Submitted to the Annals of Applied Statistics

PREDICT, CORRECT, SELECT: A GENERAL STRATEGY TO IDENTIFY CAUSAL EFFECTS OF GUN POLICY CHANGES

BY THOMAS LEAVITT[1,a], AND LAURA A. HATFIELD[2,b]

[1]*Marxe School of Public and International Affairs, Baruch College, City University of New York (CUNY),*
[a]*Thomas.Leavitt@baruch.cuny.edu*

[2]*Department of Health Care Policy, Harvard Medical School,* [b]*hatfield@hcp.med.harvard.edu*

Whether policies that expand access to firearms decrease or increase crime is a question of fierce debate. To address it, researchers often use a controlled pre-post design in which they compare over-time changes in crime in a population exposed to a change in firearm law to an unexposed comparison group. With some counterfactual assumptions, this enables causal conclusions about the effects of gun laws. However, these empirical investigations have reached varying conclusions depending on the specifics of their methods. The policy debate is therefore stymied by disagreements over the "correct" causal model. In this paper, we propose a novel identification framework that offers a way to settle the model specification debates. We propose to use models that predict untreated outcomes and correct the treated group's predictions using the comparison group's observed prediction errors. Our identifying assumption is that the treated and comparison groups would have equal prediction errors (in expectation) under no treatment. To select the best prediction model, we propose a data-driven procedure that is motivated by design sensitivity. We choose the prediction model that is most robust to violations of the identification assumption by observing the differential average prediction errors in the pre-period. Our approach offers a way out of the debate over the "correct" model by choosing the most robust model instead. It also has the desirable property of being feasible in the "locked box" of pre-period data only and accommodates the range of prediction models that applied researchers employ. We use our procedure to select from a set of candidate models and estimate the effect on homicide of Missouri's 2007 repeal of its permit-to-purchase law.

1. Introduction. Opposite sides of the gun control debate claim that increased firearm access either reduces crime or exacerbates crime. To test these ideas empirically, we can study how crime changes after gun policy changes, perhaps contrasting the changes in an exposed population to the changes in an unexposed comparison group. Such analyses yield causal conclusions under assumptions about how crime would have evolved in the two populations absent the gun policy change. For example, difference-in-differences (DID) assumes crime would have evolved in parallel and comparative interrupted time series assumes similar evolution of parameters in a linear model. We call these "controlled pre-post designs" to emphasize that they leverage pre- to post-intervention changes and a control group.

In this paper, we apply controlled pre-post designs to study how homicide rates changed after Missouri repealed its permit-to-purchase law in 2007. The law, in place since 1921, had required people purchasing handguns from private sellers to obtain a license that verified the purchaser had passed a background check. We compare changes in Missouri's homicide rate to changes in eight bordering states that did not repeal their permit-to-purchase laws (see Figure 1).

Keywords and phrases: causal inference, difference-in-differences, estimation, longitudinal analysis, predictive models, robustness.

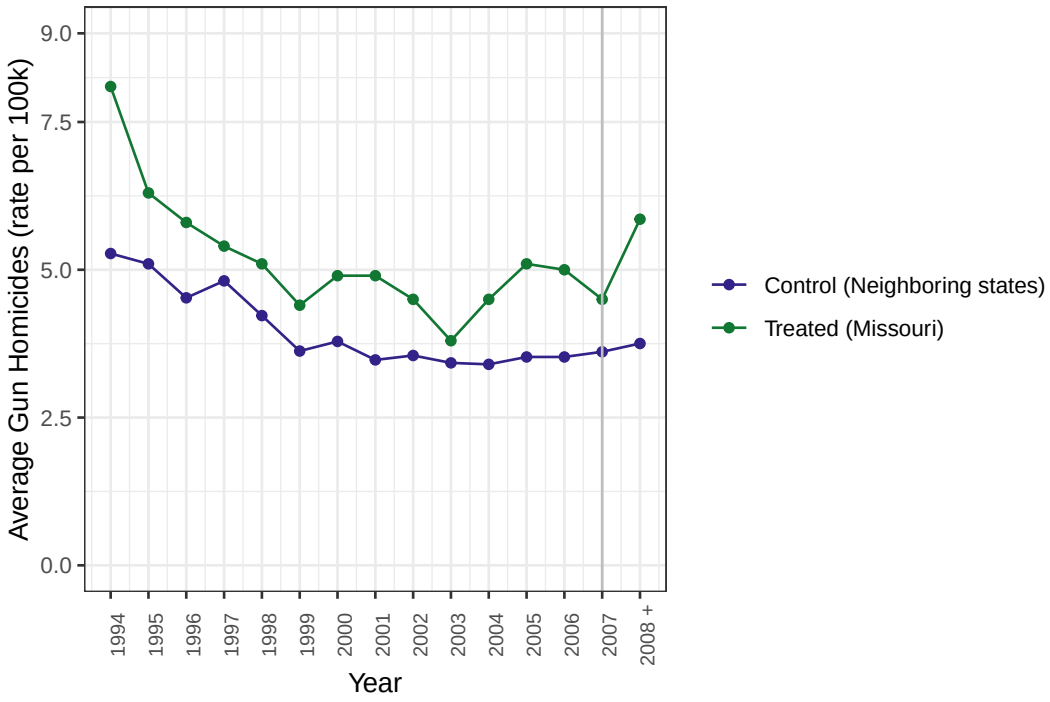

FIG 1. *Average gun homicides (rate per 100,000) before and after the 2007 permit-to-purchase repeal in Missouri (treated state) and control states (Arkansas, Illinois, Iowa, Kansas, Kentucky, Nebraska, Oklahoma, and Tennessee)*

To choose among the various controlled pre-post designs, conventional wisdom holds that we should choose the one that relies on the most plausible assumptions (Roth and Sant'Anna, 2023; Lopez Bernal, Soumerai and Gasparrini, 2018; Ryan, Burgess and Dimick, 2015; Kahn-Lang and Lang, 2020). However, reasonable people may disagree about which model is most plausible or "correct". It is impossible to establish the "correctness" of *any* set of causal assumptions (and here, modeling assumptions amount to causal assumptions), so researchers tend to use methods that are popular in their disciplines. For instance, comparative interrupted time series dominates in education policy research, while DID is more popular in health policy research (Fry and Hatfield, 2021).

Disagreement over analysis choices is not merely academic; it impedes progress on the policy front. A 2004 report by the National Research Council concluded that "it is not possible to reach any scientifically supported conclusion because of the sensitivity of the empirical results to seemingly minor changes in model specification" (National Research Council of the National Academies, 2005, p. 151). More recent syntheses of gun policy research have reached similar conclusions (Morral et al., 2018; Smart et al., 2020).

Exemplifying this diversity of analyses and conclusions, at least two previous papers have studied the effect of Missouri's permit-to-purchase repeal on homicides. Webster, Crifasi and Vernick (2014) fit a Poisson regression model with unit and time fixed effects, while Hasegawa, Webster and Small (2019) used a non-parametric difference-in-differences estimator. Each argued that their assumptions were plausible and their conclusions correct.

In this paper, we move away from the question of model "correctness" and focus on another criterion: robustness. Robustness in our framework is the (lack of) change in our estimand under violations of the counterfactual assumptions. Our method proceeds in three steps: predict, correct, select. First, using data from the period before the policy change, we

train a model that *predicts* untreated outcomes. Then, to account for time-varying shocks that affect both groups, we use the comparison group's prediction errors to *correct* the treated group's predictions after the policy change. Third, using a validation set of pre-policy data, we *select* among competing models to maximize robustness. Since the most robust model in the population is unknown, our selection process is decidedly Bayesian in that we "select" a weighted combination of the candidate models with respect to their posterior probabilities that they are optimal. Finally, using the corrected predictions from our "selected" model, we estimate our causal target quantity, the average effect of treatment on the treated (ATT). The key causal assumption is that the prediction errors would be equal (in expectation) in treated and comparison groups, absent the policy change, or that this counterfactual difference in prediction errors is bounded by some magnitude of the worst pre-period (expected) difference in prediction errors.

This identification strategy accommodates a wide variety of prediction models; in fact, we show that by careful choice of prediction model, we can reproduce many familiar "brand name" designs. For instance, difference-in-differences is a special case when the prediction model is simply the pre-period group mean. However, we can consider a wide variety of potential prediction models without requiring bespoke identifying assumptions for each model. This is because our procedure recasts the usual choice among causal assumptions (competing on plausibility) as a choice among prediction models (competing on robustness).

Our conception of robustness builds on Manski and Pepper (2018); Rambachan and Roth (2023) who formalize an idea that is implicit in pre-period parallel trends tests (Granger, 1969; Angrist and Pischke, 2008; Roth, 2022; Egami and Yamauchi, 2022): departures from the assumed causal model in the pre-period inform us about violations in the post-period. Since the true relationship between untreated outcomes in the two periods is unknown, these sensitivity analyses take observed departures in the pre-period, assume a relationship to departures in the post-period, and observe the impact on the estimate of interest. A sensitive procedure's conclusions will be undermined at less severe violations than a robust one's.

Other authors have also developed methods for causal inference from longitudinal data and applied them to study gun/policing policies and violence/crime outcomes. With a similar focus on prediction models, Antonelli and Beck (2023) use Bayesian spatio-temporal models to produce posterior predictive distributions for unit-specific treatment effects, focusing on heterogeneous treatment effects in a staggered adoption setting. Ben-Michael et al. (2023) use multitask Gaussian process models to do causal inference in panel data with one treated unit and count outcomes, contributing to the literature in synthetic controls. Our methodology is related in that we use a Bayesian approach to selection of the unknown, optimal model (in terms of robustness) for generating predictions.

In the rest of this paper, we elaborate on our approach to controlled pre-post designs applied to gun policy evaluation. Section 2 details our general identification strategy and establishes that the assumptions of some popular designs can be considered special cases of our framework. In Section 3, we introduce a sensitivity analysis framework that motivates our model selection procedure. Section 4 describes our proposed estimation and inference procedures. We implement our methods to estimate the effect of Missouri's PTP repeal on homicide in Section 5. Finally, we conclude in Section 6 and point to open questions for future research.

2. General identification strategy. Suppose a population-level data generating process with two groups, a treated group ($G = 1$) and comparison group ($G = 0$), as well $t = 1, \ldots, T$ periods of which T is the only post-treatment period. That is, between periods $T - 1$ and T, the treated group is exposed to treatment and the comparison group is not. Let the treatment indicator in period t be $D_t = G\mathbb{1}\{t = T\}$, where $\mathbb{1}\{\cdot\}$ is the indicator function that equals 1

if its argument is true and 0 if not. For the treated group, $D_t = 0$ for all $t < T$ and $D_T = 1$. For the comparison group, $D_t = 0$ for all periods.

We use potential outcomes to define our causal target. Let $Y_t(0)$ denote the untreated potential outcome in period $t = 1, \ldots, T$ and $Y_T(1)$ denote the treated potential outcome in the post-treatment period, T. Our causal target is the ATT,

$$\text{ATT} \coloneqq \mathrm{E}_{\mathcal{P}}\left[Y_T(1) \mid G = 1\right] - \mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G = 1\right], \tag{1}$$

where $\mathrm{E}_{\mathcal{P}}[\cdot]$ denotes expectation with respect to a population-level joint cumulative distribution function.

To express the ATT in terms we can estimate from data, we need assumptions about how potential outcomes relate to observable quantities. The first such assumption is consistency between potential outcomes and the observed outcome, Y_t.

ASSUMPTION 1 (Consistency). For $t = 1, \ldots, T$,

$$Y_t = D_t Y_t(1) + (1 - D_t)\, Y_t(0). \tag{2}$$

Assumption 1 ensures that the observed outcome at a given time is the potential outcome corresponding to the treatment condition at that time. This rules out treatment anticipation (i.e., the treated group manifests treated outcomes before treatment begins) and spillovers/interference (i.e., the untreated group manifests treated potential outcomes).

With Assumption 1, we can express the ATT as

$$\text{ATT} = \mathrm{E}_{\mathcal{P}}\left[Y_T - Y_T(0) \mid G = 1\right], \tag{3}$$

replacing the treated potential outcome with the observed outcome since the treated potential outcome can be observed in the post-period. It remains to replace the (unobservable) untreated potential outcome with an observable quantity.

Suppose we predict the untreated potential outcome in period t, $Y_t(0)$, via $f(\boldsymbol{X}_t)$, where f is a model belonging to class of candidate models $\mathcal{F}$ and $\boldsymbol{X}_t$ is the collection of predictors for untreated potential outcomes in period t.[1] The predictors of untreated potential outcomes in period t are quantities whose values are determined prior to t. When we have only one post-period, T, "prior to" T and "pre-period" are the same. For extensions to multiple post-treatment outcomes, we must also limit the inputs to the prediction model to quantities whose values are determined prior to the start of treatment. If the prediction function were perfect, the ATT could be identified as

$$\text{ATT} = \mathrm{E}_{\mathcal{P}}\left[Y_T - f\left(\boldsymbol{X}_T\right) \mid G = 1\right].$$

This identification assumption is the basis for single interrupted time series designs (e.g., Wagner et al., 2002; Bloom, 2003; Zhang and Penfold, 2013; McDowall, McCleary and Bartos, 2019; Shadish, Cook and Campbell, 2002).

However, untreated outcomes are subject to shocks that f cannot predict (Britt, Kleck and Bordua, 1996). Therefore, we rely on an identification assumption that uses the comparison group to inform us about what our prediction model misses. We assume that a model's prediction errors are equal in the treated and comparison groups (in expectation) or, said another way, that unexpected shocks affect both groups' outcomes equally.

[1] Since the collection of predictors may depend on the model f, we should index the predictors $\boldsymbol{X}_t$ by the corresponding model; however, we leave this dependence implicit in our notation since the corresponding model should be clear from context.

ASSUMPTION 2 (Equal expected prediction errors).

$$\mathrm{E}_{\mathcal{P}}\left[Y_T(0) - f\left(\boldsymbol{X}_T\right) \mid G = 1\right] = \mathrm{E}_{\mathcal{P}}\left[Y_T(0) - f\left(\boldsymbol{X}_T\right) \mid G = 0\right]. \tag{4}$$

The following theorem establishes that, with this additional assumption, we can identify the ATT.

THEOREM 1 (Causal identification by equal expected prediction errors). *If Assumptions 1 and 2 hold, then the ATT in Eq. (1) is identified as*

$$E_{\mathcal{P}}\left[Y_T - f\left(\boldsymbol{X}_T\right) \mid G = 1\right] - E_{\mathcal{P}}\left[Y_T - f\left(\boldsymbol{X}_T\right) \mid G = 0\right]. \tag{5}$$

The proof, given in Section 1.1 of the online supplement, is straightforward, by linearity of expectation and substitution of observed outcomes using Assumptions 1 and 2.

2.1. *Existing designs as special cases.* Under what circumstances would equal expected prediction errors hold? It turns out that several popular non-parametric identification assumptions and structural causal models imply Assumption 2. That is, we show that when these assumptions hold, Assumption 2 will also hold, for particular choice of prediction model. We consider two such situations below and detail several more in online supplement.

2.1.1. *Nonparametric identifying assumptions.* Difference-in-differences (DID) is a popular method for observational causal inference in fields such as economics, health policy, education policy, and political science. The methodological literature has exploded recently, with research on assessing causal assumptions (Roth, 2022; Bilinski and Hatfield, 2020; Freyaldenhoven, Hansen and Shapiro, 2019), matching estimators (Ham and Miratrix, 2022; Daw and Hatfield, 2018; Lindner and McConnell, 2019; Basu and Small, 2020), treatment mis-classification (Denteh and Kédagni, 2022), assumption violations (Chan and Kwok, 2022; Marcus and Sant'Anna, 2021), and extensions to new outcome types (Liu, Wang and Xu, 2024; Graves et al., 2022). (See Roth et al. 2023 for a review of many recent developments).

In the literature, identification may be shown either using nonparametric assumptions or structural models. We regard difference-in-differences as "design-based" (Angrist and Pischke, 2010, p. 14) and thus use a non-parametric identification assumption to show that our identification assumption also holds given a careful choice of prediction function.

We use a simple case in which there are two groups (treated and comparison) and two periods (pre-period $T-1$ and post-period T). DID's crucial counterfactual assumption is that untreated potential outcomes would have evolved in parallel in the two groups:

$$\begin{aligned}\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G = 1\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G = 1\right] = \\ \mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G = 0\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G = 0\right].\end{aligned} \tag{6}$$

In the Appendix (Section A.1), we show that if parallel trends holds, Assumption 2 does also, for prediction model $f\left(\boldsymbol{X}_T\right) = Y_{T-1}$.

Of course, this is only the simplest example of a DID strategy. We can extend to more cases by conditioning on covariates (in both the assumption of parallel trends and the prediction model) or, as shown in the online supplement, using alternative assumptions such as those of sequential difference-in-differences (see Section 2.1).

