depends on expectations ($z = E[z] + v$) and one where expectations depend on realizations ($E[z] = z + v$). Economic theory should dictate which path to take, but the choice will affect the structure of the error term in the resulting Euler equations.

2.3.1 Expectations about outcomes

Consider the case where the agent is able to form expectations about outcomes.[16] Initially we assume realizations are a function of expectations: $y_{t+j} = E_t[y_{t+j}] + v^y_{t,t+j}$, where $v^y_{t,t+j}$ indicates the error given time t expectations about outcomes in time $t+j$ and $E[E_t[y_{t+j}]\, v^y_{t,t+j}] = 0$. 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

$$\theta E_t[y_{t+1}] = E_t\left[\theta\beta\Sigma_{j=0}^{\infty}\theta^j E_{t+1}[d_{t+1+i}]\right] \tag{4}$$

since $E_t[e_{t+1}] = 0$. Subtracting (4) from (3) yields the Euler equation

$$y_t = \theta E_t[y_{t+1}] + \beta d_t + e_t \tag{5}$$

Plugging in our rational expectations assumption produces the estimation equation

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

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

The error term has two components: model error (e_t) and unexpected, mean-zero forecast error ($v^y_{t,t+1}$) 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 $v^y_{t,t+1}$. Furthermore, if $\{d_t\}$ are serially correlated, then d_t would be correlated with y_{t+1} and thus $v^y_{t,t+1}$ through d_{t+1}.

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

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

2.3.2 Expectations about treatments

Now consider the case where agents form expectations about treatment (not outcomes) and these expectations are a function of actual treatments: $E_t[d_{t+j}] = d_{t+j} + v^d_{t,t+j}$ where $E[d_{t+j}v^d_{t,t+j}] = 0$. This model is appropriate where treatments are exogenously assigned, as might be the case when the treatment is a tort reform (and outcome is physician supply) or medical demand (and the outcome is R&D investment). In this case we can substitute the rational expectations assumption directly into the basic forward-looking model to obtain

$$y_t = \beta\Sigma^{\infty}_{j=0}\theta^j d_{t+j} + e_t + \beta\Sigma^{\infty}_{j=1}\theta^j v^d_{t,t+j}$$

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

$$y_t = \theta y_{t+1} + \beta d_t + w_t \tag{7}$$

where

$$\begin{aligned} w_t &= e_t - \theta e_{t+1} + \beta\Sigma^{\infty}_{j=1}\theta^j v^d_{t,t+j} - \beta\Sigma^{\infty}_{j=2}\theta^j v^d_{t+1,t+j} \\ &= [e_t - \theta e_{t+1}] + \beta\theta v^d_{t,t+1} + \beta\Sigma^{\infty}_{j=2}\theta^j [v^d_{t,t+j} - v^d_{t+1,t+j}] \end{aligned}$$

The error term now 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 in 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 the $t+1$ outcome y_{t+1} and the error term: the model error e_{t+1} in period $t+1$. There is no endogeneity from $v^d_{t,t+1}$ because, although y_{t+1} is a function of d_{t+1}, we have assumed that d_{t+1} is orthogonal to $v^d_{t,t+1}$.[17] Nor

[17] Recall that y_{t+1} is not a direct function of $v^d_{t,t+1}$ because there is no error in the contemporary forecast of d_{t+1}, i.e., $E_{t+1}[d_{t+1}] = d_{t+1}$.

is there endogeneity from the change in forecasts ($v^d_{t,t+j} - v^d_{t+1,t+j}, j > 1$) because under rational expectations these forecast updates are orthogonal to prior forecast errors ($v^d_{t,t+j}$) and thus orthogonal to $E_t[d_{t+j}]$ too. Indeed, there is no additional endogeneity even if $\{d_t\}$ are serially correlated because $E[d_{t+j}v^d_{t,t+j}] = 0$ by assumption.

If instead we assumed treatments are endogenous, i.e., $d_{t+j} = E_t[d_{t+j}] + v^d_{t,t+j}$, then there is a second source of endogeneity because y_{t+1} is a function of d_{t+1} but $E\left[d_{t+1}v^d_{t,t+1}\right] \neq 0$. If additionally $\{d_t\}$ are serially correlated, there is third source of endogeneity since w_t contains $v^d_{t,t+1}$. Again, there remains no endogeneity from $v^d_{t,t+j}$ for $j > 1$ because under rational expectations updates of forecasts $v^d_{t,t+j} - v^d_{t+1,t+j}$ are orthogonal to prior forecast errors $v^d_{t,t+j}$ and thus d_{t+j}.

Findings about which set of assumption generate endogeneity (and why) are summarized in Table 1.

3 Unobserved expectations problem

The quasi-myopic model discussed in Section 2.2 and the Euler equations derived in the previous section reduce the dimensionality of the forward-looking model to a finite number of expectation terms. This brings us to the second challenge with estimating the forward-looking model: even the remaining expectation terms may not be observable.

A common approach is to isolate time t shocks that alter expectations about treatment but do not change treatment at time t. The canonical example is a regulation that is enacted at time t but not implemented until time $t+k$. Examples of such laws abound (e.g., Gersen and Posner 2007, Huber 2011)[18] and numerous studies employ a quasi-myopic model to examine outcomes between enactment and implementation (e.g., Alpert 2010, Gruber and Koszegi 2001, Lueck and Michael 2003, Blundell, Francesconi, and Van der Klaauw 2010). Another example is a time t disclosure of possible future treatment that does not ever change timing of treatment or is not followed by actual treatment. Applications are dominated by

[18] Most statutes have a slight gap between enactment date and implementation date, with the latter falling on the start of a month or quarter. Other times the gap is longer and is intended either to allow adaptation – which are the same as anticipation effects – or to temporarily grandfather existing violators and buy their political support.

event studies that examine the effect of news announcements on stock prices (e.g., Stango 2003, Karpoff, Lott, and Wehrly 2005).

This approach falls short for two reasons. First, unless the shock is unanticipated, the researcher will still underestimate the full anticipation effects of treatment. For example, if the enactment of the law was partly anticipated, then expectations about the probability of treatment just before enactment will be greater than zero even if implementation is delayed. The change in outcomes will reflect a change in expectations from a positive value to a greater positive value rather than from zero to a positive number. This will cause the researcher to underestimate full anticipation effects and, for any given degree of variance in unobservables, increase the probability of an insignificant result.[19]

Second, even if the researcher is able to identify whether there is any anticipation effect, she cannot identify the slope of the relationship between outcomes and expectations. This means that researcher cannot combine the slope estimate with parametric restrictions on the forward looking model to estimate the full anticipation effects of treatment. To see why the researcher cannot identify the slope, imagine running a two-stage least squares regression with the shock as an instrument for expectations. Unless the econometrician can actually observe agents' expectations, she cannot estimate a first-stage regression of expectations on the shock. Instead, she can only estimate a reduced-form regression of outcomes on the shock. Thus, the researcher cannot identify the slope of the relationship between outcomes and expectations. In short, a shock to expectations does not resolve the problem that expectations are not observed.

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. Unless the forecast error is orthogonal to the observable variables, however, the researcher will have to instrument for them. 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).[20] We

[19] In fact, if the enactment is not as big a surprise as the delay in implementation, agents' expectations may actually fall upon enactment. This would cause the researcher to estimate the wrong sign on anticipation effects!

[20] For a review of how to estimate a forward-looking rational expectations model, see Blake (1991).

offer a new set of instruments: leads of the endogenous variable. For continuity with the existing literature as well as the previous section, we review these instruments in the context of the Euler equations from the last section.

3.1 Instruments from the agent's information set

One candidate for an instrument is to replicate the agent's forecast, $E_t[y_{t+1}]$. McCallum (1976) notes that rational expectations implies that $E_t[y_{t+1}] = y_{t+1} - v^y_{t,t+1}$, where $v^y_{t,t+1}$ is uncorrelated with y_{t+1}. This means the econometrician can regress y_{t+1} (which is essentially a proxy for $E_t[y_{t+1}]$) on Ω_t:

$$y_{t+1} = \delta\Omega_t + v^y_{t,t+1}$$

The prediction from this first stage regression, $\widehat{y}_{t+1} = \widehat{\delta}\Omega_t$, is by construction equal to $E_t[y_{t+1}] = y_{t+1} - v^y_{t,t+1}$. It can therefore be substituted into the Euler equation without causing an endogeneity problem:

$$y_{t+1} = \theta\widehat{y}_{t+1} + \beta d_t + e_t$$

The challenge, however, is to obtain the agent's whole information set. McCallum used all the data available to him and assumed it was identical to – or fully captured – the information available to the agent. This will not generally be the case for most researchers unfortunately.