2.1.2. *Structural models.* Suppose that we have multiple outcome measurement occasions in the pre- and post-intervention periods and multiple units in the treated and comparison groups. In this case, researchers often fit two-way fixed effects (TWFE) linear regression models, where "two-way" refers to unit and time fixed effects (de Chaisemartin and D'Haultfœuille, 2023). The models contain an interaction between an indicator of the post-intervention period and treated group, the coefficient of which is interpreted as an estimator of the ATT. This approach may be justified by the equivalence of the TWFE estimator and the DID estimator above (Angrist and Pischke, 2008; Egami and Yamauchi, 2022; Imai and Kim, 2019; Kropko and Kubinec, 2020; Sobel, 2012; Wooldridge, 2005), which leads to the popular impression that TWFE model identification is also by a parallel trends assumption. However, the equivalence does not extend to the more general setting. Imai and Kim (2021) show that the TWFE model's promise of simultaneous adjustment for unobserved unit and time confounders depends crucially on linearity and additivity. Moreover, several papers have described problems with using TWFE regression estimators in the setting of staggered treatment timing and treatment effect heterogeneity (Goodman-Bacon, 2021; Borusyak, Jaravel and Spiess, 2022; de Chaisemartin and D'Haultfœuille, 2023; Sun and Abraham, 2021). Kropko and Kubinec (2020) showed that while one-way fixed effects cleanly capture either over-time or cross-sectional dimensions, the TWFE model unhelpfully combines within-unit and cross-sectional variation.

Therefore, we assume that identification of TWFE models is via the following structural model:

$$Y_{u,t}(0) = \alpha_u + \gamma_t + \epsilon_{u,t} \tag{7}$$

in which $\mathrm{E}_{\mathcal{P}}[\epsilon_{u,t} \mid \alpha_u, G_u] = 0$, where $u = 1, \ldots, U$ indexes units and $t = 1, \ldots, T$ indexes periods. We show in the Appendix (see Section A.2) that when this structural model holds, Assumption 2 also holds if the prediction model is $f(\boldsymbol{X}_T) = \arg\min_{\alpha_u} \sum_{t=1}^{T-1} (Y_{u,t} - \alpha_u)^2$. This prediction model is the population-level ordinary least squares solution to the unit fixed effects model's objective function fit to data before period T, which is equivalent to the mean of a unit's outcomes in all pre-treatment periods.

Again, this is a simple instance of a TWFE structural model. In the online supplement, we also show that this idea extends to similar structural models that include unit- or group-specific time trends (see Section 2.2) or lagged dependent variables (see Section 2.3). Many researchers fit more complicated models that include covariates, more complicated time functions, etc. and obtain estimators from those models. We have not proved that these are also special cases of Assumption 2.

2.2. *Existing designs that are not special cases.* The designs considered above all use a pre-vs-post contrast (to account for time-invariant group differences) and a treated-vs-comparison contrast (to account for common shocks). Likewise, our proposed framework uses a prediction step (leveraging predictable features of each group's outcome trajectories) and a correction step (leveraging the comparison group to correct for unexpected shocks). By contrast, some designs lack an analog of either the prediction or correction steps. The interrupted time series uses only a pre-post contrast; there are no comparison units with which to perform our correction step. Synthetic control uses only a treated-comparison contrast, omitting the pre-post contrast. In the online supplement, we provide more detail on the question of synthetic control (Section 2.4), showing that it is not a special case of our framework.

2.3. *Staggered adoption.* We have assumed that all treated units receive intervention at the same time, but now extend to staggered adoption settings, taking the perspective of Callaway and Sant'Anna (2021). That is, we consider each treatment adoption time as its own

design, estimate the treatment effect in each, and weight these estimates together in a sensible way.

Define the multiple treated groups by their time of treatment adoption, g, and for the never-treated group, let $g = \infty$. Define $Y_t(0)$ as the potential outcome at time t under assignment to being never-treated and $Y_t(g)$ as the potential outcome at time t under assignment to treatment starting at time g. Then we can re-state consistency (Assumption 1) as

$$Y_t = Y_t(0) + \sum_{g=2}^{T} \left[Y_t(g) - Y_t(0)\right] G_g \,,$$

where $G_g = \mathbb{1}\{G = g\}$ is an indicator for membership in treatment timing group g. Our target estimand is the average treatment effect on the treated for each treatment time g,

$$\text{ATT}(g) \coloneqq \text{E}_{\mathcal{P}}\left[Y_g(g) - Y_g(0) \mid G_g = 1\right].$$

That is, the ATT is the difference in potential outcomes under the condition of being treated at time g versus being never-treated for units in treatment timing group g. To identify this, we re-state the equal expected prediction errors assumption as,

$$\text{E}_{\mathcal{P}}\left[Y_t(0) - f\left(\boldsymbol{X}_t\right) \mid G_g = 1\right] = \text{E}_{\mathcal{P}}\left[Y_t(0) - f\left(\boldsymbol{X}_t\right) \mid G_\infty = 1\right], \text{ for } t = g.$$

Then we can use any of the several ideas in Section 3 of Callaway and Sant'Anna (2021) to weight the resulting estimates together.

This simplifies the approach of Callaway and Sant'Anna (2021) in two ways. First, we exclude the possibility of using not-yet-treated units in the comparison group. Second, we assume there is a single post-treatment time $t = g$ at which we estimate the treatment effect for each treatment timing group. Of course, both of these could be relaxed. The point is that identifying (and estimating) each ATT(g) reduces to the simple case of treated versus comparison.

3. Selecting models for robustness. We have described an assumption that can identify the ATT in controlled pre-post settings and showed that under several familiar nonparametric or structural identifying assumptions, our identifying assumption would also hold. Because our assumption frames the problem in terms of a prediction model, we want a principled basis on which to choose among potential prediction models. Next, we propose to assess models' robustness (the complement of sensitivity) and discuss the difference between a robust model and a "correct" model.

3.1. *Design sensitivity.* The design sensitivity framework, originally developed for matched observational studies (Rosenbaum, 2004), established that violations of key assumptions lead to a *range* of point estimates that are consistent with sample data; therefore, an estimator limits not to a point, but rather an *interval* (Rosenbaum, 2005, 2012). For a given violation, a sensitive design has a wider limiting interval than a more robust one. Conversely, in our framework, a more robust prediction model leads to a narrower limiting interval.

The robustness of a model to violations of an identifying assumption is different from the plausibility that a model is "correct," i.e., satisfies an identifying assumption. In the name of assessing plausibility that a model is "correct", researchers often study whether a version of an identification assumption holds in the pre-period. For example, in DID designs, it is common to test for non-parallel trends in the pre-period, which resembles a Granger causality test (Granger, 1969) and other forms of "placebo" tests (see, e.g., Angrist and Pischke, 2008, p. 237). This practice implicitly assumes that patterns observed in the pre-period would have continued into the post-period in the absence of treatment. In other words, this approach

replaces one unverifiable assumption about counterfactual outcomes with another (Egami and Yamauchi, 2022). The framework of design sensitivity offers a practical way out of this bind: we study the robustness of our inference to violations of the identifying assumption, grounded in empirical evidence about the potential magnitude of those violations.

3.2. *Robustness criterion.* We build on Rambachan and Roth (2023) who, following Manski and Pepper (2018), set-identify the ATT by bounding the possible violations of parallel trends. They posit that the violation lies in a set defined by the observed pre-period differential trends, yielding sensitivity bounds on the ATT. Similarly, we suppose violations of our identifying assumption lie in a set defined by the pre-period differential prediction errors. Denote the observable population-level differential prediction errors in period t under model specification $f \in \mathcal{F}$ by

$$\delta_{f,t} \coloneqq \mathrm{E}_{\mathcal{P}}\left[Y_t - f\left(\boldsymbol{X}_t\right) \mid G = 1\right] - \mathrm{E}_{\mathcal{P}}\left[Y_t - f\left(\boldsymbol{X}_t\right) \mid G = 0\right]. \tag{8}$$

The point identification of Assumption 2 under model f can now be expressed as $\delta_{f,T} - \text{ATT} = 0$. For set identification, we would instead suppose that $\delta_{f,T} - \text{ATT}$, i.e., the population-level difference in counterfactual prediction errors, lies in a compact set for some $f \in \mathcal{F}$.

To define a relevant set, we follow Rambachan and Roth (2023) in supposing that the violation of equal expected prediction errors is up to M times the largest absolute differential prediction error in a set of pre-treatment *validation periods*, $\mathcal{V} \subseteq \{2, \ldots, T-1\}$. That is, for any model f, we suppose that the ATT lies in the interval given by

$$\left[\delta_{f,T} - M \max_{v \in \mathcal{V}} |\delta_{f,v}|,\ \delta_{f,T} + M \max_{v \in \mathcal{V}} |\delta_{f,v}|\right] \tag{9}$$

with $M \geq 0$. This leads to our definition of sensitivity, which is simply the length of the interval in Eq. 9. A smaller length of this interval implies less sensitivity (i.e., greater robustness).

We can imagine alternatives to this set restriction that entail different relationships between pre- and post-periods. For instance, we could create an asymmetric set restriction. Or, if we think more recent validation periods are more informative, we might replace $\max_{v \in \mathcal{V}} |\delta_{f,v}|$ in Eq. 9 with $|\delta_{f,V}|$, where $V = \max \mathcal{V}$, i.e., bound the violation by M times the *most recent* absolute difference in prediction errors. Alternatively, if we think the average pre-treatment deviation matters, we could use $1/|\mathcal{V}| \sum_{v \in \mathcal{V}} |\delta_{f,v}|$. We proceed with the set restriction in Eq. 9, but these alternatives are straightforward to implement.

The sensitivity parameter M controls how tightly we constrain the assumptions. Point identification of Assumption 2 holds under $M = 0$ and set identification holds under $M > 0$. Proposition 1 establishes that we can use pre-period data to see which model, f, in a set of candidate models, $\mathcal{F}$, is most robust.

PROPOSITION 1. *Let f and f' be two prediction model specifications in the set of candidate model specifications, $\mathcal{F}$. Under the sensitivity model in Eq. (9), model f is more robust than f' if and only if* $\max_{v \in \mathcal{V}} |\delta_{f,v}| \leq \max_{v \in \mathcal{V}} |\delta_{f',v}|$.

The proof is in the online supplement (see Section 1.2).

Proposition 1 shows that, as long as there is some nonzero pre-treatment difference in prediction errors for all $f \in \mathcal{F}$, the most robust model for any $M > 0$ will be the one with the smallest maximum absolute difference in prediction errors. By defining robustness in terms of observable pre-period quantities, we can choose among candidate models using the data. If we had a different set restriction (e.g., the most recent or mean across validation periods), the procedure for selecting models on robustness is the same: the one with the narrowest sensitivity bounds.

How is choosing the most robust model different from choosing the "correct" model? Suppose that Assumption 2 holds exactly for one model that nonetheless is less robust (by our criterion) than another candidate model. Proposition 2 quantifies the consequences of this trade-off between "correctness" and robustness.

PROPOSITION 2. *Suppose Assumption 2 holds for f' but not f and that f' is less robust than f, as defined in Proposition 1. The difference between the ATT of the "correct" model and the robust model is*

$$E_{\mathcal{P}}\left[f\left(\boldsymbol{X}_T\right) - f'\left(\boldsymbol{X}_T\right) \mid G=0\right] - E_{\mathcal{P}}\left[f\left(\boldsymbol{X}_T\right) - f'\left(\boldsymbol{X}_T\right) \mid G=1\right] . \tag{10}$$

The proof is in the online supplement (see Section 1.3).

Proposition 2 shows that when a model's differential prediction errors in the validation periods provide "misleading" information about its (unobservable) differential prediction error in the post-period, our conclusions will suffer. This is related to the idea that conclusions are more robust if point estimates are stable across competing models (Brown and Atal, 2019; O'Neill et al., 2016). In our framework, two prediction models that yield identical point estimates for $M = 0$ can have quite different robustness for $M > 0$. However, if two models yield identical point estimates, Eq. (10) is equal to 0. Therefore, stable point estimates across prediction models do not imply our conclusions are more robust, but they do mitigate a potential trade-off in which choosing the most robust model could come at the expense of choosing the "correct" model.

4. Model selection, estimation, and inference. Thus far, we have considered population quantities only. To extend our ideas to estimation and inference in finite samples, we cannot simply plug in sample analogs of population quantities. This is because we use the data twice: first to choose a robust prediction model and again to estimate our target parameter. We therefore develop a procedure that accounts for this, illustrating our ideas in an important and accessible class of prediction models: ordinary least squares (OLS) linear regression. This class of models is sufficiently rich to capture a range of models that researchers employ in the gun policy literature. It would be straightforward to show that our conclusions apply to other models, such as logistic, Poisson, transformed-outcome and isotonic regression (see, e.g., Guo and Basse, 2023), but leave this as a topic for future research.

First, we set up the data structure. Suppose we have a sample of units indexed by $i = 1, \ldots, n$ (rather than u, as in the TWFE structural model, to emphasize that we are now talking about a finite sample). Each unit's observed data up to period t are

$$\boldsymbol{D}_{i,t} = \{Y_{i,t}, \boldsymbol{Y}_{i,<t}, \boldsymbol{X}_{i,t}, \boldsymbol{X}_{i,<t}, G_i\} . \tag{11}$$

where $\boldsymbol{Y}_{i,<t}$ and $\boldsymbol{X}_{i,<t}$ are the outcomes and predictors from $t = 1, \ldots, t-1$ and $Y_{i,t}$ and $\boldsymbol{X}_{i,t}$ are outcomes and predictors in period t. We collect these over units into $\boldsymbol{D}_t = \{\boldsymbol{D}_{1,t}, \ldots, \boldsymbol{D}_{n,t}\}$ and over time into $\boldsymbol{D} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_T\}$. The collections of outcomes and predictors across units and time are defined analogously; for instance, $\boldsymbol{X}_{<t}|G = g$ denotes $\{\boldsymbol{X}_{i,<t}\}_{i:G_i=g}$.

Next, we set up the prediction model in the OLS framework. We write the model f for group g in period t as a function of both predictors and parameters, $f(\boldsymbol{X}_{i,t}; \boldsymbol{\beta}_{f,g,t})$. The parameter vector $\boldsymbol{\beta}_{f,g,t} \in \Re^K$ (where K is the dimension of $\boldsymbol{X}_{i,t}$).[2] We collect the estimated

[2]To be very precise, we should index the predictors, $\boldsymbol{X}_t$ by model f, since the predictors vary across models. However, for simplicity, we omit these and assume the subscripts on β will make it clear.

parameters over groups into $\hat{\boldsymbol{\beta}}_{f,t} = \left\{\hat{\boldsymbol{\beta}}_{f,1,t}, \hat{\boldsymbol{\beta}}_{f,0,t}\right\}$, over times into $\hat{\boldsymbol{\beta}}_f = \left\{\hat{\boldsymbol{\beta}}_{f,t}\right\}_{t=1}^{T}$, and over models into $\hat{\boldsymbol{\beta}} = \left\{\hat{\boldsymbol{\beta}}_f\right\}_{f\in\mathcal{F}}$.

Before we proceed, we need a few additional assumptions. First, we place conditions on the population moments.

ASSUMPTION 3 (Population moment conditions). For groups $G = 0$ and $G = 1$, $\mathrm{E}_{\mathcal{P}}[\boldsymbol{Y}_t \,|\, G = g] < \infty$ and $\mathrm{E}_{\mathcal{P}}[\|\boldsymbol{X}_t\|^2 \,|\, G = g] < \infty$ for all $t = 1, \ldots, T$, and $\mathrm{E}_{\mathcal{P}}[\boldsymbol{X}_{<t}\boldsymbol{X}_{<t}^{\top} \,|\, G = g]$ is positive definite for all $t = 2, \ldots, T$.

The first two conditions are standard, and the third condition implies that we can generate predictions in period t based on the OLS solution to a linear regression model's objective function in periods before t.

Next, we place conditions on the sampling mechanism.

ASSUMPTION 4. For all $i = 1, \ldots, n$ and $t = 1, \ldots, T$, the sample data, $\{\boldsymbol{D}_{i,t}\}$ are independent and identically distributed (i.i.d.).

With these in hand, we write our point estimator of $\delta_{f,t}$ as

$$\begin{aligned}\hat{\delta}\left(\boldsymbol{D}_t, \hat{\boldsymbol{\beta}}_{f,t}\right) = &\left(\frac{1}{n_1}\right)\sum_{i=1}^{n}\mathbb{1}\{G_i = 1\}\, Y_{i,t} - \left(\frac{1}{n_1}\right)\sum_{i=1}^{n}\mathbb{1}\{G_i = 1\}\, \boldsymbol{X}_{i,t}\hat{\boldsymbol{\beta}}_{f,1,t} \\ &- \left[\left(\frac{1}{n_0}\right)\sum_{i=1}^{n}\mathbb{1}\{G_i = 0\}\, Y_{i,t} - \left(\frac{1}{n_0}\right)\sum_{i=1}^{n}\mathbb{1}\{G_i = 0\}\, \boldsymbol{X}_{i,t}\hat{\boldsymbol{\beta}}_{f,0,t}\right],\end{aligned} \tag{12}$$

where $n_g \coloneqq \sum_{i=1}^{n}\mathbb{1}\{G_i = g\}$. The estimator of lower and upper bounds of the ATT in period T for any $M \geq 0$ and $f \in \mathcal{F}$, is

$$\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}_f, M\right) = \hat{\delta}\left(\boldsymbol{D}_T, \hat{\boldsymbol{\beta}}_{f,T}\right) \pm M \max_{v\in\mathcal{V}}\left|\hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}_{f,v}\right)\right|. \tag{13}$$

When $M = 0$, we simply use $\hat{\delta}\left(\boldsymbol{D}_T, \hat{\boldsymbol{\beta}}_{f,T}\right)$.

A simple approach to estimation would be to 1) estimate $\hat{\delta}\left(\boldsymbol{D}_t, \hat{\boldsymbol{\beta}}_{f,v}\right)$ for each model and validation period, 2) choose the model with the smallest worst-case absolute difference in prediction errors over the validation periods, and 3) use that model to estimate the ATT and its bounds. However, because the chosen model depends on our particular sample, we want to incorporate this uncertainty about the model into our procedure.

The usual approach of splitting data into testing and training subsets is not feasible. We cannot split the data "vertically" (i.e., in time) because our estimators and model selection criterion use the same data *by construction*: terms in Eq. (13) use data from pre-treatment validation periods $\mathcal{V}$. Nor can we rely on splitting the data "horizontally": many applications (including the one we consider here) have only a single or a few treated units, so we cannot afford to split the units.

Therefore, we propose to use a Bayesian model averaged (BMA) estimator, which averages the estimates across models, weighting each by the model's posterior probability that it is the most robust in the population. We write this estimator as

$$\widehat{\mathrm{E}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M\right)\right] = \sum_{f\in\mathcal{F}}\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}_f, M\right)\hat{p}_f, \tag{14}$$

where $\hat{p}_f$ is the posterior probability that model f is the most robust model, given the sample data. This alternative to the "pick the winner" approach outlined above has statistical advantages (Piironen and Vehtari, 2017; Madigan and Raftery, 1994; Draper, 1995; Moulton, 1991; Raftery, Madigan and Hoeting, 1997).

How do we estimate these posterior probabilities? Rather than a fully Bayesian approach, we extend the quasi-Bayesian procedure of Gelman and Hill (2006). This has been employed by many researchers (e.g., King, Tomz and Wittenberg, 2000; Tomz, Wittenberg and King, 2003), including in interrupted time series designs (Miratrix, 2022). The idea is to generate samples from the "quasi-posterior" of all the parameters across all prediction models. For this distribution, we use a multivariate Normal with mean equal to the estimated parameters $\hat{\boldsymbol{\beta}}$ (collected over times and models) and their estimated (robust) variance-covariance matrix, $\widehat{\boldsymbol{\Sigma}}$. This is equivalent to the posterior distribution of the models' parameters if the prior were flat. To estimate the variance-covariance of all the parameters across all the model and time periods simultaneously, we use seemingly unrelated regression tools pioneered by Zellner (1962, 1963), detailed in the Appendix (Section B).

Our procedure for estimating the posterior probability that a model is optimal in the population is this: for each draw from the quasi-posterior, predict outcomes and calculate differential prediction errors over the validation periods, then select the best model. Doing this many times generates a distribution for the best model. That is, the number of times each model is selected by this procedure is proportional to the strength of the evidence that each model is the most robust.

To formally characterize this procedure, let $\hat{\boldsymbol{\beta}}^{(b)}$ for $b = 1, \ldots, B$ be draws from $\mathcal{N}(\hat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$. Then, for all $f \in \mathcal{F}$, write $\hat{p}_f$ as

$$\hat{p}_f \coloneqq \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{ f = \arg\min_{f \in \mathcal{F}} \max_{v \in \mathcal{V}} \left| \delta\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}_{f,v}^{(b)} \right) \right| \right\}, \tag{15}$$

which is the proportion of draws in which f is the most robust model. Below we show that, in a sufficiently large sample, this proportion will be close to one with high probability for the truly most robust model.

LEMMA 1. *Let $f^\dagger \in \mathcal{F}$ denote the most robust model in the population. Under Assumptions 1, 3, and 4,*

$$\hat{p}_{f^\dagger} \xrightarrow{p} 1.$$

The proof is given in the online supplement (see Section 1.4). As we show next, this lemma implies that our BMA estimator is consistent.

PROPOSITION 3. *Under Assumptions 1, 3, and 4,*

$$\widehat{E}_{\mathcal{F} \mid \boldsymbol{D}}\left[\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M \right) \right] \xrightarrow{p} \delta_{f^\dagger, T} \pm M \max_{v \in \mathcal{V}} |\delta_{f^\dagger, v}|.$$

The proof is given in the online supplement (see Section 1.5). Proposition 3 shows that the BMA estimator converges in probability to the same limit as that of an estimator in which the optimal model in the population were known before observing data. We provide a conceptual diagram of the overall estimation process in the Appendix (Section C).

For inference, we build on the approach from Antonelli, Papadogeorgou and Dominici (2022). Those authors establish that we can estimate the uncertainty about both the model and the data in a computationally tractable way by summing two components: variance of the model posterior (holding the sample fixed) and variance of the sample (holding the model

posterior fixed). Denote the variance of B draws from the observed posterior, holding the sample fixed, by

$$\widehat{\text{Var}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}\left(\boldsymbol{D},\hat{\boldsymbol{\beta}},M\right)\right] \coloneqq \sum_{f\in\mathcal{F}}\left(\hat{\Delta}\left(\boldsymbol{D},\hat{\boldsymbol{\beta}}_f,M\right)-\widehat{\text{E}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}\left(\boldsymbol{D},\hat{\boldsymbol{\beta}},M\right)\right]\right)^2\hat{p}_f, \tag{16}$$

where, recalling Eq. 15, $\hat{p}_f$ is

$$\frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\left\{f=\arg\min_{f\in\mathcal{F}}\max_{v\in\mathcal{V}}\left|\delta\left(\boldsymbol{D}_v,\hat{\boldsymbol{\beta}}_{f,v}^{(b)}\right)\right|\right\}.$$

Then denote the variance of our estimator over R resamples of data, holding fixed the observed posterior, as

$$\begin{aligned}\widehat{\text{Var}}_{\boldsymbol{D}^{(r)}\,|\,\mathcal{F}}\left[\widehat{\text{E}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}(\boldsymbol{D}^{(r)},\widehat{\boldsymbol{\beta}}^{(r)},M)\right]\right] \coloneqq \frac{1}{R}\sum_{r=1}^{R}\Bigg(&\widehat{\text{E}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}(\boldsymbol{D}^{(r)},\widehat{\boldsymbol{\beta}}^{(r)},M)\right]\\ &-\frac{1}{R}\sum_{r=1}^{R}\widehat{\text{E}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}(\boldsymbol{D}^{(r)},\widehat{\boldsymbol{\beta}}^{(r)},M)\right]\Bigg)^2.\end{aligned} \tag{17}$$

The sum of Eq. (16) and (17) is the variance estimator of the BMA estimator in Eq. (14), accounting for both sampling and model uncertainty. Confidence intervals can then be constructed by drawing on a Normal approximation. We show via simulations in the online supplement (see Section 4) that this approach to inference yields 95% confidence intervals with coverage at least as great as nominal rates in moderately large samples. We also observe the conservatism that Antonelli, Papadogeorgou and Dominici (2022, p. 103) note.

5. The effect of gun laws on violent crime. We now return to our analysis of Missouri's repeal of its permit-to-purchase (PTP) law. Our data comprise state-year observations of the homicide rate in Missouri and each of its eight neighboring comparison states. To estimate the repeal's impact, we form a set of candidate prediction models drawn from the gun policy literature. Researchers agree on a basic model with unit fixed effects (as in Webster, Crifasi and Vernick (2014)), but disagree on other model components. Based on our survey of the literature, we divide the relevant model components into three categories:

1. Unit-specific time trends. Researchers often include unit-specific time trends, usually linear but sometimes more complicated forms (Black and Nagin, 1998; French and Heagerty, 2008). Others explicitly advocate against their inclusion (Aneja, Donohue III and Zhang, 2014; Wolfers, 2006). We consider models that include unit-specific linear or quadratic trends. (It is straightforward to include higher-order trends, e.g., cubic, quartic, quintic, etc.)

2. Lagged dependent variables (LDV). Some researchers include lags of the dependent variable, (Duwe, Kovandzic and Moody, 2002; Moody et al., 2014) while others advocate against their inclusion because of the possibility of bias in short time series (Nickell, 1981). Following the applied literature, we consider only models that include values of the dependent variable at one time lag; however, multiple time lags are straightforward to incorporate.

3. Outcome transformations. Linear regression is popular for outcome regressions, but can be problematic because many outcomes of interest (including the homicide rate that we consider) are naturally bounded (Moody, 2001; Plassmann and Tideman, 2001). We use only linear models, but do consider transformations of the outcome variable, specifically logs and first differences (Black and Nagin, 1998). However, because we want to compare across models, we back-transform our prediction to the original outcome scale to compute prediction errors.

Obviously, this framework leaves out some modeling variations. For example, some studies in the gun policy literature employ random effects (Crifasi et al., 2018) and two-stage models (Rubin and Dezhbakhsh, 2003). However, given the prominence of these three model components, as well asa unit fixed effects and linear models, we believe the resulting set of candidate models is reasonably broad and relevant to the gun policy literature.

From the model components above (summarized in Table 1), we take all possible combinations to derive a set of 18 candidate models. Because of their use in virtually all prediction models we surveyed, we include unit fixed effects for all 18 prediction models.

Time trend	Lagged dependent variables	Outcome transformations
None	None	None
t	Y_{t-1}	$\log(Y_t)$
t^2		$Y_t - Y_{t-1}$

TABLE 1

Model components used to create a set of candidate prediction models.

To select among the 18 prediction models, we estimate the differences in average prediction errors between treated and comparison groups. For each year prior to the law's passage in 2007, we train our prediction models on the previous years. For example, in 2006, we train a model on data from 1999 to 2005, predict in 2006, and compute the difference in average prediction errors between treated and comparison groups. To ensure adequate years of training data, we follow Hasegawa, Webster and Small (2019) in beginning the validation period in 1999. Thus, we have 5 or more years of training data, even in first validation year (1999, for which we train the model on data from 1994 – 1998).

Figure 2 shows the absolute differential average prediction errors for all 18 models, with the maximum for each model highlighted in black. The LDV model fit to the log of the outcome without unit-specific time trends (row 1, column 4) minimizes our sensitivity criterion on the sample data. The baseline mean model (row 2, column 4), which most closely corresponds to the model of choice in Webster, Crifasi and Vernick (2014) and Hasegawa, Webster and Small (2019), is the fifth-best model.

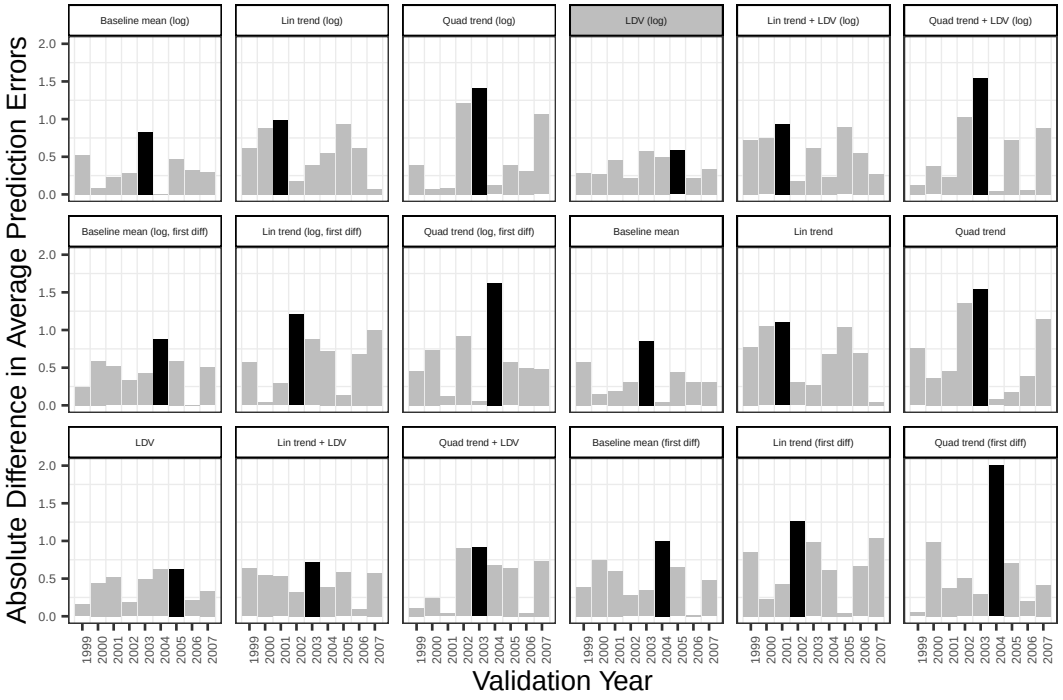

FIG 2. *Absolute difference in average prediction errors for all candidate models. The maximum for each model is highlighted in black.*

From Figure 2, we can also see which prediction models would be optimal under different sensitivity criteria. For example, the prediction model with the smallest absolute difference in average prediction errors in the last pre-period (2007) is the linear time trend model (row 2, column 5). By contrast, the prediction model with the smallest absolute difference in average prediction errors, averaged over all validation periods, is the baseline mean model on the outcome's log scale (row 1, column 1). These different loss functions for choosing the optimal model can be justified by an appropriate sensitivity analysis model. Given the sensitivity analysis in Eq. (9), which aligns with the sensitivity analysis proposed in recent research (Rambachan and Roth, 2023), the aforementioned LDV model on the log scale is optimal.

Figure 3 shows the predictions and their errors for all validation periods with this optimal prediction model.

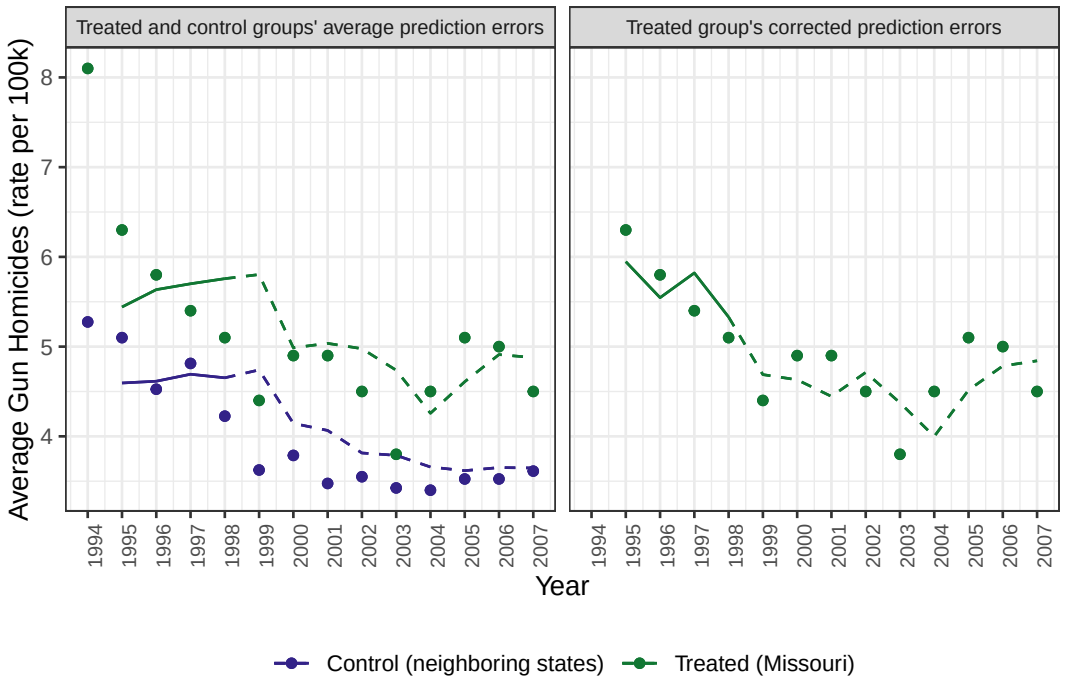

FIG 3. *Average outcomes and the optimal model's average predictions for all pre-treatment validation periods in treated and control states. The points are observed outcomes, the solid lines are model-fitted values in the training periods and the dashed lines are modeled predictions in the validation periods.*

Figure 4 suggests that potential trade-offs between models' "correctness" and robustness are not especially severe. The variability in point estimates across models is not especially large (sd = 0.38).