A second, more realistic candidate is to instrument for the agent's forecast with a subset Ω_{1t} of his full information set $\Omega_t = (\Omega_{1t}, \Omega_{2t})$ where for simplicity we assume $\Omega_{1t} \perp \Omega_{2t}$. In that case the first stage regression becomes

$$y_{t+1} = \delta_1\Omega_{1t} + \varepsilon_t$$

where $\varepsilon_t = \delta_2\Omega_{2t} + v^y_{t,t+1}$. The predicted value $\widehat{y}_{t+1} = \widehat{\delta}_1\Omega_{1t}$ measures the expectation with error: $E_t[y_{t+1}] = \widehat{y}_{t+1} + \delta_2\Omega_{2t}$. That measurement error, however, is orthogonal to the predicted value by construction. So, in the case where agents generate forecasts of outcomes and $E\left[y_{t+1}v^y_{t,t+1}\right] \neq 0$, corresponding to the top left cell of Table 1, we can consistently estimate the Euler equation

$$y_t = \theta\widehat{y}_{t+1} + \beta d_t + w_t$$

where $w_t = \delta_2 \Omega_{2t} + e_t$. Some modification is required to address the other specifications of rational expectations. These are summarized in Table 2.

3.2 Leads of the endogenous variable as instruments

We propose an alternative set of instruments for y_{t+1}: leads of y_{t+1}. Since leads of y_{t+1} and y_{t+1} itself both depend, according to the forward-looking model (3), on expectations about future treatment, shocks to leads of y_{t+1} will also move y_{t+1}. 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 bulk of this section is concerned with carefully describing the conditions under which consistent estimation is possible. We will work with a modified form of our Euler equation that allows for controls x_t and random effects η_i:

$$y_t = \theta y_{t+1} + \alpha x_t + \beta d_t + \eta_i + w_t \tag{8}$$

This is not only more general than what we have considered up to now, but also allows the researcher to take full advantage of panel data and dynamic panel estimators in particular.[21]

3.2.1 Dynamic panel estimation

There is a large literature, notably Arellano and Bond (1991), Arellano and Bover (1995) and Blundell and Bond (1998), that seeks to estimate autoregressive error components models of the form

$$y_t = \theta y_{t-1} + \alpha x_t + \eta_i + e_t \tag{9}$$

where $i = 1...N$, $t = 1...T$, and $|\theta| < 1$. There is an obvious link with equation (8), which simply substitutes y_{t+1} for y_{t-1}. In (9), it is assumed that η_i and e_t are independently distributed across i with $E[\eta_i] = E[e_t] = E[\eta_i e_t] = 0$. The number of time periods T is fixed and the number of individuals N is large. We assume for simplicity that x_t is strictly

[21] Estimation is simplified if we assume away the random effect η_i because then there is no need to take differences as done below. Nevertheless, restrictions on the serial correlation of the error term w_t are still necessary in order for leads of y_{t+1} to be valid instruments.

exogenous.[22] Direct OLS estimation of equation (9) is inconsistent because $E[y_{t-1}\eta_i] \neq 0$. Estimating first differences fails because $E[\Delta y_{t-1}\Delta e_t] \neq 0$.[23] Using an instrumental variable solves this problem but that requires additional data the researcher may not have. Alternatively, she can instrument for Δy_{t-1} using lags of y_{t-1} or instrument for y_{t-1} using lags of Δy_{t-1} if the following standard assumptions are met:

A1: $E[y_1 e_t] = 0 \; \forall \; i, \; \forall \; t \geq 2$

A2: $E[e_t e_s] = 0 \; \forall \; t \neq s$

A3: $E[\eta_i \Delta e_2] = 0 \; \forall \; i$

Assumption A1 requires the initial condition y_1 to be uncorrelated with future disturbances. Assumption A2 requires these disturbances to be uncorrelated. These two assumptions together imply the following moment conditions:

$$E[y_{t-j}\Delta e_t] = 0 \; \forall \; j \geq 2, \; \forall \; t \tag{10}$$

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

$$E[\Delta y_{t-1} e_t] = 0 \; \forall \; t \tag{11}$$

Equation (9) is overidentified if $T > 3$ but can be estimated using the Generalized Method of Moments (GMM) framework developed by Hansen (1982). "Difference GMM" estimation exploits the moment conditions (10) while "system GMM" estimation exploits both (10) and (11).

Assumption A2 is critical for the validity of these estimation procedures. As currently stated, 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 there remain enough lags for identification. We therefore loosen A2:

[22] This exogeneity assumption can be relaxed.

[23] A within estimator suffers from this same problem, although the bias disappears as $T \longrightarrow \infty$.

A2′: $E[e_t e_{t-j}] = 0 \; \forall \; j > H, \; H \geq 0$

This changes our moment conditions to

$$E[y_{t-j} \Delta e_t] = 0 \; \forall \; j \geq H + 2, \; \forall \; t$$

$$E[\Delta y_{t-H-1} e_t] = 0 \; \forall \; t$$

To exploit the analogy between the Euler equation from the forward-looking model and the autoregressive error components model, we must derive analogues to assumptions A1, A2′, and perhaps A3 and show they are satisfied. This will depend on the content of the error term in the Euler equation, which in turn depends on how rational expectations are specified. We consider the two main cases – expectations about outcomes and expectations about treatments – in turn.

3.2.2 Expectations about outcomes

Suppose that expectations are a function of outcomes and $E_t[y_{t+1}] = y_{t+1} + v^y_{t+1}$.[24] Our error term is

$$w_t = e_t + \theta v^y_{t+1}$$

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

$$E[w_t w_{t+j}] = E[\left(e_t - \theta v^y_{t+1}\right)\left(e_{t+j} - \theta v^y_{t+j+1}\right)] = 0 \; \forall \; j > H$$

If e_t and v^y_t 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.

3.2.3 Expectations about treatments

Suppose expectations are a function of treatments and $E_t[d_{t+1}] = d_{t+1} + v^d_{t,t+1}$. Our error term becomes substantially more complicated:

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

$$w_t = e_t - \theta e_{t+1} + \beta\theta v^d_{t,t+1} + \beta(\Sigma_{k=2}^{\infty}\theta^i[v^d_{t,t+k} - v^d_{t+1,t+k}]) \tag{12}$$

This error term follows an $MA(1)$ process so we must discard one lead for our estimators. The analogue to Assumption A2′ is $E[w_t w_{t+j}] = 0 \; \forall \; j > H$ for some constant $H \geq 1$. This will hold if the following four conditions are satisfied for all periods t and individuals i:

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

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 $(v^d_{t,t+k} - v^d_{t+1,t+k}) = 0$) or no serial correlation (so $E[v^d_{t+j,t+j+1}v^d_{t+l,t+k}] = 0 \; \forall \; j,k,l$). 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 (10) and (11) as

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

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.[25] 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.

Our Monte Carlo simulations (available upon request) suggest that serial correlation in the forecast errors are unlikely to cause much bias in practice. Regardless, the researcher can and should perform Hansen, Difference-in-Sargan, and Arellano-Bond autocorrelation tests to validate her dynamic panel assumptions.

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 the appendix.

3.3 Comparing instruments

Whereas section 3 discussed using elements of the agent's information set – typically, lags of the outcome variable – as instruments, we proposed in section 3.2 to use leads of the outcome variable as instruments. Without knowing the agent's full information set, however, it is not possible to determine which set of instruments are superior.

The power of a set of instruments depends on first stage fit, which in turn depends on the variance of the first stage errors. Consider the case where the agent is assumed to be able to forecast outcomes, so that the researcher wants to instrument for $E_t[y_{t+1}]$. When instrumenting with a portion Ω_{1t} of the agent's information set, the first stage is $E_t[y_{t+1}] = \delta_1\Omega_{1t} + \delta_2\Omega_{2t}$. Since the researcher cannot observe $E_t[y_{t+1}]$, she employs y_{t+1} as a proxy. This makes the first stage

$$y_{t+1} = \delta_1\Omega_{1t} + \varepsilon_t$$

[25] It 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 our maximization problem.