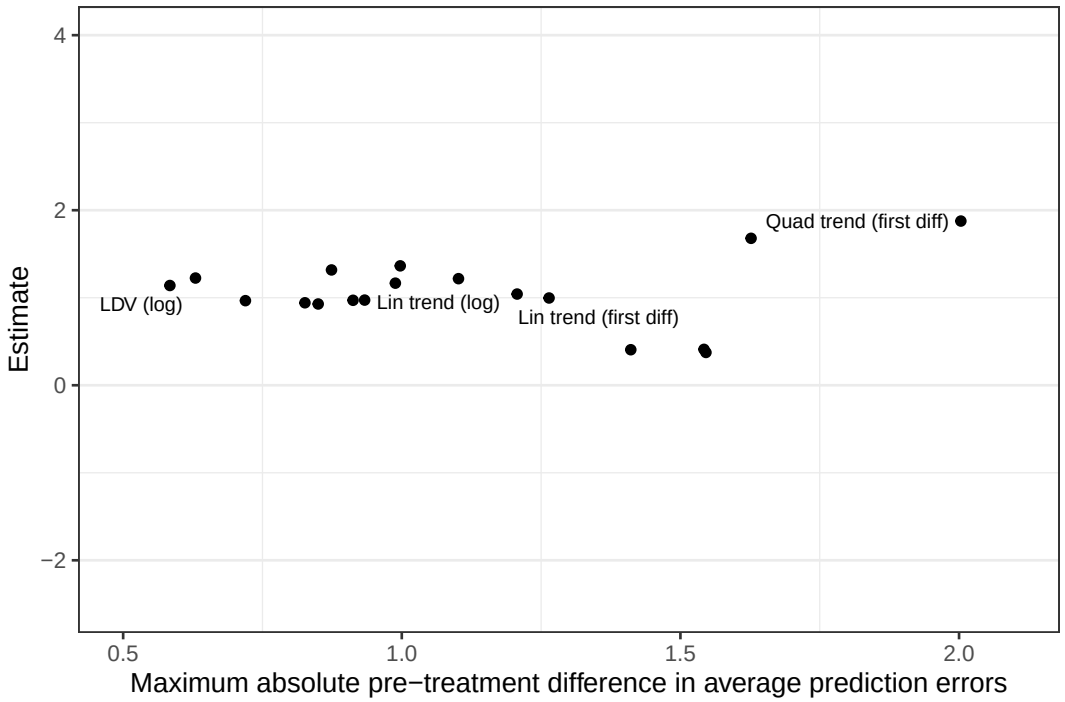

FIG 4. *Estimates under each model (y axis) and corresponding maximum absolute differential prediction errors in the pre-period (x axis). The optimal model is shaded in gray.*

Much value remains in the similarity of point estimates across models even if they differ dramatically in terms of robustness. (For instance, the most and least robust models yield relatively similar point estimates of 1.14 and 1.88, respectively.) As Proposition 2 shows, if point identification of $M = 0$ happens to be true under one model that is not the most robust, then the point estimate under the most robust model will not be too misleading insofar as the estimates under both models are similar.

Turning to estimation and inference, this empirical setting requires careful attention to the sources of randomness. In the setting of gun policy research, an influential article by Manski and Pepper (2018) argues that it is difficult to conceive the units of analysis — typically state-years — as randomly sampled from a target population of interest: "Random sampling assumptions, however, are not natural when considering states or countries as units of observations" (Manski and Pepper, 2018, p. 235). In the setting of most gun policy research, as Manski and Pepper (2018) argue, uncertainty is driven by a fundamental ambiguity over whether counterfactual point identification assumptions hold — i.e., what Rambachan and Roth (2023, p. 2556) call "identification uncertainty."

In a setting characterized by only identification rather than sampling uncertainty, Rambachan and Roth (2023, p. 2563) argue that a natural starting point for controlled pre-post designs is one of set identification with $M = 1$. In this set identification framework (as opposed to point identification in which $M = 0$), researchers can then gradually increase M in a subsequent sensitivity analysis. The crucial feature of this inferential setting is the absence of uncertainty over which model is truly optimal. In this setting one could deterministically select the truly optimal model. Then, given the selection of this optimal model, it would be straightforward to calculate bounds on the ATT under $M = 1$ and to assess the sensitivity of these bounds under increasing values of M. Under this approach, the bounds of the ATT (with $M = 1$) under the most robust model is $[0.56, 1.72]$. The changepoint value of M, i.e.,

the smallest value of M at which the estimated lower and upper bounds of the ATT bracket 0, is 1.95.

The analysis above supposes the setting that Manski and Pepper (2018) argue is most sensible for our application. However, if we suppose that states are independent and identically distributed draws from a target population, then the estimation and inferential procedure in Section 4 is appropriate. The BMA point estimate (under $M = 0$) of 1.16 is nearly identical to the point estimate under the optimal model (1.14) in the realized sample. This optimal model in the sample (LDV fit on the log scale) also receives the greatest posterior probability that it is optimal, 0.46. The model with the second greatest posterior probability of 0.4 is the next best model in the sample data (the LDV model fit without the log transformation). Figure 5 below shows the full posterior distribution given the sample data.

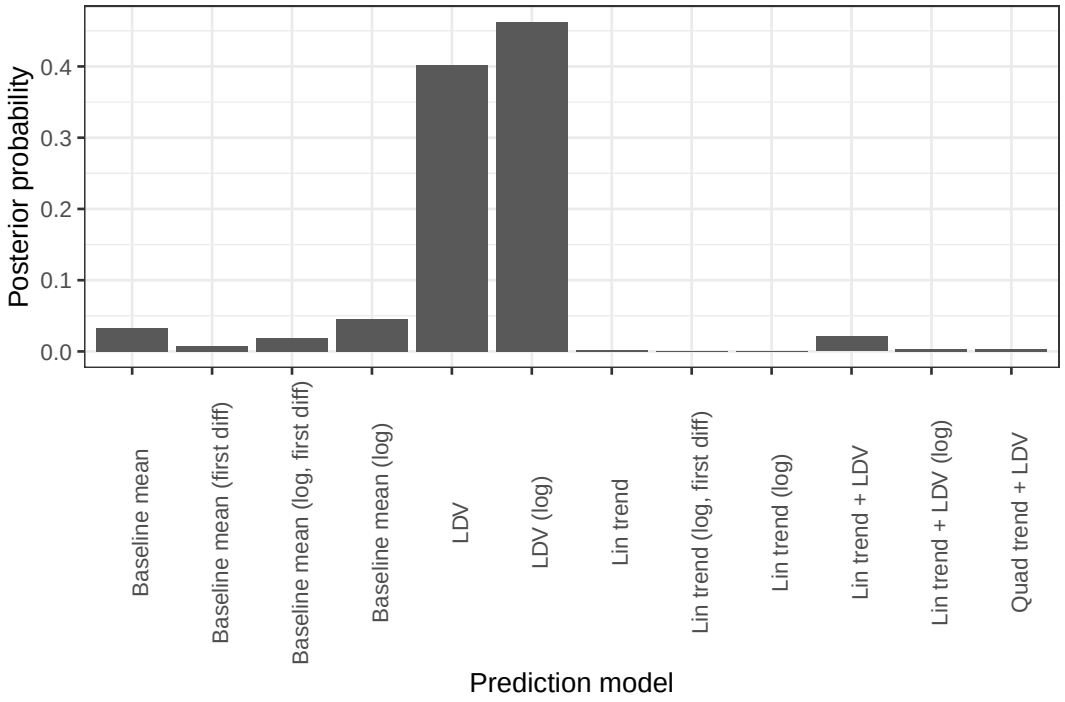

FIG 5. *Posterior plausibility that each candidate model is truly optimal model*

Using our proposed variance estimation procedure, which we would expect not to perform at its best in small samples, yields an estimated standard error (accounting for both model and sampling uncertainty) of 0.14 and corresponding 95% confidence interval of $[0.88, 1.44]$. That is, we conclude that the repeal of Missouri's permit-to-purchase law increased the state's gun homicide rate by somewhere between 0.88 to 1.44 per 100,000 population. For context, the observed homicide rate in 2007 (just before the repeal) in Missouri was 4.5, so the point estimate of 1.16 represents a 26% increase.

The changepoint value of M is 1.82. The same changepoint value of M at which the lower bound estimator's 95% confidence interval no longer excludes 0 is approximately 1.07. The latter changepoint M's small value is to be expected in a a small study like this application.

6. Conclusion and open questions. In this paper, we introduce a new method for identifying the ATT based on a "predict, correct, select" procedure. We *predict* untreated potential outcomes, *correct* them using the observable prediction error in the comparison group, and

select the optimal prediction model using a robustness criterion. Our causal identification of the ATT based on these predictions assumes equal prediction errors (in expectation) in the treated and comparison groups. We have shown that several popular designs are special cases of our general identification framework.

We developed these ideas to reconcile disparate results from gun policy evaluations. Specifically, we studied the repeal of Missouri's permit-to-purchase law in 2007 using models drawn from the literature. Rather than make claims that any one underlying causal model is "correct", we selected the optimal model based on robustness. We found that a lagged dependent variable model, fit on the log scale, minimized our robustness criterion in our sample, making this model the most likely to be the optimal model in the population (although other models are plausible as well); our overall point estimate, averaging over the posterior probability that each model is optimal in the population, was an increase of 1.16 homicides per 100,000 population.

Our sensitivity bounds would include 0 for $M \geq 1.82$. That is, the violation of Assumption 2 would have to be more than 1.82 times greater than the worst violation (for a weighted combination of models) in the 9 validation years. By contrast, in the absence of our Bayesian model selection procedure, the value of M that leads the sensitivity bounds to include 0 could be much smaller under any given model, as low as 0.24, with an unweighted average (across all models) of 1.06.

Our approach has several limitations. First, like all causal inference methods, our identifying assumption is untestable because it involves counterfactual quantities. Studying the differential prediction errors of a set of models in the pre-period has similar conceptual problems to testing for differential pre-trends in difference-in-differences. This is why we use a sensitivity perspective to choose a prediction model based on robustness.

Second, our method is scale-dependent because we measure prediction error as a linear difference on the scale of the outcome variable. This limits our approach. However, we believe this limitation is not specific to our particular framework, as scale dependence is a well-known issue in controlled pre-post designs as a whole.

Third, prediction models can only use variables that are measured prior to treatment. For some data-generating models, such as interactive fixed effects, the correction step will not debias the estimator because the shocks do not affect treated and comparison groups equally. However, as pointed out by a reviewer, an interesting extension of our ideas might separate the comparison units into some for the prediction step and others for the correction step. For instance, the contemporaneous outcomes of some comparison units could be allowed into the prediction function for the treated units' post-period outcomes, while other comparison units' post-period outcomes are used to correct for unexpected common shocks.

Fourth, by switching to a robustness criterion for model selection, we induce a possible "correctness" versus robustness trade-off (Proposition 2). Rather than claim that we can choose the "correct" model, we choose a model that maximizes our robustness criterion. A model for which our identifying (Assumption 2) holds exactly need not maximize robustness. However, since there is no data-driven way to choose a model that satisfies a causal identification assumption, we believe choosing based on robustness offers an appealing alternative.

Finally, our inferential procedure, which attempts to account for many sources of uncertainty, may not adequately address all of them. Bootstrap methods perform poorly when there are few clusters, as in our analysis with only one treated unit and eight comparison units (Bertrand, Duflo and Mullainathan, 2004; MacKinnon and Webb, 2020; Conley and Taber, 2011; Rokicki et al., 2018). However, we still believe that our proposal for formally accounting for the model selection procedure is an improvement over the status quo, in which model selection is usually hidden from view and outside the bounds of inference entirely. Post-selection inference is an active area of research, and as a recent review article noted, "has

a long and rich history, and the literature has grown beyond what can reasonably be synthesized in our review" (Kuchibhotla, Kolassa and Kuffner, 2022, p. 506). Future research should explore the application of these simultaneous inference and conditional selective inference methods to problems like ours in which sample splitting in infeasible.

Our proposal also has several key strengths. First, our conception of robustness allows us to choose a prediction model using pre-treatment observations only. This may discourage fishing, i.e., picking a prediction model that yields the most desirable or "statistically significant" result. Contrast this with selecting a model based on "correctness," which involves assumptions about unknowable counterfactual outcomes and therefore introduces the temptation to claim that the model with the most favorable results is the "correct" model.

Second, many researchers already interpret robustness in terms of "correctness." In difference-in-differences, for instance, researchers interpret parallel trends in the pre-period as evidence for the plausibility of the true identifying assumption of parallel trends from the pre- to post-periods. Yet pre-period parallel trends provide evidence of counterfactual parallel trends only under additional assumptions, and violations of pre-period parallel trends can still be consistent with the identifying assumption (Kahn-Lang and Lang, 2020; Roth and Sant'Anna, 2023). Therefore, our proposal offers a more transparent version of this practice, recasting the evaluation of pre-period violations as a sensitivity analysis rather than as a test of untestable assumptions.

Third, we show that some familiar designs are special cases of this assumption for particular choices of prediction models. Thus, to generate the set of candidate prediction models, the existing literature can provide a rich set of models that already have the imprimatur of plausibility.

Fourth, as noted by one of the reviewers, one could use this framework for estimators that rely on an ignorability assumption by selecting models based on the prediction error for the treated units only during the pre-intervention period, rather than *differential* prediction error. This type of "placebo test" is common in the synthetic controls literature (Robbins, Saunders and Kilmer, 2017).

However, we need not be limited to models already in use. A last and potentially significant benefit of our proposed method is its ability to draw upon flexible and modern prediction models, e.g., machine learning methods. Recall that we need not believe the model; in fact, the inner workings of the prediction models can remain a black box. As long as it generates equally good predictions in the treated and comparison group, we can identify our target causal estimand. However, we note that our estimation and inferential procedure would need to be substantially updated to accommodate such models, and believe this is a fruitful line of future inquiry.

APPENDIX A: EXISTING MODELS AS SPECIAL CASES (OR NOT)

Our proofs each follow the steps sketched out below.

1. Use the design's identification assumptions to re-express the treated and comparison groups' untreated potential outcomes (in expectation) in the post-period, $\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G = 1\right]$ and $\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G = 0\right]$.
2. Write the prediction errors in treated and comparison groups (in expectation):
   a) First, use Assumption 1 to substitute untreated potential outcomes for any observed outcomes in the argument $\boldsymbol{X}_t$ to the prediction model, $f\left(\boldsymbol{x}\right)$.[3]

[3]Since the prediction model can only use pre-treatment outcomes, any outcomes in $\boldsymbol{X}_t$ are untreated potential outcomes.

b) Next, take expectation (with respect to the identification assumptions) of the prediction models in each group, $\mathrm{E}_{\mathcal{P}}\left[f\left(\boldsymbol{X}_T\right) \mid G=1\right]$ and $\mathrm{E}_{\mathcal{P}}\left[f\left(\boldsymbol{X}_T\right) \mid G=0\right]$
c) Finally, compute the differential prediction error (in expectation),

$$\mathrm{E}_{\mathcal{P}}\left[Y_T(0)-f\left(\boldsymbol{X}_T\right) \mid G=1\right]-\mathrm{E}_{\mathcal{P}}\left[Y_T(0)-f\left(\boldsymbol{X}_T\right) \mid G=0\right].$$

3. Show that this is equal to 0, thereby implying Assumption 2 and, consequently, the identified estimand in Eq. (5).

A.1. Difference-in-Differences. If the prediction function is

$$f(\boldsymbol{X}_t)=Y_{t-1}, \tag{18}$$

then Assumption 2 will be true whenever parallel trends holds.

First, use parallel trends in Eq. (6) to write the treated and comparison groups untreated potential outcomes (in expectation) in the post-treatment period as

$$\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right]=\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right]+\left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right]-\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right]\right)$$
$$\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right]=\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right]+\left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right]-\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right]\right).$$

Next, using Assumption 1, the expectations of the prediction model in Eq. (18) in each group are

$$\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=1\right]=\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right]$$
$$\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=0\right]=\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right].$$

Hence, the differential prediction error (in expectation) is

$$\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right]-\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right]-\left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right]-\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right]\right),$$

which is equal to 0 by parallel trends in Eq. (6). Hence, Assumption 2 also holds.

A.2. Two-way Fixed Effects. If the prediction function is

$$f(\boldsymbol{X}_t)=\arg\min_{\alpha_u} \sum_{s=1}^{t-1}\left(Y_{u,s}-\alpha_u\right)^2, \tag{19}$$

then Assumption 2 will be true whenever the TWFE structural model in Eq. 7 holds.

First, the structural model in Eq. (7) yields the following untreated potential outcomes (in expectation) in the post-period:

$$\mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) \mid G_u=1\right]=\mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u=1\right]+\gamma_T$$
$$\mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) \mid G_u=0\right]=\mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u=0\right]+\gamma_T.$$

The prediction model in Eq. (19) is simply each unit's average outcome in the pre-period,

$$f\left(\boldsymbol{X}_T\right)=\arg\min_{\alpha_u} \sum_{t=1}^{T-1}\left(Y_{u,t}-\alpha_u\right)^2=\frac{1}{(T-1)} \sum_{t=1}^{T-1} Y_{u,t}, \tag{20}$$

so substituting $Y_{u,t}(0)$ for the observed outcomes (by Assumption 1) and taking expectation with respect to the structural model in Eq. (7) yields

$$\mathrm{E}_{\mathcal{P}}\left[f\left(\boldsymbol{X}_T\right) \mid G_u=1\right]=\mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u=1\right]+\left(\frac{1}{T-1}\right) \sum_{t=1}^{T-1} \gamma_t$$

$$\mathrm{E}_{\mathcal{P}}\left[f\left(\boldsymbol{X}_T\right) \mid G_u=0\right]=\mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u=0\right]+\left(\frac{1}{T-1}\right) \sum_{t=1}^{T-1} \gamma_t.$$

By substitution, we write the differential prediction error (in expectation) as

$$\begin{aligned}\delta_T &= \mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) - f\left(\boldsymbol{X}_T\right) \mid G_u = 1\right] \\ &\quad - \mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) - f\left(\boldsymbol{X}_T\right) \mid G_u = 0\right] \\ &= \left(\mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 1\right] + \gamma_T - \mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 1\right] - \left(\frac{1}{T-1}\right)\sum_{t=1}^{T-1}\gamma_t\right) \\ &\quad - \left(\mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 0\right] + \gamma_T - \mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 0\right] - \left(\frac{1}{T-1}\right)\sum_{t=1}^{T-1}\gamma_t\right),\end{aligned}$$

which is equal to 0, thereby implying Assumption 2.

Thus, the popular TWFE structural model implies our identification condition when the prediction function is OLS with unit fixed effects. This result would still hold if one were to fit both unit and time fixed effects, but doing so is unnecessary because the latter are constant across units and, hence, eliminated by the treated-minus-control difference between groups.

On the other hand, other structural models require more careful thought about the appropriate prediction function. For example, with a unit- or group-specific linear time trend model, use of the prediction function in Eq. 19 would not imply equal expected prediction errors, but use of the OLS analog of this model does. Other models, such as that of interactive fixed effects, typically used to justify the synthetic control method (Abadie, Diamond and Hainmueller, 2010), have no clear corresponding prediction function that implies equal expected prediction errors. This should be unsurprising since the synthetic control design, which is based on a treated-versus-control contrast, is outside the scope of controlled pre-post designs.

Embedding potential outcomes in structural models or specific parametric distributions can provide intuition about when equal expected prediction errors holds. However, our identification condition does not require such assumptions. The prediction functions, which may or may not use OLS, should be interpreted as just that — algorithms without the assumptions of corresponding structural models. This approach to prediction models is common in design-based settings wherein randomness stems from either an assignment (Rosenbaum, 2002; Sales, Hansen and Rowan, 2018) or sampling (Huang et al., 2023) mechanism.

APPENDIX B: JOINT VARIANCE-COVARIANCE MATRIX FOR BAYESIAN MODEL SELECTION

The estimated joint variance-covariance matrix for all coefficients across all models and validation periods is

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_{(f_1,v_1),(f_1,v_1)} & \cdots & \widehat{\boldsymbol{\Sigma}}_{(f_1,v_1),(f_{|\mathcal{F}|},v_{|\mathcal{V}|})} \\ \vdots & \ddots & \vdots \\ \widehat{\boldsymbol{\Sigma}}_{(f_{|\mathcal{F}|},v_{|\mathcal{V}|}),(f_1,v_1)} & \cdots & \widehat{\boldsymbol{\Sigma}}_{(f_{|\mathcal{F}|},v_{|\mathcal{V}|}),(f_{|\mathcal{F}|},v_{|\mathcal{V}|})}, \end{bmatrix} \tag{21}$$

where

$$\widehat{\boldsymbol{\Sigma}}_{(f,v),(f',v')} = \left(\boldsymbol{X}_{f,v}^\top \boldsymbol{X}_{f,v}\right)^{-1} \boldsymbol{X}_{f,v}^\top \boldsymbol{W}_{f,v} \boldsymbol{W}_{f',v'} \boldsymbol{X}_{f',v'} \left(\boldsymbol{X}_{f',v'}^\top \boldsymbol{X}_{f',v'}\right)^{-1} \tag{22}$$

and $\boldsymbol{W}_{f,v}$ is the $n \times n$ diagonal matrix whose ith diagonal element is the prediction error (residual) for unit i in validation period v under model f.

We can equivalently express (22) as

$$\widehat{\boldsymbol{\Sigma}}_{(f,v),(f',v')} = (1/n)\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_{f,v,i}^\top \boldsymbol{X}_{f,v,i}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} e_{f,v,i} e_{f',v',i} \boldsymbol{X}_{f,v,i}^\top \boldsymbol{X}_{f',v',i}\right) \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_{f',v',i}^\top \boldsymbol{X}_{f',v',i}\right)^{-1},$$

from which it is straightforward to see that, under suitable regularity conditions, $\widehat{\boldsymbol{\Sigma}}_{(f,v),(f',v')}$ tends to 0 as n increases indefinitely.

APPENDIX C: CONCEPTUAL DIAGRAM OF ESTIMATION PROCESS

Figure 6 provides a conceptual diagram of the overall estimation process. All of the mathematical quantities in Figure 6 are defined in the main text. However, to reiterate, the index $b = 1, \ldots, B$ runs over the quasi-posterior draws from $\mathcal{N}\left(\hat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}\right)$. In addition, $|\hat{\delta}_f^{*(b)}|$ denotes the largest absolute differential prediction error for model f over all validation periods, $\mathcal{V}$, where $V = \max \mathcal{V}$, under the bth draw from $\mathcal{N}\left(\hat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}\right)$. The optimal model under the bth draw is denoted by $f^{\dagger(b)}$. The elements in the set of candidate models, $\mathcal{F}$, are denoted by $f_1, f_2, \ldots, f_{|\mathcal{F}|}$. All other quantities — namely, $\hat{\Delta}(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M)$, $\widehat{\mathrm{E}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M\right)\right]$ and $\hat{p}_f$ — are as defined in Eqs. 13, 14 and 15 in the main text.

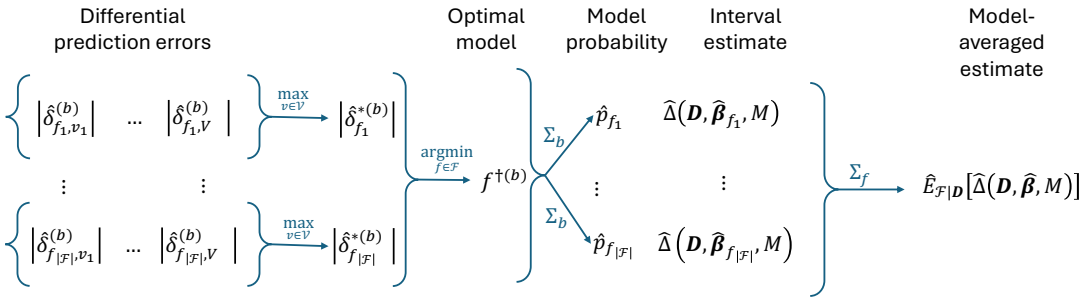

FIG 6.

Funding. This work was supported by the Agency for Healthcare Research and Quality (R01HS028985). Research reported in this publication was also supported by National Institute on Aging of the National Institutes of Health under award number P01AG032952. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association* **105** 493–505.

ANEJA, A., DONOHUE III, J. J. and ZHANG, A. (2014). The Impact of Right to Carry Laws and the NRC Report: The Latest Lessons for the Empirical Evaluation of Law and Policy Technical Report No. NBER Working Paper No. 18294, https://www.nber.org/papers/w18294, National Bureau of Economic Research, Cambridge, MA.

ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ.

ANGRIST, J. D. and PISCHKE, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *The Journal of Economic Perspectives* **24** 3–30.

ANTONELLI, J. and BECK, B. (2023). Heterogeneous Causal Effects of Neighbourhood Policing in New York City with Staggered Adoption of the Policy. *Journal of the Royal Statistical Society Series A: Statistics in Society* **186** 772–787. https://doi.org/10.1093/jrsssa/qnad058

ANTONELLI, J., PAPADOGEORGOU, G. and DOMINICI, F. (2022). Causal inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties. *Biometrics* **78** 100-114.

BASU, P. and SMALL, D. S. (2020). Constructing a More Closely Matched Control Group in a Difference-in-Differences Analysis: Its Effect on History Interacting with Group Bias. *Observational Studies* **6** 103-130.

BEN-MICHAEL, E., ARBOUR, D., FELLER, A., FRANKS, A. and RAPHAEL, S. (2023). Estimating the Effects of a California Gun Control Program with Multitask Gaussian Processes. *The Annals of Applied Statistics* **17** 985–1016. https://doi.org/10.1214/22-AOAS1654

BERTRAND, M., DUFLO, E. and MULLAINATHAN, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics* **119** 249–275.

BILINSKI, A. and HATFIELD, L. A. (2020). Nothing to See Here? Non-Inferiority Approaches to Parallel Trends and Other Model Assumptions. arXiv Preprint, https://arxiv.org/pdf/1805.03273v5.pdf.

BLACK, D. A. and NAGIN, D. S. (1998). Do Right-to-Carry Laws Deter Violent Crime? *The Journal of Legal Studies* **27** 209–219.

BLOOM, H. S. (2003). Using "Short" Interrupted Time-Series Analysis to Measure the Impacts of Whole-School Reforms: With Applications to a Study of Accelerated Schools. *Evaluation Review* **27** 3–49.

BORUSYAK, K., JARAVEL, X. and SPIESS, J. (2022). Revisiting Event Study Designs: Robust and Efficient Estimation. Working Paper, https://www.econstor.eu/bitstream/10419/260392/1/1800643624.pdf.

BRITT, C. L., KLECK, G. and BORDUA, D. J. (1996). A Reassessment of the D.C. Gun Law: Some Cautionary Notes on the Use of Interrupted Time Series Designs for Policy Impact Assessment. *Law & Society Review* **30** 361-380.

BROWN, T. T. and ATAL, J. P. (2019). How Robust are Reference Pricing Studies on Outpatient Medical Procedures? Three Different Preprocessing Techniques Applied to Difference-in Differences. *Health Economics* **28** 280–298.

CALLAWAY, B. and SANT'ANNA, P. H. C. (2021). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics* **225** 200–230.

CHAN, M. K. and KWOK, S. S. (2022). The PCDID Approach: Difference-in-Differences When Trends Are Potentially Unparallel and Stochastic. *Journal of Business & Economic Statistics* **40** 1216-1233.

CONLEY, T. G. and TABER, C. R. (2011). Inference with 'Difference in Differences' with a Small Number of Policy Changes. *The Review of Economics and Statistics* **93** 113–125.

CRIFASI, C. K., MERRILL-FRANCIS, M., MCCOURT, A. D., VERNICK, J. S., WINTEMUTE, G. J. and WEBSTER, D. W. (2018). Association between Firearm Laws and Homicide in Urban Counties. *Journal of Urban Health* **95** 383-390.

DAW, J. R. and HATFIELD, L. A. (2018). Matching in Difference-in-Differences: Between a Rock and a Hard Place. *Health Services Research* **53** 4111–4117.

DE CHAISEMARTIN, C. and D'HAULTFŒUILLE, X. (2023). Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey. *The Econometrics Journal*.

DENTEH, A. and KÉDAGNI, D. (2022). Misclassification in Difference-in-differences Models. arXiv Preprint, https://arxiv.org/pdf/2207.11890.pdf.

DRAPER, D. (1995). Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)* **57** 45-70.

DUWE, G., KOVANDZIC, T. and MOODY, C. E. (2002). The Impact of Right-to-Carry Concealed Firearm Laws on Mass Public Shootings. *Homicide Studies* **6** 271-296.

EGAMI, N. and YAMAUCHI, S. (2022). Using Multiple Pre-treatment Periods to Improve Difference-in-Differences and Staggered Adoption Designs. *Political Analysis*.

FRENCH, B. and HEAGERTY, P. J. (2008). Analysis of Longitudinal Data to Evaluate a Policy Change. *Statistics in Medicine* **27** 5005-5025.

FREYALDENHOVEN, S., HANSEN, C. and SHAPIRO, J. M. (2019). Pre-event Trends in the Panel Event-Study Design. *American Economic Review* **109** 3307–3338.

FRY, C. E. and HATFIELD, L. A. (2021). Birds of a feather flock together: Comparing controlled pre-post designs. *Health Services Research* **56** 942–952.

GELMAN, A. and HILL, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY.

GOODMAN-BACON, A. (2021). Difference-in-Differences with Variation in Treatment Timing. *Journal of Econometrics* **225** 254–277.

GRANGER, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **37** 424–438.

GRAVES, J. A., FRY, C., MCWILLIAMS, J. M. and HATFIELD, L. A. (2022). Differenceindifferences for Categorical Outcomes. *Health Services Research* **57** 681-692.

GUO, K. and BASSE, G. W. (2023). The Generalized Oaxaca-Blinder Estimator. *Journal of the American Statistical Association* **118** 524-536.

HAM, D. W. and MIRATRIX, L. (2022). Benefits and costs of matching prior to a Difference in Differences analysis when parallel trends does not hold. arXiv Preprint, https://arxiv.org/pdf/2205.08644.pdf.

HASEGAWA, R. B., WEBSTER, D. W. and SMALL, D. S. (2019). Evaluating Missouri's Handgun Purchaser Law: A Bracketing Method for Addressing Concerns About History Interacting with Group. *Epidemiology* **30** 371–379.

HUANG, M., EGAMI, N., HARTMAN, E. and MIRATRIX, L. (2023). Leveraging Population Outcomes to Improve the Generalization of Experimental Results: Application to the JTPA Study. *Annals of Applied Statistics*.

IMAI, K. and KIM, I. S. (2019). When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? *American Journal of Political Science* **63** 467–490.

IMAI, K. and KIM, I. S. (2021). On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data. *Political Analysis* **29** 405–415.

KAHN-LANG, A. and LANG, K. (2020). The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications. *Journal of Business & Economic Statistics* **38** 613–620.

KING, G., TOMZ, M. and WITTENBERG, J. (2000). Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science* **44** 341–355.

KROPKO, J. and KUBINEC, R. (2020). Interpretation and Identification of within-unit and cross-sectional variation in panel data models. *PLoS ONE* **15** e0231349.

KUCHIBHOTLA, A. K., KOLASSA, J. E. and KUFFNER, T. A. (2022). Post-Selection Inference. *Annual Review of Statistics and Its Application* **9** 505–527. https://doi.org/10.1146/annurev-statistics-100421-044639

LINDNER, S. and MCCONNELL, K. J. (2019). Difference-in-Differences and Matching on Outcomes: A Tale of Two Unobservables. *Health Services and Outcomes Research Methodology* **19** 127–144.

LIU, L., WANG, Y. and XU, Y. (2024). A Practical Guide to Counterfactual Estimators for Causal Inference with TimeSeries CrossSectional Data. *American Journal of Political Science* **68** 160-176.

LOPEZ BERNAL, J., SOUMERAI, S. and GASPARRINI, A. (2018). A Methodological Framework for Model Selection in Interrupted Time Series Studies. *Journal of Clinical Epidemiology* **103** 82-91.

MACKINNON, J. G. and WEBB, M. D. (2020). Randomization Inference for Difference-in-Differences with Few Treated Clusters. *Journal of Econometrics* **218** 435–450.

MADIGAN, D. and RAFTERY, A. E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association* **89** 1535-1546.

MANSKI, C. F. and PEPPER, J. V. (2018). How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions. *The Review of Economics and Statistics* **100** 232–244.

MARCUS, M. and SANT'ANNA, P. H. C. (2021). The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics. *Journal of the Association of Environmental and Resource Economists* **8** 235-275.

MCDOWALL, D., MCCLEARY, R. and BARTOS, B. J. (2019). *Interrupted Time Series Analysis*. Oxford University Press, New York, NY.

MIRATRIX, L. W. (2022). Using Simulation to Analyze Interrupted Time Series Designs. *Evaluation Review* **46** 750-778.

MOODY, C. E. (2001). Testing for the Effects of Concealed Weapons Laws: Specification Errors and Robustness. *The Journal of Law and Economics* **44** 799-813.

MOODY, C. E., MARVELL, T. B., ZIMMERMAN, P. R. and ALEMANTE, F. (2014). The Impact of Right-to-Carry Laws on Crime: An Exercise in Replication. *Review of Economics & Finance* **4** 33-43.

MORRAL, A. R., RAMCHAND, R., SMART, R., GRESENZ, C. R., CHERNEY, S., NICOSIA, N., PRICE, C. C., HOLLIDAY, S. B., SAYERS, E. L. P. and SCHELL, E. A. TERRY L (2018). *The Science of Gun Policy: A Critical Synthesis of Research Evidence on the Effects of Gun Policies in the United States*, 1st ed. RAND Corporation, Santa Monica, CA.

MOULTON, B. R. (1991). A Bayesian Approach to Regression Selection and Estimation, with Application to a Price Index for Radio Services. *Journal of Econometrics* **49** 169-193.

NICKELL, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica* **49** 1417-1426.

NATIONAL RESEARCH COUNCIL OF THE NATIONAL ACADEMIES (2005). *Firearms and Violence: A Critical Review*. The National Academic Press, Washington, D. C.

O'NEILL, S., KREIF, N., GRIEVE, R., SUTTON, M. and SEKHON, J. S. (2016). Estimating Causal Effects: Considering Three Alternatives to Difference-in-Differences Estimation. *Health Services and Outcomes Research Methodology* **16** 1-21.

PIIRONEN, J. and VEHTARI, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing* **27** 711-735.

PLASSMANN, F. and TIDEMAN, T. N. (2001). Does the Right to Carry Concealed Handguns Deter Countable Crimes? Only a Count Analysis Can Say. *The Journal of Law and Economics* **44** 771-798.

RAFTERY, A. E., MADIGAN, D. and HOETING, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association* **92** 179-191.

RAMBACHAN, A. and ROTH, J. (2023). A More Credible Approach to Parallel Trends. *Review of Economic Studies*.

ROBBINS, M. W., SAUNDERS, J. and KILMER, B. (2017). A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention. *Journal of the American Statistical Association* **112** 109-126.

ROKICKI, S., COHEN, J., FINK, G., SALOMON, J. A. and LANDRUM, M. B. (2018). Inference with Difference-in-Differences with a Small Number of Groups: A Review, Simulation Study and Empirical Application Using SHARE Data. *Medical Care* **56** 97–105.

ROSENBAUM, P. R. (2002). Covariance Adjustment in Randomized Experiments and Observational Studies. *Statistical Science* **17** 286–327.

ROSENBAUM, P. R. (2004). Design Sensitivity in Observational Studies. *Biometrika* **91** 153–164.

ROSENBAUM, P. R. (2005). Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies. *The American Statistician* **59** 147–152.