Consider the analogous problem for an OLS problem: $\min_{\beta_0,\beta_1} (y - \beta_1 x_1 - \beta_0 - k(\beta_1))^2$ where $k(\beta_1)$ is some constant that is independent of x_1. Identification of β_0 is impossible but OLS can still identify β_1 even without knowing the form of $k(\beta_1)$.

where $\varepsilon_t = \delta_2 \Omega_{2t} + v^y_{t,t+1}$.

When instrumenting for $E_t[y_{t+1}]$ with y_{t+j} for $j > 1$, however, the first stage is $E_t[y_{t+1}] = \gamma y_{t+j} + u_t$. After replacing the dependent variable with y_{t+1}, we get

$$y_{t+1} = \gamma y_{t+j} + \omega_t$$

where $\omega_t = u_t + v^y_{t,t+1}$. The difference between the errors in the two first stages depends on $\delta_2 \Omega_{2t}$ and u_t, neither of which is likely to be known a priori. At a minimum, arguing that the Ω_{1t} yields superior instruments requires knowing how much of the agent's information set is excluded. If one knew that, however, one could simply replicate the agent's exact forecast.

To be clear, our claim is not that instrumenting with leads is as good as instrumenting with part of the agent's information set. Our claim is simply that there is no reason, *a priori* to prefer one to the other. Indeed, the researcher should select strong candidates from both sets of instruments so as to maximize her power.

There are common situations, however, when either lags or leads are invalid instruments. For instance, if agents do not continuously update their forecasts of future variables with new, orthogonal information, lags may no longer be valid instruments. To illustrate why, we examine the case where agents have rational (i.e., unbiased) forecasts of treatment but never update these forecasts.[26] This implies the forecast error no longer depends on the date the forecast was made, so that $E_t[d_{t+j}] = d_{t+j} + v^d_{t,t+j} = d_{t+j} + v^d_{t+j}$. The exponential discounting model may now be written

$$y_t = \beta \Sigma_{j=0}^{\infty} \theta^j (d_{t+j} + v^d_{t+j}) + e_t$$

Subtracting y_{t+1} yields the Euler equation

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

Because y_{t-j} for any $j > 1$ is correlated with future values of y_{t-j}, one of which is y_t, y_{t-j} will be correlated with v_{t+1}. Although lags are no longer orthogonal to the error term in the Euler equation, leads remain valid instruments.

[26] Carroll (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.

Conversely, if the researcher derives an Euler condition that includes lagged dependent 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}. Because of serial correlation in y_t, y_{t+j} for any $j > 1$ is correlated with y_t and thus w_t. Now leads are no longer orthogonal to the error term. Yet lags remain valid instruments.

Although the researcher may not know *a priori* the appropriate specification of the forward-looking model, she can use tests for whether instruments are orthogonal to the error, such as Hansen's J statistic, along with the two previous results to refine her specification.

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 describing the data we employ. Section 4.2 explains that the effect of tort liability on physician supply is not theoretically identified. It could increase or decrease equilibrium supply depending on the efficiency of physician-patient contracting. Empirical work is required to identify the net effect of liability. Section 4.3 explains that we focus on punitive damage caps as a treatment because physician supply changed prior to adoption of caps. We argue that this pre-period response is better interpreted as an anticipation effect rather than as endogeneity (or a simple pre-period trend), making it a good candidate for our proposed methods of estimating 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.) 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 Data

Our analysis uses physician count data from the American Medical Association's Physician Masterfile.[27] The data are available at the state-year level, span the period 1980-2001, and include any physician with a state identifier. This includes private practitioners, hospital staff, residents, and locum tenens.[28] Military doctors are not included. There are no data available in 1984 or 1990 but otherwise the data are complete.

Physicians are categorized into one of 20 possible specialties. 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 4. 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 4 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.

Our tort reform data come from Avraham (2010).[29] 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 3 and coded as 0-1 indicator variables.

4.2 Background on tort liability and theory

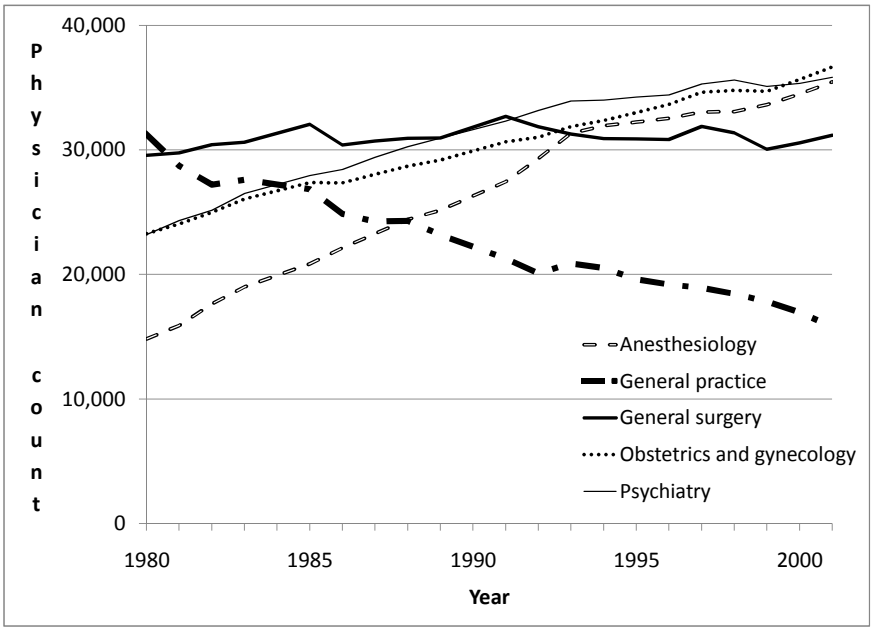
Tort liability is akin to a legally-mandated alteration of the implicit labor contract between a patient and her physician. In most cases, it requires the physician to provide the quality of care that a "reasonable physician" would provide and compensates patients who suffer injuries due to inadequate care. Compensation may include economic damages for lost wages

[27] This data is also analyzed in Klick and Stratmann (2007). We are grateful to the authors for sharing their code for cleaning the AMA data.

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

[29] Klick 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.

Figure 4: Physician supply from 1980 to 2001



Note: data for 1984 and 1990 are interpolated.

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 and broadly to deter the doctor from similar misbehavior in the future.

Tort reform refers to various changes to these mandatory contract terms. Table 3 provides a description of the most common reforms. Most of them, such as caps on punitive damages, lower the liability of doctors.[30] 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 access to care for patients. This topic has gotten substantial attention from scholars and the media.[31] Economic theory, by itself, cannot confirm or refute this concern. The effect

[30]Seventeen states imposed limits on punitive damages during our sample period. These caps either impose a specific dollar upper bound such as $250,000 on punitive awards or require that punitive awards be no more than a fraction or multiple of economic damages.

[31]See 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 Economist (2005).

of an increase in tort liability on the equilibrium quantity of physicians is theoretically ambiguous.

For example, transaction costs may prevent a patient and physician from writing a complete contract that covers all contingencies, including specific instances of malpractice. In that case, mandatory terms imposed by tort liability have two effects. They may incentivize physicians to take more care, increasing the quality of their supply. This would shift the physician demand curve outwards. But the mandatory terms may also increase the cost of each unit of physician supply because improving quality is costly and there may be litigation costs. This would shift the physician supply curve inwards. If the demand effect dominated, equilibrium physician supply would rise. But if the supply effect dominated, equilibrium physician supply would fall.

One might object that medical malpractice insurance insulates physicians from the effects of tort liability. However, such insurance covers only the financial costs of liability, not the psychic and hassle costs of litigation, which anecdotally physicians assert is large (see Mello, Chandra, Gawande, and Studdert 2010). Another objection is that health insurance insulates patients from the costs of additional quality, at least on the margin. Thus tort liability should only increase physician demand, thereby unambiguously increasing the equilibrium quantity of physicians. If physicians cannot, however, pass the costs of higher quality onto health insurers, supply may contract even if patients do not face higher costs on the margin. Physicians will simply exit and consumers will have less access. Finally, one might object that physicians face large relocation costs that block their exit. These costs, however, will not affect adjustments on the intensive margin of hours worked. Moreover, 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).[32]

Several recent studies employ a myopic model to analyze the impact of tort liability on physician supply. Kessler, Sage, and Becker (2005) perform a difference-in-differences

[32]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.

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 rural physicians by about 10%. 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.

4.3 Evidence for anticipation effects

Among the set of tort reforms that may affect physician supply, we focus on punitive damage caps because this reform is a good candidate for our model of anticipation effects. More specifically, punitive damage caps meet the following three criteria: (1) physician supply changes prior to enactment of the reform; (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 this 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.