ROSENBAUM, P. R. (2012). An Exact Adaptive Test with Superior Design Sensitivity in an Observational Study of Treatments for Ovarian Cancer. *The Annals of Applied Statistics* **6** 83-105.

ROTH, J. (2022). Pretest with Caution: Event-Study Estimates After Testing for Parallel Trends. *American Economic Review: Insights* **4** 305-322.

ROTH, J. and SANT'ANNA, P. H. C. (2023). When Is Parallel Trends Sensitive to Functional Form? *Econometrica* **91** 737-747.

ROTH, J., SANT'ANNA, P. H. C., BILINSKI, A. and POE, J. (2023). What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. *Journal of Econometrics*.

RUBIN, P. H. and DEZHBAKHSH, H. (2003). The Effect of Concealed Handgun Laws on Crime: Beyond the Dummy Variables. *International Review of Law and Economics* **23** 199-216.

RYAN, A. M., BURGESS, J. F. and DIMICK, J. B. (2015). Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences. *Health Services Research* **50** 1211–1235.

SALES, A. C., HANSEN, B. B. and ROWAN, B. (2018). Rebar: Reinforcing a Matching Estimator With Predictions From High-Dimensional Covariates. *Journal of Educational and Behavioral Statistics* **43** 3–31.

SHADISH, W. R., COOK, T. D. and CAMPBELL, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company, Boston, MA.

SMART, R., MORRAL, A. R., SMUCKER, S., CHERNEY, S., SCHELL, T. L., PETERSON, S., AHLUWALIA, S. C., CEFALU, M., XENAKIS, L., RAMCHAND, R. and GRESENZ, C. R. (2020). *The Science of Gun Policy: A Critical Synthesis of Research Evidence on the Effects of Gun Policies in the United States*, 2nd ed. RAND Corporation, Santa Monica, CA.

SOBEL, M. E. (2012). Does Marriage Boost Men's Wages?: Identification of Treatment Effects in Fixed Effects Regression Models for Panel Data. *Journal of the American Statistical Association* **107** 521–529.

SUN, L. and ABRAHAM, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* **225** 175–199.

TOMZ, M., WITTENBERG, J. and KING, G. (2003). Clarify: Software for Interpreting and Presenting Statistical Results. *Journal of Statistical Software* **8** 1–30.

WAGNER, A. K., SOUMERAI, S. B., ZHANG, F. and ROSS-DEGNAN, D. (2002). Segmented Regression Analysis of Interrupted Time Series Studies in Medication Use Research. *Journal of Clinical Pharmacy and Therapeutics* **27** 299–309. https://doi.org/10.1046/j.1365-2710.2002.00430.x

WEBSTER, D., CRIFASI, C. K. and VERNICK, J. S. (2014). Effects of the Repeal of Missouri's Handgun Purchaser Licensing Law on Homicides. *Journal of Urban Health: Bulletin of the New York Academy of Medicine* **91** 293-302.

WOLFERS, J. (2006). Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results. *The American Economic Review* **96** 1802–1820.

WOOLDRIDGE, J. M. (2005). Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models. *The Review of Economics and Statistics* **87** 385–390.

ZELLNER, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association* **57** 348-368.

ZELLNER, A. (1963). Estimators for Seemingly Unrelated Regression Equations: Some Exact Finite Sample Results. *Journal of the American Statistical Association* **58** 977-992.

ZHANG, F. and PENFOLD, R. B. (2013). Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements. *Academic Pediatrics* **13** S38–S44.

1 Proofs

1.1 Proof of Theorem 1

Proof. The proof is analogous to that of the ATT's identification under parallel trends in the canonical DID design. The descriptive difference between treated and control populations is

$$\mathrm{E}_{\mathcal{P}}\left[Y_T \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=1\right] - \left(\mathrm{E}_{\mathcal{P}}\left[Y_T \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=0\right]\right).$$

Then, given the model relating observed to potential outcomes in Assumption 1, namely, $Y_t = D_t Y_t(1) + (1 - D_t)\, Y_t(0)$, this descriptive difference can be expressed as

$$\mathrm{E}_{\mathcal{P}}\left[Y_T(1) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=1\right] - \left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=0\right]\right). \tag{1}$$

Equal-expected-prediction-errors in Assumption 2 then implies that

$$\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=1\right] = \mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right] - \left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=0\right]\right),$$

which, upon substituting this expression for $\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=1\right]$ in Eq. (1), yields

$$\begin{aligned}
&\mathrm{E}_{\mathcal{P}}\left[Y_T(1) \mid G=1\right] - \underbrace{\left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right] - \left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=0\right]\right)\right)}_{=\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=1\right]} \\
&\quad - \left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=0\right]\right) \\
&\quad = \mathrm{E}_{\mathcal{P}}\left[Y_T(1) - Y_T(0) \mid G=1\right] \\
&\quad = \mathrm{ATT},
\end{aligned}$$

thereby completing the proof. □

1.2 Proof of Proposition 1

Proof. The proof is immediate from the the ATT's lower and upper bounds in Eq. 9 in the manuscript: Given the sensitivity model in Eq. (9) in the manuscript, the difference between the upper and lower bounds of the ATT is

$$\begin{aligned}
&\delta_{f,T} + M \max_{v \in \mathcal{V}} |\delta_{f,v}| - \left(\delta_{f,T} - M \max_{v \in \mathcal{V}} |\delta_{f,v}|\right) \\
&\qquad = 2M \max_{v \in \mathcal{V}} |\delta_{f,v}|.
\end{aligned} \tag{2}$$

It follows immediately from Eq. (2) that, for a fixed $M \geq 0$, one model, f, will be (weakly) more robust than another model, f', if and only if

$$\max_{v \in \mathcal{V}} |\delta_{f,v}| \leq \max_{v \in \mathcal{V}} |\delta_{f',v}|.$$

□

1.3 Proof of Proposition 2

Proof. The supposition that equal expected prediction errors in Assumption 2 holds for f' implies, following Theorem 1, that we can express the true ATT as

$$\begin{aligned} \text{ATT} = &\mathrm{E}_{\mathcal{P}}\left[Y_T(1) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[f'(\boldsymbol{X}_T) \mid G=1\right] \\ &- \left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[f'(\boldsymbol{X}_T) \mid G=0\right]\right). \end{aligned} \tag{3}$$

Then taking the difference between

$$\mathrm{E}_{\mathcal{P}}\left[Y_T(1) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=1\right] - \left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=0\right]\right)$$

and the ATT in Eq. (3) yields

$$\left(\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[f'(\boldsymbol{X}_T) \mid G=0\right]\right) - \left(\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[f'(\boldsymbol{X}_T) \mid G=1\right]\right) \ ,$$

thereby completing the proof. □

1.4 Proof of Lemma 1

Proof. The proof proceeds in the following steps.

1. First, it shows that draws of coefficients from the multivariate Normal centered at the estimated coefficients with variance-covariance matrix equal to Eq. (22) in the manuscript are close to the population-level coefficients with probability limiting to 1.

2. Then the proof shows that the difference in average prediction errors, calculated over random draws from the aforementioned multivariate Normal, will be close to the population-level difference in expected prediction errors with probability limiting to 1.

3. Finally, the proof concludes by showing that the event in step 2 (occurring with probability limiting to 1) implies the event that the most robust model in the population minimizes the maximum absolute difference in average prediction errors over draws of coefficients from the multivariate Normal. Hence, in our procedure, the posterior probability of the truly most robust model will converge in probability to 1.

To carry out the proof via the steps above, first note that the weak law of large numbers (WLLN) implies that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ and $\hat{\boldsymbol{\Sigma}} \xrightarrow{p} 0$, which, by the continuous mapping theorem (CMT), implies that $\mathcal{N}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}\right)$ converges in probability to a constant whereby the probability that any

draw, $\hat{\boldsymbol{\beta}}^*$, is equal to $\boldsymbol{\beta}$ is 1. (This property can be established by taking the multivariate Normal's MGF and showing that it limits to the MGF of a multivariate constant.) Hence, it follows that, for all $\varepsilon > 0$,

$$\overset{*}{\Pr}\left(\left\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\right\|^2 \leq \varepsilon\right) \xrightarrow{p} 1,$$

where $\hat{\boldsymbol{\beta}}^*$ is a draw from $\mathcal{N}\left(\hat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}\right)$ conditional on sample data and $\Pr^*$ denotes conditional probability given sample data.

To show convergence in probability of the regression prediction in a sample to its population-level analogue, first write the average of the squared differences in predictions between $\hat{\boldsymbol{\beta}}^*_{f,v}$ and the population-level $\hat{\boldsymbol{\beta}}_{f,v}$ for any (f, v) as

$$\frac{1}{n_g}\sum_{i=1}^{n} \mathbb{1}\left\{G_i = g\right\}\left[\boldsymbol{X}_{i,t}\left(\hat{\boldsymbol{\beta}}^*_{f,v} - \boldsymbol{\beta}_{f,v}\right)\right]^2. \tag{4}$$

The Cauchy-Schwarz inequality implies that

$$\frac{1}{n_g}\sum_{i=1}^{n} \mathbb{1}\left\{G_i = g\right\}\left[\boldsymbol{X}_{i,v}\left(\hat{\boldsymbol{\beta}}^*_{f,v} - \boldsymbol{\beta}_{f,v}\right)\right]^2 \leq \left\|(\hat{\boldsymbol{\beta}}^*_{f,v} - \boldsymbol{\beta}_{f,v})\right\|^2 \frac{1}{n_g}\sum_{i=1}^{n} \mathbb{1}\left\{G_i = g\right\}\boldsymbol{X}_{i,v}\boldsymbol{X}_{i,v}^\top.$$

The WLLN implies that the second factor, $\frac{1}{n_g}\sum_{i=1}^{n} \mathbb{1}\left\{G_i = g\right\}\boldsymbol{X}_{i,v}\boldsymbol{X}_{i,v}^\top$, limits in probability to $\mathrm{E}_{\mathcal{P}}\left[\boldsymbol{X}_v^2 \mid G = g\right]$, where the regularity condition in Assumption 3 that $\mathrm{E}_{\mathcal{P}}\left[\|\boldsymbol{X}_v\|^2 \mid \mid G = g\right] < \infty$ implies that $\mathrm{E}_{\mathcal{P}}\left[\boldsymbol{X}_v^2 \mid G = g\right] < \infty$. Consequently, since $\left\|(\hat{\boldsymbol{\beta}}^*_{f,v} - \boldsymbol{\beta}_{f,v})\right\|^2 \xrightarrow{p} 0$, the CMT implies that

$$\left\|(\hat{\boldsymbol{\beta}}^*_{f,v} - \boldsymbol{\beta}_{f,v})\right\|^2 \frac{1}{n_g}\sum_{i=1}^{n} \mathbb{1}\left\{G_i = g\right\}\boldsymbol{X}_{i,v}\boldsymbol{X}_{i,v}^\top \xrightarrow{p} 0.$$

Since the upper-bound of (4) converges in probability to 0, so, too, must (4) itself.

The CMT then implies that

$$\hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}^*_{f,v}\right) \xrightarrow{p^*} \delta_{f,v}.$$

That is, for all $\varepsilon > 0$,

$$\overset{*}{\Pr}\left(\left|\hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}^*_{f,v}\right) - \delta_{f,v}\right| \leq \varepsilon\right) \xrightarrow{p} 1, \tag{5}$$

for all $(f, v) \in \mathcal{F} \times \mathcal{V}$.

Now define $\overline{v}_f$ as the validation period with the greatest absolute difference in expected

prediction errors under model f. Since (5) holds for all $\varepsilon > 0$, we can pick an $\varepsilon > 0$ such that, for all $f \in \mathcal{F}$,

$$\delta_{(f,\overline{v}_f)} - \varepsilon > \delta_{(f,\overline{v}_f)} + \varepsilon \tag{6}$$

for all $v \in \{\mathcal{V} \setminus \overline{v}_f\}$ and

$$\delta_{(f^\dagger,\overline{v}(f^\dagger))} - \varepsilon < \delta_{(f,\overline{v}_f)} + \varepsilon \tag{7}$$

for all $f \in \{\mathcal{F} \setminus f^\dagger\}$.

With $\varepsilon > 0$ satisfying (6) and (7), it follows that the event

$$\left|\hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}^*_{f,v}\right) - \delta_{f,v}\right| \text{ for all } (f,v) \in \mathcal{F} \times \mathcal{V}$$

implies the event that

$$f^\dagger = \arg\min_{f\in\mathcal{F}} \max_{v\in\mathcal{V}} \hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}^*_{f,v}\right). \tag{8}$$

Hence, (5) implies that

$$\overset{*}{\Pr}\left(f^\dagger = \arg\min_{f\in\mathcal{F}} \max_{v\in\mathcal{V}} \hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}^*_{f,v}\right)\right) \xrightarrow{p} 1, \tag{9}$$

thereby completing the proof.

□

1.5 Proof of Proposition 3

Proof. Lemma 1 and the law of total probability imply that $\hat{p}_f \xrightarrow{p} 0$ for all $f \in \left\{\mathcal{F} \setminus f^\dagger\right\}$. Since

$$\hat{\delta}\left(\boldsymbol{D}_t, \hat{\boldsymbol{\beta}}_{f,t}\right) \xrightarrow{p} \delta_{(f,t)}$$

for all $(f,t) \in \mathcal{F} \times \mathcal{T}$, the CMT implies that

$$\hat{\delta}\left(\boldsymbol{D}_t, \hat{\boldsymbol{\beta}}_{f^\dagger,T}\right) \xrightarrow{p} \delta_{(f^\dagger,T)},$$

from which another application of the CMT implies the result. □

2 Additional special cases of identification assumption

2.1 Sequential DID

Sequential DID relies on a parallel-trends-in-trends assumption in which each group's average outcome in period $T-1$ plus the group's change in average outcomes from periods $T-2$ to $T-1$ is equal to each group's expected untreated potential outcome [7, 12, 14, 15, 17]. We formally write parallel trends-in-trends as

Parallel trends-in-trends $\coloneqq$

$$\begin{aligned}(10)\quad &\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] - \left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right]\right) = \\ &\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=1\right] - \left(\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=0\right]\right).\end{aligned}$$

This can be generalized to K time-wise differences [see, e.g., 12], but for simplicity, we focus on $K=2$.

If the prediction function is

$$(11)\qquad f(\boldsymbol{X}_t) = Y_{t-1} + (Y_{t-1} - Y_{t-2}) \text{ for } t = 3, \ldots, T$$

then Assumption 2 will be true whenever Eq. (10) holds.

First, parallel trends-in-trends in Eq. (10) implies that

$$\begin{aligned}\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right] &= \left(\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] + \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=1\right]\right) + \mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] \\ &\quad - \left(\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right] + \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=0\right]\right) \\ \mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] &= \left(\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right] + \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=0\right]\right) + \mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right] \\ &\quad - \left(\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] + \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=1\right]\right).\end{aligned}$$

Then the prediction function in Eq. (11) and Assumption 1 imply that

$$\begin{aligned}&\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=1\right] = \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] + \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=1\right] \text{ and} \\ &\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=0\right] = \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right] + \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=0\right],\end{aligned}$$

which implies that the expected prediction errors in each group are

$$\begin{aligned}\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=1\right] &= -\left(\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right] + \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=0\right]\right) \\ \mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=0\right] &= -\left(\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] + \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=1\right]\right).\end{aligned}$$

Therefore, the difference in expected prediction errors is

$$\begin{aligned}&\left(\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right]+\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right]-\mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=1\right]\right)\\&-\left(\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right]+\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right]-\mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=0\right]\right),\end{aligned}$$

which parallel trends-in-trends in Eq. (10) implies is equal to 0, thereby completing the proof.

2.2 Unit- or group-specific time trends

In contrast to methods that assume similar time dynamics in treated and comparison groups, comparative interrupted time series methods explicitly model differential time trends in the two groups. A fully linear implementation measures changes in intercepts and slopes across the two groups, but a more flexible version of comparative interrupted time series measures period-by-period differences from an extrapolated linear trend in each individual or group [5, 19].

Like TWFE, this method assumes a parametric structural model for the untreated potential outcomes

$$Y_t(0)=\xi_u t+\gamma_t+\epsilon_{u,t} \tag{12}$$

where ξ_u is the linear time slope of the u^{th} unit and $\mathrm{E}[\epsilon_{u,t} \mid \xi_u, G_u]=0$ for all $u=1,\ldots,U$ and $t=1,\ldots,T$. With this model, we can show that there exists a prediction function such that when Eq. (12) holds, equal expected prediction errors holds also.

If the prediction function is

$$f(\boldsymbol{X}_t)=\hat{\xi}_u t \text{ where } \hat{\xi}_u=\underset{\xi_u}{\arg\min}\sum_{s=1}^{t-1}\left(Y_{u,s}-\xi_u s\right)^2, \tag{13}$$

then Assumption 2 will hold whenever the structural model in Eq. (12) is true.