4.3.1 Physician supply changed prior to enactment of reform

Figure 5 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.[33] 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.

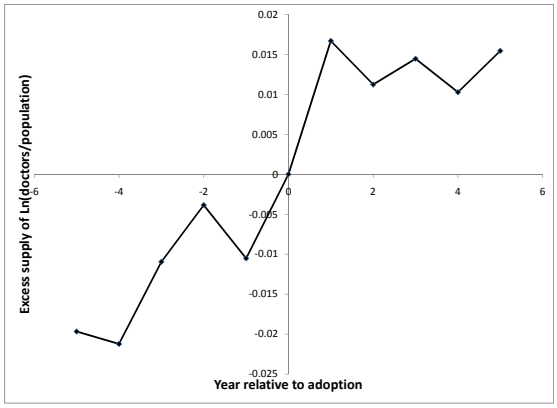
4.3.2 Reform is exogenous to physician supply

A change in outcomes prior to treatment is consistent with both anticipation effects and endogeneity. However, there are strong reasons to believe that caps on punitive damages are

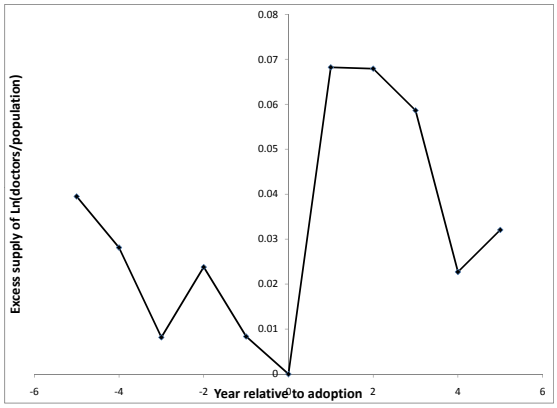
[33]Figure 5a replicates Figure 1 from the introduction.

Figure 5: Tort reforms exhibiting pre-period changes in behavior

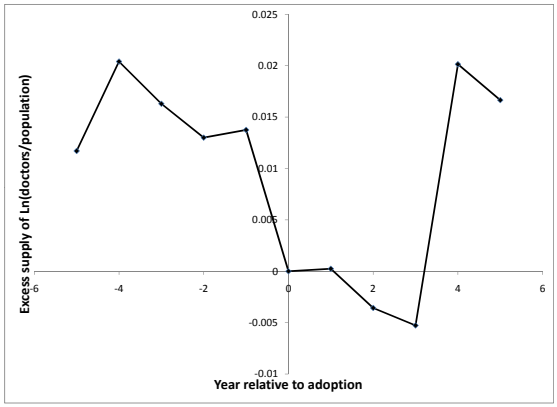
(a) Caps on punitive damages



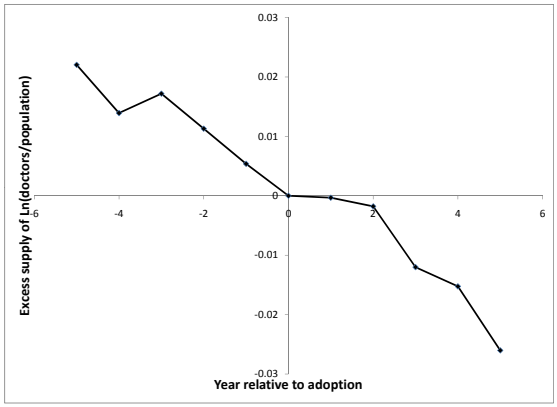
(b) Caps on total damages



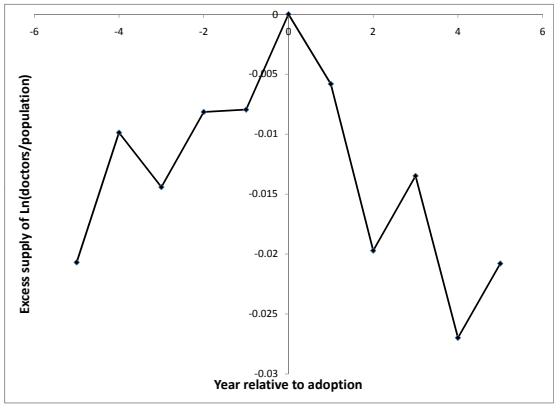
(c) Contingency fee reform



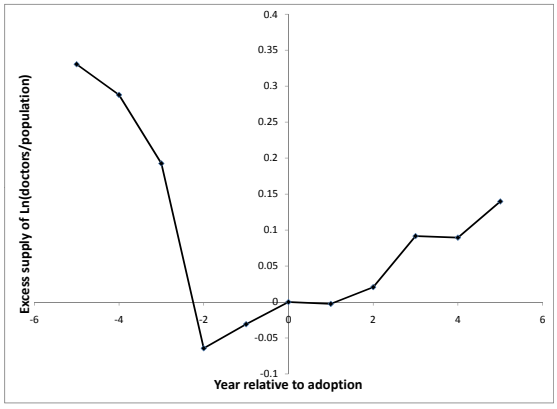
(d) Joint and several liability reform



(e) Split recovery



(f) Victims' fund



Note: These figures plot the normalized coefficients λ_j from the following regression: $\ln y_{ist} = \Sigma_{j=-5}^{5} \lambda_j D_{st+j} + \gamma X_{ist} + u_{ist}$, where y_{ist} is physician supply 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.

exogenous to physician supply, our second criteria. Unlike other reforms, punitive damage caps are targeted at all tort suits, not just medical malpractice suits.[34] This is verified in Table 5, 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.

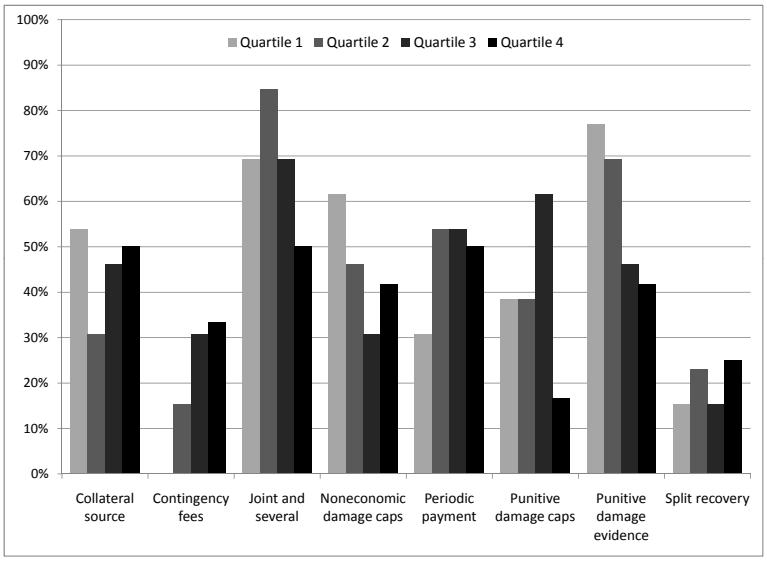
Furthermore, we can rule out specific channels by which physician supply might be thought to effect adoption of punitive damage caps. For example, one might suppose that state legislatures are public-spirited and decrease liability only when physician supply falls. Figure 5b demonstrates this phenomenon for states that adopted caps on total damages: a steady decline in physician supply is followed by a large increase once this liability-reducing reform is adopted. By contrast, Figure 5a 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 6, 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 possibility raised by the pre-post plots in Figure 5 is that pre-treatment changes in outcomes reflect pre-period trends in the treatment states rather than anticipation effects. If there were a pre-treatment trend aside from anticipation effects, however, one might expect it for all reforms in our data. Yet Figure 7 shows that the other four reforms in our data display no discernible changes in supply prior to anticipation. Moreover, Figure 5 shows that different reforms have different pre-period trends. This is true even if we examine only the reforms that reduce liability (all but Figure 5d).

[34] Currie and MacLeod (2008), which uses an older version of our tort reform data and supplements it with additional data, makes a similar argument for the exogeneity this reform.

Figure 6: 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.