First, the structural model in Eq. (12) implies that

$$\begin{aligned}\mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) \mid G_u=1\right]&=\mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u=1\right]+\gamma_T\\\mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) \mid G_u=0\right]&=\mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u=0\right]+\gamma_T.\end{aligned}$$

Second, note that the solution to the empirical risk minimization problem for ξ_u in periods before T is

$$\hat{\xi}_u=\frac{\sum_{t=1}^{T-1} tY_{u,t}}{\sum_{t=1}^{T-1} t^2},$$

which, from the linear time trend model in Eq. (12), can be expressed as

$$\begin{aligned}\hat{\xi}_u &= \frac{\sum_{t=1}^{T-1} t\left(\xi_u t + \gamma_t + \epsilon_{u,t}\right)}{\sum_{t=1}^{T-1} t^2} \\ &= \xi_u + \frac{\sum_{t=1}^{T-1} t\gamma_t}{\sum_{t=1}^{T-1} t^2} + \frac{\sum_{t=1}^{T-1} t\epsilon_{u,t}}{\sum_{t=1}^{T-1} t^2}\end{aligned}$$

It follows further that the prediction for unit u in period T is

$$\begin{aligned}f(\boldsymbol{X}_{u,T}) &= \hat{\xi}_u T \\ &= \xi_u T + \left(\frac{\sum_{t=1}^{T-1} t\gamma_t}{\sum_{t=1}^{T-1} t^2}\right) T + \left(\frac{\sum_{t=1}^{T-1} t\epsilon_{i,t}}{\sum_{t=1}^{T-1} t^2}\right) T.\end{aligned}$$

Then, due to the structural model in Eq. (12) and since all $t = 1, \ldots, T$ are fixed constants, it follows that

$$\begin{aligned}\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_{u,T}) \mid G_u = 1\right] &= \mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u = 1\right] + \left(\frac{\sum_{t=1}^{T-1} t\gamma_t}{\sum_{t=1}^{T-1} t^2}\right) T \\ \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_{u,T}) \mid G_u = 0\right] &= \mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u = 0\right] + \left(\frac{\sum_{t=1}^{T-1} t\gamma_t}{\sum_{t=1}^{T-1} t^2}\right) T.\end{aligned}$$

Finally, the model in Eq. (12) implies that the difference in expected prediction errors is equal to 0:

$$\begin{aligned}&\mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) - f(\boldsymbol{X}_{u,T}) \mid G_u = 1\right] - \left(\mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) - f(\boldsymbol{X}_{u,T}) \mid G_u = 0\right]\right) \\ &= \mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u = 1\right] + \gamma_T - \mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u = 1\right] - \left(\frac{\sum_{t=1}^{T-1} t\gamma_t}{\sum_{t=1}^{T-1} t^2}\right) T \\ &- \left(\mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u = 0\right] + \gamma_T - \mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u = 0\right] - \left(\frac{\sum_{t=1}^{T-1} t\gamma_t}{\sum_{t=1}^{T-1} t^2}\right) T\right) \\ &= 0,\end{aligned}$$

where the last line follows from the fact that t and γ_t are always equal across units.

A complementary question is, as a reviewer wondered, "whether it is possible for nonparametric identification to not hold, but the proposed assumption to hold." Indeed, we can use the structural model above to illustrate such a case. Suppose the true structural model is comparative interrupted time series as implemented by Bloom and Riccio [5], which is equivalent to a flexible difference-in-differences specification that uses time fixed effects and group-specific linear trends [8]. The nonparametric parallel trends assumption in Eq. (6) of the manuscript, clearly will not hold because the truth is differential trends in the two groups. However, Assumption 2 will still hold for the prediction function that incorporates differential trends in the two groups into its predictions.

2.3 Lagged dependent variable model

Lagged dependent variable (LDV) models incorporate across-time dependence of outcomes within units. In political science, authors have debated the merits of autoregressive distributed lag (ADL) models that include lags of both outcomes and treatments [3]. The relationship between lags of treatment and lags of the outcome is complicated by a classic observation about the bias of unit fixed effects in autoregressive models [16]. Thus, one recent comparison across model specifications argued that LDV models should use first differences [9]. Other authors have argued for a specification that includes the full vector of pre-treatment outcomes (like a regression analogue of synthetic controls) [18]. Other authors have emphasized the causal assumptions, including whether past treatments can affect current outcomes and whether past outcomes can affect current treatment [10], the problem of conditioning on post-treatment outcomes [4], and the relationship between the causal assumptions of difference-in-differences and methods that, like LDV, condition on past outcomes [6].

As with the approaches above, we focus on a basic implementation of LDV methods that uses a structural model for the untreated potential outcomes

$$Y_t(0) = \gamma_t + \lambda Y_{t-1}(0) + \epsilon_t \ , \tag{14}$$

where λ is a parameter that controls the strength of the dependence and $\mathrm{E}_{\mathcal{P}}[\epsilon_t] = 0$ for $t = 1, \ldots, T$.[1] Notice that this resembles the two-way fixed effects model of Eq. (7), but instead of unit-level (time-invariant) fixed effects, it includes unit-level dependence on past outcomes. Then we assume a form of exogeneity conditional on past outcomes,

$$\textbf{Exogeneity conditional on past outcomes} \coloneqq \mathrm{E}_{\mathcal{P}}\left[\epsilon_t \mid Y_{t-1}, Y_{t-2}, \ldots, Y_1, G\right] = 0 \ . \tag{15}$$

There exists a prediction function such that when Equations (14) and (15) both hold, so does equal

[1] We could generalize this to dependence on outcomes with lag 2, 3, etc. We use lag-1 outcome dependence for simplicity.

expected prediction errors.

If the prediction function is

$$f(\boldsymbol{X}_t) = \hat{\lambda} Y_{t-1} \text{ where } \hat{\lambda} = \arg\min_{\lambda} \sum_{s=2}^{t-1} \left(\tilde{Y}_s - \lambda \tilde{Y}_{s-1} \right)^2, \tag{16}$$

where $\tilde{Y}_t \coloneqq Y_t - \mathrm{E}_{\mathcal{P}}[Y_t]$ for all $t = 1, \ldots, T$, then Assumption 2 will hold whenever the outcome model in Eq. (14) and exogeneity in Eq. 15 are true.

First, note that the structural model in Eq. (14) implies that

$$\begin{aligned} \mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G = 1\right] &= \mathrm{E}_{\mathcal{P}}\left[\lambda Y_{T-1}(0) \mid G = 1\right] + \gamma_T \\ \mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G = 0\right] &= E\left[\lambda Y_{T-1}(0) \mid G = 0\right] + \gamma_T. \end{aligned}$$

Second, the solution to the empirical risk minimization problem for λ in periods before T is

$$\hat{\lambda} = \frac{\sum_{t=2}^{T-1} \tilde{Y}_{t-1} \tilde{Y}_t}{\sum_{t=2}^{T-1} \tilde{Y}_{t-1}^2}. \tag{17}$$

Given the equivalent representation of the LDV model in Eq. (14) in which outcomes, predictors and the error term are centered by their means across units for each time period [11], the solution to the empirical risk minimization problem in Eq. (17) can be expressed as

$$\hat{\lambda} = \lambda + \frac{\sum_{t=2}^{T-1} \tilde{Y}_{t-1} \tilde{\epsilon}_t}{\sum_{t=2}^{T-1} \tilde{Y}_{t-1}^2}.$$

It follows that the prediction in period T is

$$\begin{aligned} f(\boldsymbol{X}_T) &= \hat{\lambda} Y_{T-1} \\ &= \lambda Y_{T-1} + \left(\frac{\sum_{t=2}^{T-1} \tilde{Y}_{t-1} \tilde{\epsilon}_t}{\sum_{t=2}^{T-1} \tilde{Y}_{t-1}^2} \right) Y_{T-1}, \end{aligned}$$

which exogeneity in Eq. (15) then implies has expectations in treated and control groups equal to

$$\begin{aligned} \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G = 1\right] &= \lambda \, \mathrm{E}_{\mathcal{P}}\left[Y_{T-1} \mid G = 1\right] \\ \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G = 1\right] &= \lambda \, \mathrm{E}_{\mathcal{P}}\left[Y_{T-1} \mid G = 0\right]. \end{aligned}$$

The LDV model in Eq. (14) further implies that the expected prediction errors in treated and

comparison groups are

$$\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=1\right] = \gamma_T + \lambda\,\mathrm{E}_{\mathcal{P}}\left[Y_{T-1} \mid G=1\right] - \lambda\,\mathrm{E}_{\mathcal{P}}\left[Y_{T-1} \mid G=1\right] = \gamma_T$$
$$\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G=0\right] = \gamma_T + \lambda\,\mathrm{E}_{\mathcal{P}}\left[Y_{T-1} \mid G=0\right] - \lambda\,\mathrm{E}_{\mathcal{P}}\left[Y_{T-1} \mid G=0\right] = \gamma_T.$$

It then follows immediately that the difference in expected prediction errors is equal to 0, thereby completing the proof.

2.4 Synthetic controls

Suppose that we are studying a single treated unit (denote it $u = 1$ without loss of generality). The synthetic control weights, denoted by w_u for unit u, is the solution to a regularized minimization of the mean squared difference between the treated unit's outcomes and the weighted control outcomes at each pre-period time,

$$\frac{1}{T-1}\sum_{t=1}^{T-1}\left(Y_{1,t} - \frac{1}{N-1}\sum_{u=2}^{N} w_u Y_{u,t}\right)^2.$$

(This is slightly simplified because it omits the penalty term.) The synthetic control estimator, as originally proposed by Abadie [1], is simply

$$Y_{1,T} - \frac{1}{U-1}\sum_{u=2}^{U} Y^{\ddagger}_{u,t}, \tag{18}$$

where $Y^{\ddagger}_{u,t} = w_u Y_{u,t}$ for $u = 2, \ldots, U$ are the comparison units' weighted outcomes.

Following the proofs above, we would want to know whether some prediction function, when combined with the identifying assumption of synthetic controls, implies that Assumption 2 also holds. What then, is the identifying assumption of synthetic controls? As far as we can tell, synthetic controls began with an estimation method and then suggested a structural model (interactive fixed effects) that would justify that estimator. Therefore, we instead propose prediction functions that, under Assumption 2, would yield the difference in conditional expectations of Eq. 18, i.e.,

$$\mathrm{E}_{\mathcal{P}}\left[Y_{1,T} \mid G_1 = 1\right] - \mathrm{E}_{\mathcal{P}}\left[\frac{1}{U-1}\sum_{u=2}^{U} Y^{\ddagger}_{u,t} \mid G_u = 0\right]. \tag{19}$$

Thus, whenever the synthetic control assumption is met, Assumption 2 will be also.

First, suppose that the prediction function in period T is $f(\boldsymbol{X}_T) = 0$ and we had weighted the comparison units by w_u as a "pre-processing" step. Then the ATT implied by Eq. 5 would be

$$ATT = \mathrm{E}_{\mathcal{P}}\left[Y_{1,T} - 0\right] - \mathrm{E}_{\mathcal{P}}\left[Y^{\ddagger}_{u,T} - 0 \mid G_u = 0\right]$$

which is the synthetic control difference in conditional expectations in Eq. 19.

Alternatively, suppose we considered the weighting to be part of the prediction function and used the comparison group's period T outcomes to "predict" for both groups,

$$f(\boldsymbol{X}_T) = \begin{cases} \mathrm{E}_{\mathcal{P}}\left[Y^{\ddagger}_{u,T} \mid G_u = 0\right] & \text{if } G_u = 1 \\ \\ Y_{u,T} & \text{if } G_u = 0 \end{cases}$$

Then the ATT implied by Eq. (5) would be

$$ATT = \left(\mathrm{E}_{\mathcal{P}}\left[Y_{1,T} \mid G_u = 1\right] - \mathrm{E}_{\mathcal{P}}\left[Y^{\ddagger}_{u,T} \mid G_u = 0\right]\right) - \mathrm{E}_{\mathcal{P}}\left[Y_{u,T} - Y_{u,T} \mid G_u = 0\right],$$

which is again the synthetic control difference in conditional expectations in Eq. 19.

Both prediction functions are trivial because they do no actual prediction: one is a constant (0) and the other is based on current outcomes across groups, not past within groups. This makes sense because the synthetic control method involves only a treated-vs-comparison contrast, not a pre-vs-post contrast. The pre-period's only contribution in synthetic controls is to inform the weights. When we try to fit synthetic controls into our "predict, correct" paradigm, we find that it involves only the correction step without the prediction step.

Nevertheless, synthetic control weights may still be useful if we believe that weighting by similarity on pre-period outcomes helps us select a more suitable comparison group. We can weight as a pre-processing step, then apply our methods to the weighted combination of comparison units. Others have combined DID and synthetic controls [e.g., 2], and we envision this to be a fruitful topic for further research.

2.5 Interactive fixed effects

We now use an interactive fixed effects structural model to demonstrate an example (inspired by a reviewer) in which a parametric structural model holds, but, given a specific prediction function, equal expected prediction errors in Assumption 2 does not. Here we show that the prediction model (corresponding to TWFE) in Eq. (19) of the manuscript does not imply equal expected prediction errors when the structural model is that of interactive fixed effects — an unsurprising result given that the interactive fixed effect models implies that time shocks differ between treated and comparison groups. That said, our argument below does not rule out the possibility that another prediction function could be found that does imply equal expected prediction errors [perhaps drawing upon 13]; however, it is unclear whether such an appropriate prediction function would conduct the "predict" step from pre-period data within groups, in accordance with controlled pre-post designs to which our argument pertains.

Suppose untreated potential outcomes are generated by an interactive fixed effects structural

model,

$$Y_{u,t}(0) = \alpha_u + \gamma_t + \nu_u F_t + \epsilon_{u,t}, \tag{20}$$

where ν_u is an unobserved, unit-specific "loading" of the unobserved common factor, F_t, and $\mathrm{E}_{\mathcal{P}}\left[\epsilon_{u,t} \mid \alpha_u, \nu_u, G_u\right] = 0$ for all $u = 1, \ldots, U$ and $t = 1, \ldots, T$. Taking expectation, the treated and comparison groups' expected untreated potential outcomes in the post-treatment period are

$$\begin{aligned}
\mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) \mid G_u = 1\right] &= \mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 1\right] + \gamma_T + \mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 1\right] \\
\mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) \mid G_u = 0\right] &= \mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 0\right] + \gamma_T + + \mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 0\right].
\end{aligned}$$

Consider the prediction function in (19), which is simply each unit's average outcome prior to t:

$$\arg\min_{\alpha_u} \sum_{s=1}^{t-1} \left(Y_{u,s} - \alpha_u\right)^2 = \frac{1}{(t-1)} \sum_{s=1}^{t-1} Y_{u,s}. \tag{21}$$

With this prediction function, the prediction in period T is

$$f(\boldsymbol{X}_T) = \frac{1}{(T-1)} \sum_{t=1}^{T-1} Y_{u,t},$$

which, by the consistency assumption, is

$$f(\boldsymbol{X}_T) = \frac{1}{(T-1)} \sum_{t=1}^{T-1} Y_{u,t}(0).$$

The IFE model in Eq. (20) implies that the expectations of the predictions in the treated and control groups are

$$\begin{aligned}
\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G_u = 1\right] &= \frac{1}{(T-1)} \left[\sum_{t=1}^{T-1} \left(\mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 1\right] + \gamma_t + \mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 1\right]\right)\right] \\
&= \frac{1}{(T-1)} \left[\sum_{t=1}^{T-1} \mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 1\right] + \sum_{t=1}^{T-1} \gamma_t + \sum_{t=1}^{T-1} \mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 1\right]\right] \\
&= \mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 1\right] + \frac{1}{(T-1)} \sum_{t=1}^{T-1} \gamma_t + \frac{1}{(T-1)} \sum_{t=1}^{T-1} \mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 1\right]
\end{aligned}$$

and

$$\mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G_u = 0\right] = \mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 0\right] + \frac{1}{(T-1)}\sum_{t=1}^{T-1}\gamma_t + \frac{1}{(T-1)}\sum_{t=1}^{T-1}\mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 0\right].$$

The IFE model in Eq. (20) also implies the expected prediction errors in each group are

$$\begin{aligned}
\mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) \mid G_u = 1\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G_u = 1\right] &= \mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 1\right] + \gamma_T + \mathrm{E}_{\mathcal{P}}\left[\nu_u F_T \mid G_u = 1\right] \\
&\quad - \Bigg(\mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 1\right] + \frac{1}{(T-1)}\sum_{t=1}^{T-1}\gamma_t \\
&\quad + \frac{1}{(T-1)}\sum_{t=1}^{T-1}\mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 1\right] \Bigg) \\
&= \gamma_T - \sum_{t=1}^{T-1}\gamma_t + \mathrm{E}_{\mathcal{P}}\left[\nu_u F_T \mid G_u = 1\right] - \sum_{t=1}^{T-1}\mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 1\right]
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}_{\mathcal{P}}\left[Y_{u,T}(0) \mid G_u = 0\right] - \mathrm{E}_{\mathcal{P}}\left[f(\boldsymbol{X}_T) \mid G_u = 0\right] &= \mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 0\right] + \gamma_T + \mathrm{E}_{\mathcal{P}}\left[\nu_u F_T \mid G_u = 0\right] \\
&\quad - \Bigg(\mathrm{E}_{\mathcal{P}}\left[\alpha_u \mid G_u = 0\right] + \frac{1}{(T-1)}\sum_{t=1}^{T-1}\gamma_t \\
&\quad + \frac{1}{(T-1)}\sum_{t=1}^{T-1}\mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 0\right] \Bigg) \\
&= \gamma_T - \sum_{t=1}^{T-1}\gamma_t + \mathrm{E}_{\mathcal{P}}\left[\nu_u F_T \mid G_u = 0\right] - \sum_{t=1}^{T-1}\mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 0\right].
\end{aligned}$$

Taking the difference in expected prediction errors yields

$$\mathrm{E}_{\mathcal{P}}\left[\nu_u F_T \mid G_u = 1\right] - \mathrm{E}_{\mathcal{P}}\left[\nu_u F_T \mid G_u = 0\right] - \frac{1}{T-1}\left(\sum_{t=1}^{T-1}\mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 1\right] - \mathrm{E}_{\mathcal{P}}\left[\nu_u F_t \mid G_u = 0\right]\right),$$

which is not necessarily equal to 0.