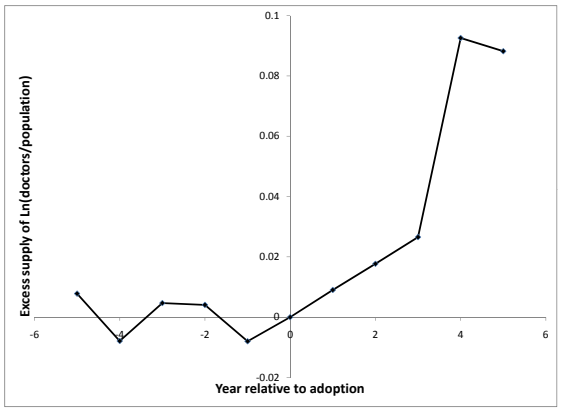
More fundamentally, the concern about pre-period trends in treatment states is just a variant of the claim that treatment was endogenous. The arguments we use to rule out the tort reforms we study as endogenous in levels can also be used to rule out endogeneity in trends. For example, it is difficult to understand why public-interest minded states would pass punitive damage caps to lower liability when physician supply was already trending upwards. Physician political power does not explain our observations either: Figure 5a suggests that – relative to the pre-trend – caps reduced physician supply. Not only is this counterintuitive (damage caps *decrease* liability, so one should expect them to increase supply), but also it implies that physicians lobbied for a reform that lowered their numbers.

We have argued that anticipation effects explain the trends we observe in Figure 5a while endogeneity does not.[35] Even if one could not rule out endogeneity in this manner, one must still justify why, among two possible explanations – endogeneity and anticipation effects – one defaults into equating pre-period trends with endogeneity. If anticipation effects are present, filtering out pre-trends with state-specific trends to address endogeneity will cause

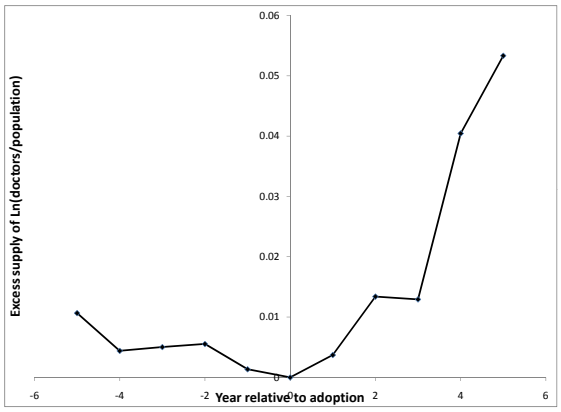
[35] We also argue in the Appendix that anticipation – but not endogeneity – explains physician supply behavior for the other two reforms we consider, joint and several reform and split recovery rules.

Figure 7: Tort reforms exhibiting no change in pre-period behavior

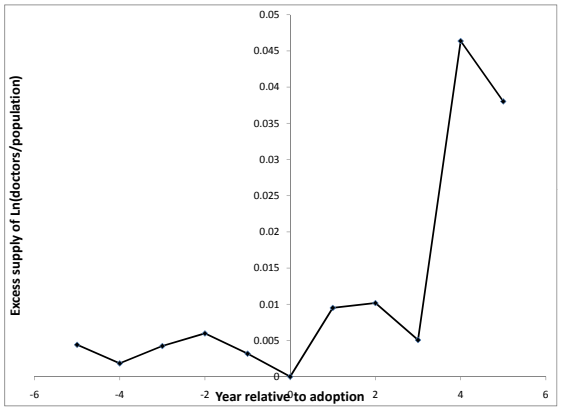
(a) Caps on noneconomic damages



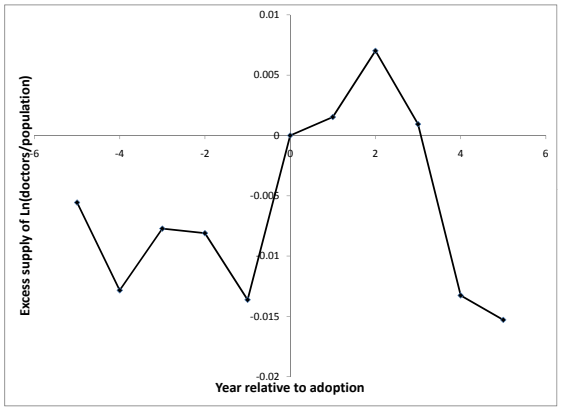
(b) Collateral source rule



(c) Periodic payment

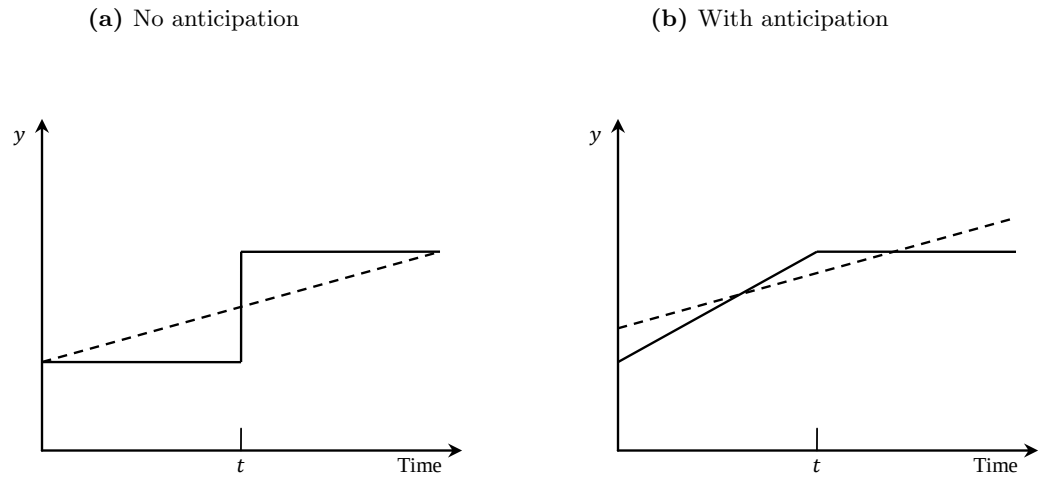


(d) Punitive damage evidence



Note: These figures plot the normalized coefficients λ_j from the following regression: $\ln y_{ist} = \Sigma_{j=-5}^{5} \lambda_j D_{st+j} + \gamma X_{ist} + u_{ist}$, where y_{ist} is physician supply 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 8: State-specific trends can bias estimation



In panel A, the state-specific trend causes the researcher to underestimate the magnitude of the treatment effect. In panel B, it causes the researcher to estimate the wrong sign of the treatment effect.

one not only to underestimate full treatment effects, but also perhaps to estimate the wrong sign on treatment effects. We illustrate this in Figure 8. Panel A displays a level increase in outcomes at time-t due to time-t treatment. It then plots a dotted state-specific trend for treatment states. It is easy to see that, even without anticipation effects, state trends reduce the magnitude of level treatment effects and, for a given level of variance in unobservables, reduce the probability that the level treatment effect will be estimated as significant, a point previously made by Wolfers (2006). Panel B modifies the treatment in Panel A by adding anticipation effects to the pre-period. Now the dotted state-specific trend will not only reduce the estimated level treatment effect, but also it may even cause the estimated effect to be negative rather than positive! The use of a state-specific pre-trend would reduce this risk, but only because it is an alternate parameterization of anticipation effects, namely a model with linear discounting rather than exponential discounting of anticipation effects. The lesson is that insertion of state-specific trends or state-specific pre-trends must have a theoretical justification, else they may introduce bias into the estimation.

4.3.3 Reform was anticipated

Ruling out endogeneity is a necessary condition for estimating anticipation effects but it may not be sufficient. We can demonstrate, however, 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 (Economist 2005). Moreover, they can be alerted to forthcoming reform through at least two possible channels: newspapers and insurance premiums.

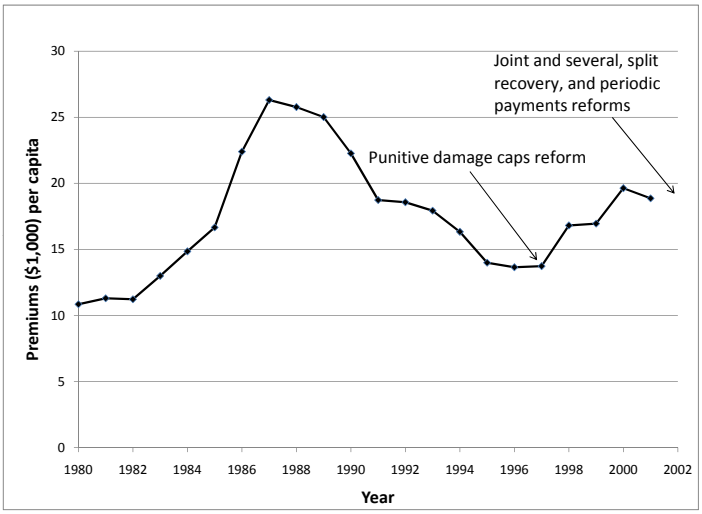
Newspaper articles discussing upcoming legislation can directly inform physicians about potential future reforms. To verify this, we searched for newspaper stories about punitive damage caps prior to adoption of that reform. For example, in Pennsylvania, a large adopter, we found over 80 articles during 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 6.3.

Even if one is skeptical that physicians are sophisticated enough to understand the impact of particular reforms on their liability exposure, it is likely that medical malpractice insurance companies are. They may then indirectly signal forthcoming reform to physicians by decreasing premiums when expected future physician liability decreases. Figure 9 displays the log of per capita premiums 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.[36]

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 physicians in the state in that year for all 50 states. Physician data for 1984 and 1990 are interpolated. Amounts are in 1984 dollars.

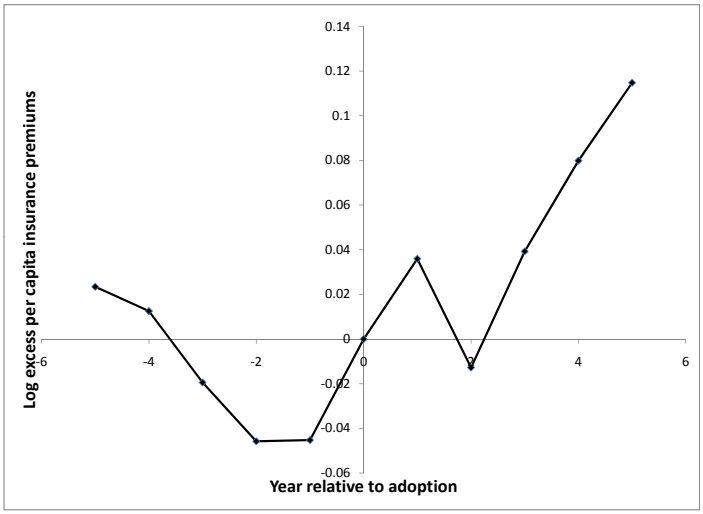
4.4 Empirical model

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 4) in states that adopt reform in a year to within-state changes in supply among states that do not adopt in that year.[37] 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

[36] Appendix 6.4 shows that corresponding pre-trends in insurance premiums also exist for joint and several and split recovery reforms).

[37] We also separately estimated our models for all four risk tiers. Including the two low-risk tiers attenuated our estimates and 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 10: Excess amount of premiums before and after reform: annual leads and lags from 5 years before to 5 years after adoption

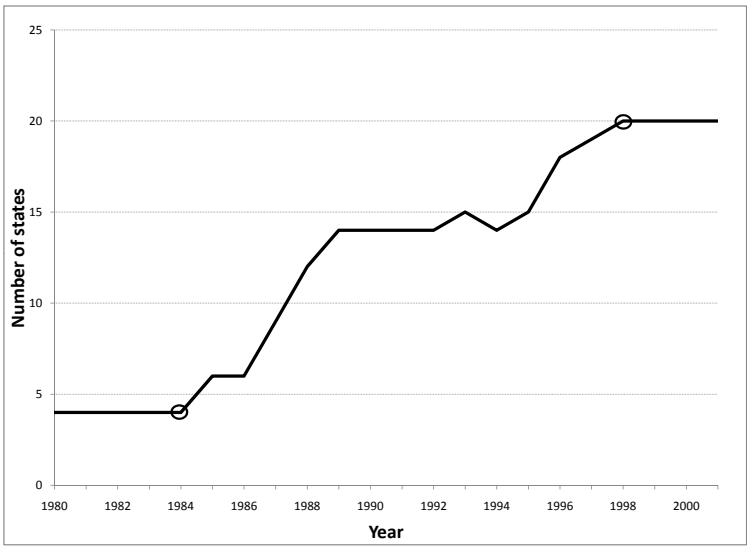


Note: premium data are from AM Best. This plot displays the normalized coefficients λ_j from the OLS regression $\ln y_{st} = \Sigma_{j=-5}^{5} \lambda_j D_{st+j} + \gamma_s + \gamma_t + u_{st}$, where y_{st} is the total amount of direct premiums earned in state s in time t divided by the number of 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. Standard errors are clustered by state.

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 4 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 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.

Figure 11: Cumulative number of states adopting punitive damage caps



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

$$\ln y_{ist} = \beta_0 d_{st} + \gamma_{is} + \gamma_{it} + u_{ist} \tag{13}$$

y_{ist} is 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 then estimate four quasi-myopic models that include up to four leading indicators for whether a law was passed:

$$\ln y_{ist} = \beta_0 d_{st} + \Sigma_{j=1}^{S} \beta_j D_{st+j} + \gamma_{is} + \gamma_{it} + u_{ist} \tag{14}$$

$S = 1 \ldots 4$ is the number of leading indicators in the regression and D_{st+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 $D_{st+1} = 1$ in period 4 and 0 otherwise. We parameterize the quasi-myopic model using treatment adoption dummies D_{st+j} rather than merely concurrent treatment dummies d_{st} so that regression coefficients directly identify the parameters of interest: The full treatment effect is estimated by $\hat{\beta}_0$ and the contemporaneous effect is estimated by $\hat{\beta}_0 - \hat{\beta}_1$.

Next we estimate a model where physician supply is modeled as a function of exponentially discounted expectations of tort reforms:

$$\ln y_{ist} = \beta d_{st} + \beta \Sigma_{j=1}^{\infty} \theta^j E_t \left[d_{st+j} \right] + \gamma_{is} + \gamma_{it} + \varepsilon_{ist}$$

We assume agents have rational expectations of future tort reforms and tort reform is exogenous:[38]

$$E_t \left[d_{st+j} \right] = d_{st+j} + v_{t,t+j}^d$$

This yields the estimable Euler equation

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

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

As we shall illustrate 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.[39] Thus, our estimable Euler equation becomes

$$\ln y_{ist} = \theta \ln y_{ist+1} + \beta d_{st} + \delta d_{st+1} + \gamma_{is} + \gamma_{it} + w_{ist} \tag{15}$$

where the regressor d_{st+1} controls for the endogeneity of d_{st}. We estimate equation (15) first using OLS, then using our proposed leads of $\ln y_{ist+1}$ as instruments, and finally using lags of $\ln y_{ist+1}$ as instruments. We employ all available instruments in each category; restricting the number of instruments does not substantively affect our results.

All our estimations weight observations by state population. Following the recommendations of Bertrand, Duflo, and Mullainathan (2004), we allow for arbitrary serial correlation

[38] We also estimated an exponential discounting model, described in Appendix 6.2, 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.

[39] 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.

in the error term as well as arbitrary cross-sectional correlation within tiers (defined in Table 4) when computing standard errors.[40] 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.[41] 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 5). We therefore exclude potentially endogenous states when performing our estimations.[42] Furthermore, we do not include other tort reforms as controls because their endogeneity could contaminate our estimates.

We perform two sets of robustness checks. First, we estimate a specification that clusters standard errors at the state rather than the state-tier level. Second, to allow for the possibility that we have been too conservative in enforcing exogeneity, we also estimate a specification that includes all states and controls for other tort reforms.

[40] Recall that we only estimate treatment effects for tiers 1 and 2 from Table 4. Thus if we include all 50 states we would have $50 \times 2 = 100$ clusters. We cluster at the tier level rather than the state level for two reasons. First, we cannot think of an unobserved variable that is correlated across tiers within a state after controlling for correlations across specialties within a tier. Allowing wider correlations than economically justified may cause one to overestimate standard errors. The second reason is that, because we exclude some states from our analysis, clustering at the state level results in fewer than 50 clusters, which may be suboptimal (Angrist and Pischke 2008 p. 323). Nevertheless, we also cluster at the state level as a robustness check (see Section 4.5).

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

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

4.5 Results

Table 6 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 full effect of reform, including anticipation effects, and can be interpreted as relative changes in physician supply. Column 1 estimates that punitive damage caps reduced physician supply by 3.9%. Moving across the first row reveals that the full effect of reform monotonically increases from 3.9% to 5.0% as we add leads. All estimates are strongly significant.