3 Model implementation in the applied analysis

Table 1: Candidate prediction models used in the analysis of Missouri's repeal of permit-to-purchase.

Baseline Mean	$Y_t \sim \beta_0$
Baseline Mean (log)	$\log(Y_t) \sim \beta_0$
Baseline Mean (first diff)	$Y_t - Y_{i,t-1} \sim \beta_0$
Lin Time Trend	$Y_t \sim \beta_0 + \beta_1 t$
Lin Time Trend (log)	$\log(Y_t) \sim \beta_0 + \beta_1 t$
Lin Time Trend (first diff)	$Y_t - Y_{i,t-1} \sim \beta_0 + \beta_1 t$
Quad Time Trend	$Y_t \sim \beta_0 + \beta_1 t^2$
Quad Time Trend (log)	$\log(Y_t) \sim \beta_0 + \beta_1 t^2$
Quad Time Trend (first diff)	$Y_t - Y_{i,t-1} \sim \beta_0 + \beta_1 t^2$
LDV	$Y_t \sim \beta_0 + \beta_2 Y_{i,t-1}$
LDV (log)	$\log(Y_t) \sim \beta_0 + \beta_2 \log(Y_{i,t-1})$
LDV (first diff)	$Y_t - Y_{i,t-1} \sim \beta_0 + \beta_2 Y_{i,t-1}$
Lin Time Trend + LDV	$Y_t \sim \beta_0 + \beta_1 t + \beta_2 Y_{i,t-1}$
Lin Time Trend + LDV (log)	$\log(Y_t) \sim \beta_0 + \beta_1 t + \beta_2 \log(Y_{i,t-1})$
Lin Time Trend + LDV (first diff)	$Y_t - Y_{i,t-1} \sim \beta_0 + \beta_1 t + \beta_2 Y_{i,t-1}$
Quad Time Trend + LDV	$Y_t \sim \beta_0 + \beta_1 t^2 + \beta_2 Y_{i,t-1}$
Quad Time Trend + LDV (log)	$\log(Y_t) \sim \beta_0 + \beta_1 t^2 + \beta_2 Y_{i,t-1}$
Quad Time Trend + LDV (first diff)	$Y_t - Y_{i,t-1} \sim \beta_0 + \beta_1 t^2 + \beta_2 Y_{i,t-1}$

4 Simulation Studies

We conduct several simulation studies to assess the performance of our Bayesian model averaged (BMA) estimator, with special focus on the coverage of its confidence intervals. We use the same simulation setup as Schell et al. [20]. This setup is especially compelling in a setting related to gun policy, and for that reason has been adopted in closely related simulation studies [5].

The simulation setup from Schell et al. [20] consists of crude death rates in all 50 states in each year from 1979 to 2014. We focus on years 1994 to 2008 and suppose that 2008 is the only post-treatment year. Akin to our application in Sec. 5 above, we let the years 1994 to 1998 serve as training years and let 1999 to 2007 serve as validation years. We randomly sample 5 states to serve as "treated," which begins in 2008. The remaining 45 states are the "comparison" states.

We consider a class of 6 candidate models: (1) baseline mean, $Y_t \sim \beta_0$, (2) linear time trend, $Y_t \sim \beta_0 + \beta_1 t$, (3) LDV, $Y_t \sim \beta_0 + \beta_2 Y_{t-1}$, (4) linear time trend + LDV, Y_t $\beta_0 + \beta_1 t + \beta_2 Y_{t-1}$, (5) baseline mean (first diff), $Y_t - Y_{t-1}$ β_0 and (6) linear time trend (first diff), $Y_t - Y_{t-1} \sim \beta_0 + \beta_1 t$. Performing our procedure on the population data shows that model (1), the simple baseline mean model minimizes our sensitivity criterion (i.e., the worst-case absolute prediction error in the pre-treatment validation periods).

To conduct our simulations, we treat the 5 treated states and 45 control states as the population of interest and consider properties of our BMA estimator over 1000 random draws with replacement of states from this population. For each draw, we sample with replacement a fixed number from the distribution of treated states and a fixed number from the distribution of control states. This sampling corresponds to the usual assumption (made in our paper) of independent and identically distributed random sampling of units (in this case states) within groups.

Over each realization of sample data, we record the posterior probability that each of the 6 candidate models is most robust. We also record the BMA estimate, the variance in estimates across models with respect to the posterior, and the bootstrap distribution of BMA estimates over repeated draws from the empirical distribution of the sample data, holding the original posterior distribution fixed. To construct 95% confidence intervals, we use a Normal approximation in which the lower bound is the BMA estimate minus 1.96 multiplied by the square root of the overall variance (the sum of the Eq. (16) and Eq. (17) in the manuscript. The upper bound of the 95% confidence interval is constructed analogously. A confidence interval covers the target if it brackets the true difference in expected trends under the truly optimal model for period T in the population. The bias of the BMA estimator also refers to this target.

We conduct our simulations under an increasing number of sampled units, holding the ratio of treated to control units fixed. We begin with 1 treated unit and 8 controls, which mirrors the setting of our application in Sec. 5. We then increase the number of treated units to 3, 15 and 35 with 24, 120 and 280 control units, respectively.

Figure 1 below shows the distribution of expected posterior probabilities for each of the 6

candidate models under varying sample sizes. In all settings, even with only 1 treated unit and 8 control units, the truly optimal model in the population (the baseline mean model) receives a substantial probability. As we would expect, the expected posterior probability of this model increases in sample size. In the largest sample with 35 treated states and 280 control states, the expected posterior probability of the truly optimal model is approximately 0.8, compared to roughly 0.36 in the smallest sample.

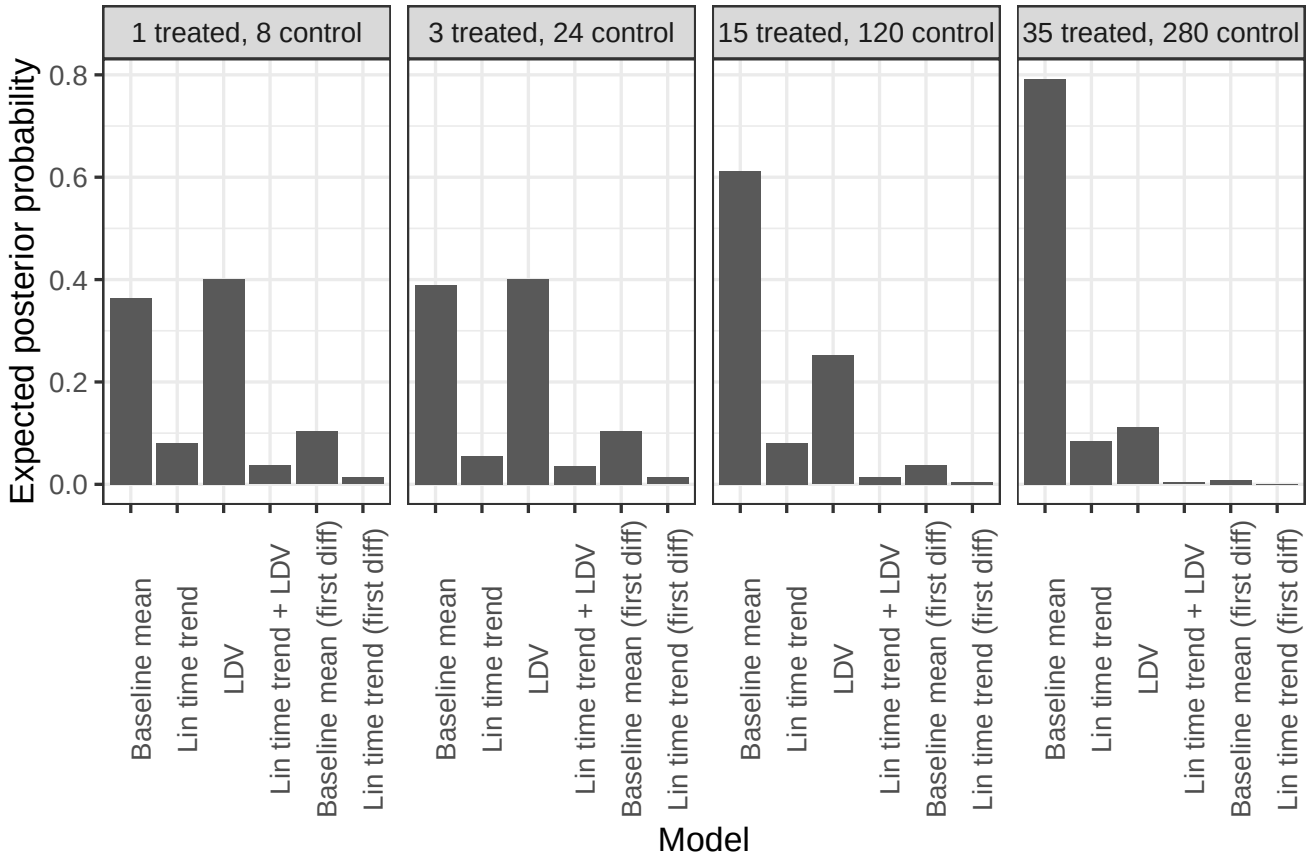

Figure 1

In Figure 2 below, we report the absolute percent bias of the BMA estimator, the coverage probabilities of 95% confidence intervals and the ratio of the expected variance estimator to the true monte carlo variance of the BMA estimator. As we would expect bias is substantial in small samples, but diminishes in large samples. In the smallest sample, the coverage of our 95 % confidence intervals is 0.85, but achieves nominal rates once the sample size becomes moderately large. In moderately large samples, our variance estimator is conservative, as Antonelli, Papadogeorgou and Dominici [6] suggest. In all but the smallest sample size, the ratio of the expected variance estimate to the true monte carlo variance of the BMA estimator is at least as great as 1.

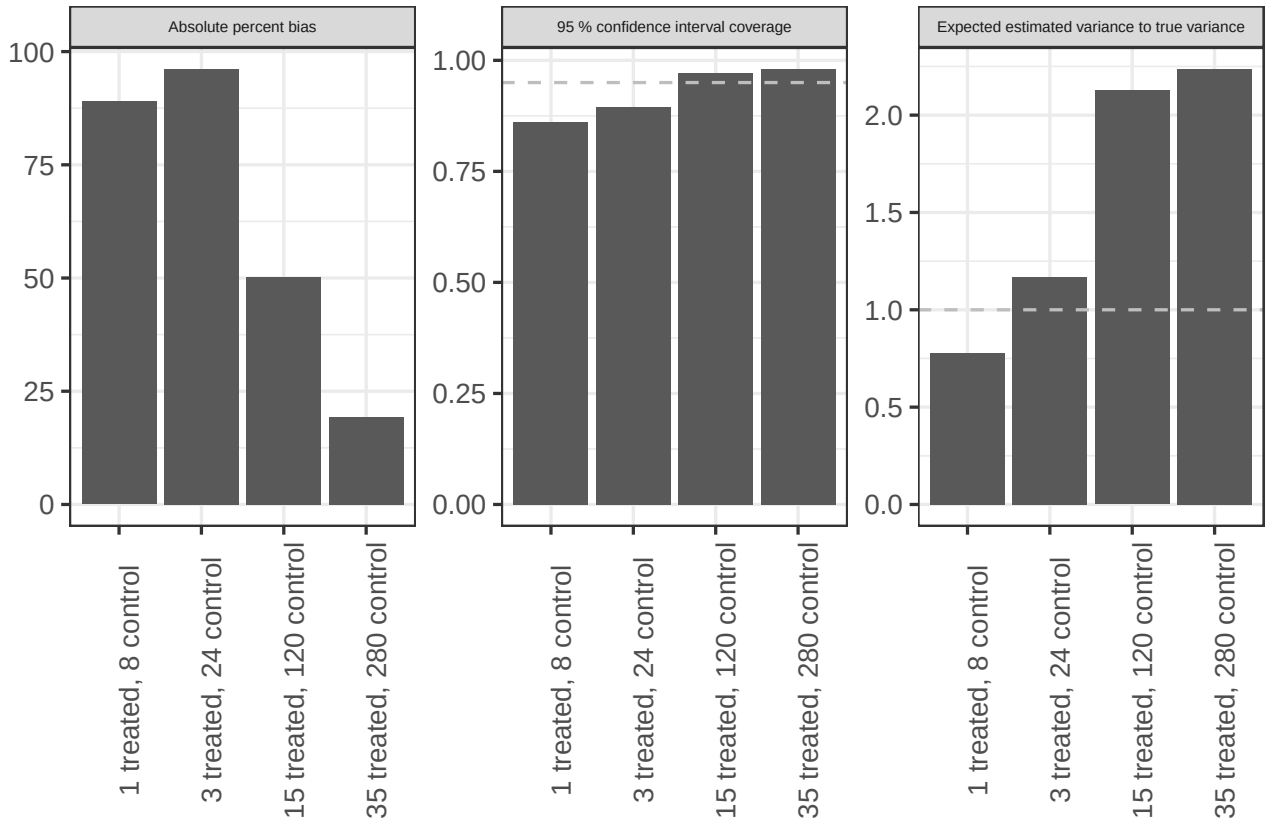
Absolute percent bias
100
75
50
25
0
95 % confidence interval coverage
1.00
0.75
0.50
0.25
0.00
Expected estimated variance to true variance
2.0
1.5
1.0
0.5
0.0
1 treated, 8 control
3 treated, 24 control
15 treated, 120 control
35 treated, 280 control

Figure 2

References

[1] A. Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005. 10

[2] D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. Synthetic difference in differences. *American Economic Review*, 111(12):4088–4118, 2021. 11

[3] N. Beck and J. N. Katz. Modeling dynamics in time-series-cross-section political economy data. *Annual Review of Political Science*, 14(1):331–352, 2011. 8

[4] M. Blackwell and A. N. Glynn. How to make causal inferences with time-series cross- sectional data under selection on observables. *The American Political Science Review*, 112(4):1067–1082, 2018. 8

[5] H. S. Bloom and J. A. Riccio. Using place-based random assignment and comparative interrupted time-series analysis to evaluate the Jobs-Plus employment program for public housing residents. *The Annals of the American Academy of Political and Social Science*, 599 (1):19–51, 2005. 6, 8

[6] P. Ding and F. Li. A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*, 27(4):605–615, 2019. 8

[7] N. Egami and S. Yamauchi. Using multiple pre-treatment periods to improve Difference-in-Differences and Staggered Adoption designs. *Political Analysis*, 2022. 5

[8] C. E. Fry and L. A. Hatfield. Birds of a feather flock together: Comparing controlled pre-post designs. *Health Services Research*, 56(5):942–952, 2021. 8

[9] B. A. Griffin, M. S. Schuler, E. A. Stuart, S. Patrick, E. McNeer, R. Smart, D. Powell, B. D. Stein, T. L. Schell, and R. L. Pacula. Moving beyond the classic difference-in-differences model: A simulation study comparing statistical methods for estimating effectiveness of state-level policies. *BMC Medical Research Methodology*, 21(279), 2021. 8

[10] K. Imai and I. S. Kim. When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, 63(2):467–490, 2019. 8

[11] J. Kropko and R. Kubinec. Interpretation and identification of within-unit and cross-sectional variation in panel data models. *PLoS ONE*, 15(4):e0231349, 2020. 9

[12] M.-j. Lee. Generalized difference in differences with panel data and least squares estimator. *Sociological Methods & Research*, 45(1):134–157, 2016. 5

[13] L. Liu, Y. Wang, and Y. Xu. A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 68(1):160–176, 2024. 11

[14] R. Mora and I. Reggio. Treatment effect identification using alternative parallel assumptions. Working Paper, Economic Series (48) 12–33, Universidad Carlos III, Getafe, Spain, December 2012. 5

[15] R. Mora and I. Reggio. Alternative diff-in-diffs estimators with several pretreatment periods. *Econometric Reviews*, 38(5):465–486, 2019. 5

[16] S. Nickell. Biases in dynamic models with fixed effects. *Econometrica*, 49(6):1417–1426, 1981. 8

[17] A. Olden and J. Møen. The triple difference estimator. *The Econometrics Journal*, 25(3): 531–553, 2022. 5

[18] S. O'Neill, N. Kreif, R. Grieve, M. Sutton, and J. S. Sekhon. Estimating causal effects: Considering three alternatives to difference-in-differences estimation. *Health Services and Outcomes Research Methodology*, 16(1):1–21, 2016. 8

[19] J. A. Riccio and H. S. Bloom. Extending the reach of randomized social experiments: New directions in evaluations of American welfare-to-work and employment initiatives. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 165(1):13–30, 2002. 6

[20] T. L. Schell, B. A. Griffin, and A. R. Morral. *Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths: A Simulation Study*. RAND Corporation, Santa Monica, CA, 2018. 15