Next we turn to estimates for our exponential discounting model. Under rational expectations, the regression model is given by the Euler equation (15). Column 1 of Table 7 reports OLS estimates of this equation. These estimates, although statistically significant, are inconsistent because OLS estimation of equation (15) does not account for the correlation between the error term and y_{ist+1}. Column 2 estimates the Euler equation using leads of y_{ist+1} as instruments for y_{ist+1}. The estimated contemporaneous and full effects of punitive damage caps are significant at 2.2% and 5.7%, respectively. Finally, column 3 reports results when we uses lags of y_{ist} rather than leads to instrument for y_{ist+1}. The estimated effects similar to those estimated using leads as instruments.

Discussion of results

Specification 1 in Table 8 summarizes the results from Tables 6 and 7. 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 full effect of reform, including anticipation effects.[43] Including leads reduces this bias by reducing omitted variable bias. Table 8 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 estimates of the full effects increase as we keep adding leads to the quasi-myopic model suggests that each additional lead moves us closer to an estimate of the full effect.

Combining this with our result that the exponential discounting model yields even larger

[43] This assumes contemporaneous treatment effects and anticipation effects have the same sign and that current treatment is positively correlated with future treatment, on average.

estimates of the full effect of tort reform strongly suggests that anticipation effects matter and for perhaps longer than our data permit in 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.

Robustness checks

Specification 2 in Table 8 reports results when we cluster by state rather than state-tier. This does not significantly affect the strength of our results. Specification 3 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 full effect for this reform as we add leads to the quasi-myopic model.

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 that individuals are forward looking when purchasing durable goods such as cars or houses or making human capital investments, for example, and that firms are forward looking when investing in physical capital or entering new markets. 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 full effect of the tort reforms we study are double that suggested by a myopic model.

The framework has some limitations, each of which is a potential topic for future research. Perhaps outcomes are not additively separable in each period's expectations. We offer no formal way to discriminate between the two sets of parametric restrictions 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 simply survey agents about their expectations or at least survey a subsample to empirically estimate the relationship between expectations and unobservables. 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.

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.

Angrist, J. D. and J.-S. Pischke (2008, December). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

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. (2010). Database of State Tort Law Reforms (DSTLR 3rd). *SSRN eLibrary*.

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.

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.

Blake, D. (1991, September). The estimation of rational expectations models: A survey. *Journal of Economic Studies 18*(3), 31–70.

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.

Chow, G. C. (1989, August). Rational versus adaptive expectations in present value models. *The Review of Economics and Statistics 71*(3), 376–84.

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.

Economist (2005, December 17). Scalpel, scissors, lawyer. *The Economist*.

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.

Gersen, J. and E. Posner (2007). Timing rules and legal institutions. *Harv. L. Rev. 121*, 543.

Gruber, J. and B. Koszegi (2001, November). Is addiction "rational"? theory and evidence. *The Quarterly Journal of Economics 116*(4), 1261–1303.

Hansen, L. P. (1982, July). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–54.

Helland, E., J. Klick, and A. Tabarrok (2005, Spring). Data watch: Tort-uring the data. *Journal of Economic Perspectives 19*(2), 207–220.

Huber, B. (2011). Transition Policy in Environmental Law. *Harv. Env. L. J. 35*.

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.

Maltzman, F. and C. Shipan (2008). Change, Continuity, and the Evolution of the Law. *American Journal of Political Science*, 252–267.

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.

Mello, M. M., A. Chandra, A. A. Gawande, and D. M. Studdert (2010). National Costs Of The Medical Liability System. *Health Aff 29*(9), 1569–1577.

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. (2009). How to do xtabond2: An introduction to difference and system gmm in stata. *Stata Journal 9*(1), 86–136.

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*.

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.

6 Appendix

6.1 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 $v^d_{t,t+j}$.[44] If treatment states are correlated over time then endogeneity occurs because $E[d_t v^d_{t,t+j}] \neq 0$ and the error term will be serially

[44]Recall that $v^d_{t,t+j}$ is defined as the forecast error from the time-t forecast of d_{t+j}

correlated, violating Assumption A2′ from Section 3.2.1. 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

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

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

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

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

6.2 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:

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

Plugging these equations into equation (5) and simplifying yields

$$y_t = \theta\phi y_t + \theta(1-\phi) E_{t-1}\left[y_t\right] + \beta d_t + \varepsilon_t \tag{16}$$

The one-step back version of equation (5) is:

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

Solve this for $E_{t-1}\left[y_t\right]$ and plug the result into (16). 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.

6.3 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 6.4) 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 articles pertaining to the reform in question.[45] 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

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

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).

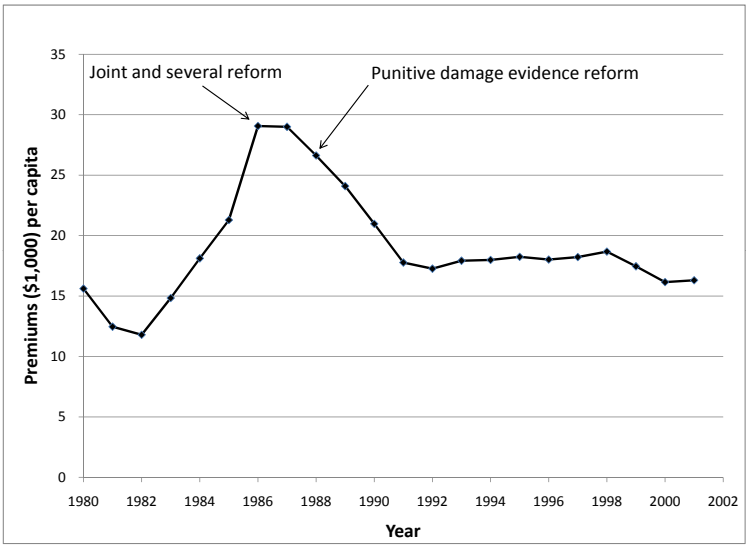
6.4 Results for joint and several reform and split recovery

In this appendix we examine two additional tort reforms: joint and several liability reform split recovery reform. The doctrine of joint and several 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 5d and 5e 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 6, which plots the fraction of states that ever adopt different tort reforms by quartile of physician supply in 1980.

Appendix 6.3 provided evidence that these two reforms were discussed in newspapers

Figure 12: Per capita insurance premiums for California



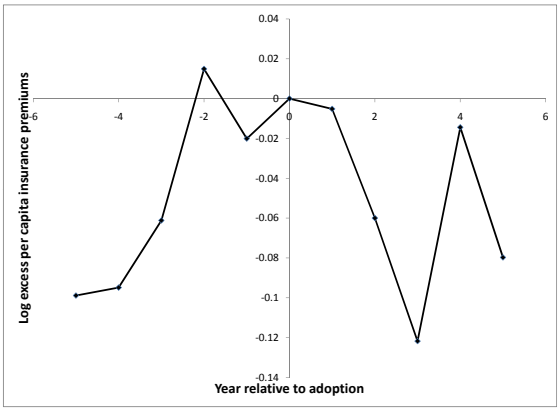
Note: Premium data are from AM Best. These plots display the direct premiums earned in a given calendar year divided by the number of physicians in the state in that year for all 50 states. Physician data for 1984 and 1990 are interpolated. Amounts are in 1984 dollars.

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 5d. 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 5e.

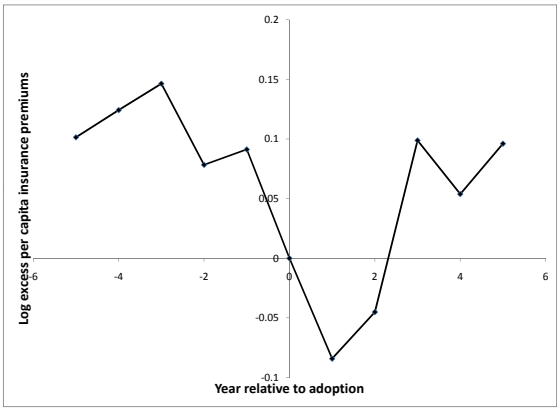
Appendix Tables 9 – 12 present our estimation results and Appendix Table 13 summarizes them. Our estimates for joint and several reform are strongly significant while those for split recovery are moderately significant. Table 13 shows that, just as with punitive

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

(a) Joint and several reform



(b) Split recovery reform



Note: premium data are from AM Best. This plot displays the normalized coefficients λ_j from the OLS regression $\ln y_{st} = \Sigma_{j=-5}^{5} \lambda_j D_{st+j} + \gamma_s + \gamma_t + u_{st}$, where y_{st} is 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. Standard errors are clustered by state.

damage caps, the estimated effects increase as we add leads in the quasi-myopic model and are largest overall for the exponential discounting model.

We also estimated alternative specifications for these reforms where we (1) clustered standard errors by state rather than tier and (2) included all states and included other reforms as controls. Those results, available upon request, show that our estimates for these two reforms are robust to clustering by state. Estimates for split recovery reform, however, become insignificant and nearly zero when we include all states and include other reforms as controls.

Table 1: Summary of endogeneity problems that arise under different formulations of rational expectations.

Formulation of rational expectations	Variable that is forecasted	
	Outcome ($z = y$)	Treatment ($z = d$)
Realization correlated with forecast error: $z = E[z] + v$ $\rightarrow E[zv] \neq 0$	1) Endogeneity because $E[y_{t+1} v^y_{t,t+1}] \neq 0$ 2) If $corr(d_t d_{t+1}) \neq 0$, then also endogeneity because $E[d_{t+1} v^y_{t,t+1}] \neq 0$	1) Endogeneity because $E[y_{t+1} e_{t+1}] \neq 0$ 2) If $corr(d_t d_{t+1}) \neq 0$, then also endogeneity because $E[d_{t+1} v^d_{t,t+1}] \neq 0$
Realization not correlated with forecast error: $E[z] = z + v$ $\rightarrow E[zv] = 0$	No endogeneity because $E_t[y_{t+1} v^y_{t,t+1}] = 0$ & $E_t[d_t v^y_{t,t+1}] = 0$	Endogeneity because $E[y_{t+1} e_{t+1}] \neq 0$ & $E[y_{t+1} E_t[d_{t+1}]] \neq 0$ $\rightarrow E[y_{t+1} v^d_{t,t+1}] \neq 0$

Table 2: Summary of IVs (drawn from the agent's information set) that should be employed under different formulations of rational expectations.

Formulation of rational expectations	Variable that is forecasted	
	Outcome ($z = y$)	Treatment ($z = d$)
Realization correlated with forecast error: $z = E[z] + v$ $\rightarrow E[zv] \neq 0$	1) IV for $E_t[y_{t+1}]$ with Ω_{1t} 2) If $corr(d_t d_{t+1}) \neq 0$, then also IV for d_t perhaps with d_{t+1} and assume $E[\hat{d}_t v^y_{t,t+1}] = 0$ (see appendix)	1) IV for y_{t+1}, with Ω_{1t} 2) If $corr(d_t d_{t+1}) \neq 0$, then also IV for d_t perhaps with d_{t+1} (see appendix)
Realization not correlated with forecast error: $E[z] = z + v$ $\rightarrow E[zv] = 0$	No IV required	IV for y_{t+1} with Ω_{1t}

7 Tables

Table 3: 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 4: 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 5: 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 6: Myopic and quasi-myopic (QM) estimates for punitive damage caps

	Number of leads				
Tort reform	(0)	(1)	(2)	(3)	(4)
Punitive damage caps	0.039**	0.042**	0.045**	0.050**	0.050**
	(0.015)	(0.016)	(0.018)	(0.020)	(0.020)
Lead (t+1)		0.012	0.015	0.020	0.020
		(0.008)	(0.010)	(0.012)	(0.013)
Lead (t+2)			0.008	0.014	0.014
			(0.009)	(0.010)	(0.011)
Lead (t+3)				0.015*	0.015
				(0.007)	(0.009)
Lead (t+4)					-0.000
					(0.005)
Model	Myopic	QM	QM	QM	QM
Observations	6,363	6,363	6,363	6,363	6,363
R^2	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-risk tier. A */** next to the coefficient indicates significance at the 10/5% level.

Table 7: Exponential discounting model estimates for punitive damage caps

Tort reform	(1)	(2)	(3)
Punitive damage caps (full effect)	0.046**	0.057**	0.064**
	(0.016)	(0.015)	(0.023)
Punitive damage caps (one-period effect)	0.015**	0.022**	0.022**
	(0.005)	(0.007)	(0.006)
Discount rate ($\widehat{\theta}$)	0.665**	0.622**	0.654**
	(0.052)	(0.096)	(0.086)
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	0.769

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-risk tier. A */** next to the coefficient indicates significance at the 10/5% level. Full effect is defined as the one-period effect divided by one minus the discount rate. The AR(3) test checks for order-3 serial correlation in the residuals.

Table 8: Summary of estimated full effects for punitive damage caps

		Specification		
Model	IV	(1)	(2)	(3)
Myopic	None	0.039**	0.039**	0.045**
Quasi-myopic (1 lead)	None	0.042**	0.042**	0.046**
Quasi-myopic (2 leads)	None	0.045**	0.045*	0.050**
Quasi-myopic (3 leads)	None	0.050**	0.050**	0.056**
Quasi-myopic (4 leads)	None	0.050**	0.050*	0.059**
Exponential discounting	Lags	0.064**	0.064**	0.049
Exponential discounting	Leads	0.057**	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 6 and 7. Specification 2 clusters standard errors by state instead of by tier. Specification 3 includes all 50 states in the analysis and adds other (potentially endogenous) tort reforms as controls.

Appendix tables

Table 9: Myopic and quasi-myopic (QM) estimates for joint and several

	Number of leads				
Tort reform	(0)	(1)	(2)	(3)	(4)
Joint and several	-0.028**	-0.032**	-0.037**	-0.038**	-0.040**
	(0.013)	(0.014)	(0.015)	(0.016)	(0.016)
Lead (t+1)		-0.014**	-0.021**	-0.023**	-0.024**
		(0.006)	(0.008)	(0.009)	(0.009)
Lead (t+2)			-0.026**	-0.028**	-0.029**
			(0.011)	(0.012)	(0.012)
Lead (t+3)				-0.004	-0.005
				(0.007)	(0.008)
Lead (t+4)					-0.004
					(0.008)
Model	Myopic	QM	QM	QM	QM
Observations	5,473	5,473	5,473	5,473	5,473
R^2	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-risk tier. A */** next to the coefficient indicates significance at the 10/5% level.

Table 10: Myopic and quasi-myopic (QM) estimates for split recovery

	Number of leads				
Tort reform	(0)	(1)	(2)	(3)	(4)
Split recovery	0.036*	0.036	0.035	0.036	0.042*
	(0.021)	(0.022)	(0.023)	(0.023)	(0.024)
Lead (t+1)		0.003	0.002	0.003	0.008
		(0.013)	(0.014)	(0.014)	(0.014)
Lead (t+2)			-0.004	-0.004	0.002
			(0.011)	(0.011)	(0.011)
Lead (t+3)				0.004	0.010
				(0.008)	(0.010)
Lead (t+4)					0.024*
					(0.013)
Model	Myopic	QM	QM	QM	QM
Observations	8,965	8,965	8,965	8,965	8,965
R^2	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-risk tier. A */** next to the coefficient indicates significance at the 10/5% level.

Table 11: Exponential discounting model estimates for joint and several

Tort reform	(1)	(2)	(3)
Joint and several (full effect)	-0.036**	-0.041**	-0.067*
	(0.017)	(0.016)	(0.035)
Joint and several (one-period effect)	-0.013**	-0.015**	-0.013**
	(0.005)	(0.006)	(0.005)
Discount rate ($\widehat{\theta}$)	0.641**	0.642**	0.802**
	(0.043)	(0.085)	(0.069)
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)		0.228	0.609

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-risk tier. A */** next to the coefficient indicates significance at the 10/5% level. Full effect is defined as the one-period effect divided by one minus the discount rate. The AR(3) test checks for order-3 serial correlation in the residuals.

Table 12: Exponential discounting model estimates for split recovery

Tort reform	(1)	(2)	(3)
Split recovery (full effect)	0.050*	0.041**	0.061*
	(0.026)	(0.021)	(0.036)
Split recovery (one-period effect)	0.014**	0.015**	0.013**
	(0.007)	(0.007)	(0.006)
Discount rate ($\widehat{\theta}$)	0.723**	0.622**	0.791**
	(0.042)	(0.096)	(0.066)
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-risk tier. A */** next to the coefficient indicates significance at the 10/5% level. Full effect is defined as the one-period effect divided by one minus the discount rate. The AR(3) test checks for order-3 serial correlation in the residuals.

Table 13: Summary of estimated full effects for joint and several (JS) and split recovery (SP)

		Reform	
Model	IV	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-risk tier.