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 may use a controlled pre-post design in which they observe over-time changes in crime among a population exposed to a change in firearm law and compare to the changes in 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 (CITS) 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 (PTP) 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 PTP laws (see Figure 1).

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

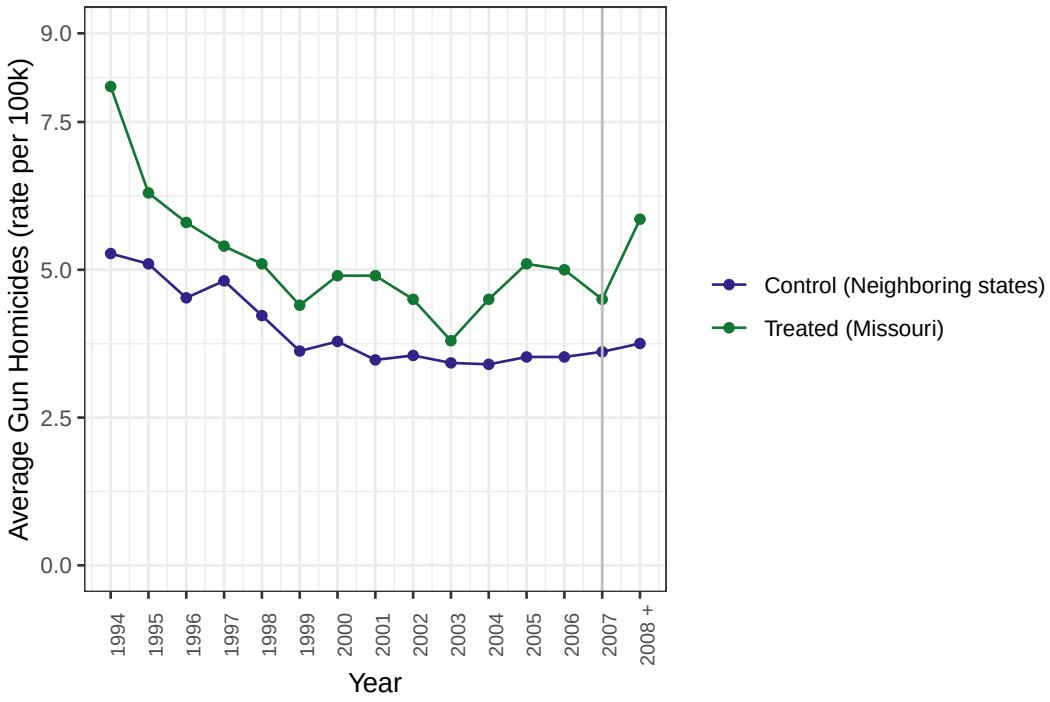

FIG 1. *Average gun homicides (rate per 100,000) before and after the 2007 permit-to-purchase repeal in Missouri (treated state) and control states (Arkansas, Illinois, Iowa, Kansas, Kentucky, Nebraska, Oklahoma, and Tennessee)*

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, CITS 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 PTP 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. Like most researchers, each argued that their assumptions were plausible and their conclusions correct.

In this paper, we move away from the question of model plausibility or correctness and focus on another criterion: robustness. Robustness in our framework is the (lack of) change in our conclusions 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. 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.

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 (i.e., we need not debate the relative plausibility of the DID identification assumption versus the CITS identification assumption). 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 method for causal inference in 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.

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 two popular designs (DID and two-way fixed effects (TWFE) models) are special cases of our framework. Then in Section 3, we introduce a sensitivity analysis framework that motivates our model selection procedure. We provide a data-driven algorithm to select a model that maximizes robustness in Section 4 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 (DGP) 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, wherein $Y_t(0)$ denotes the untreated potential outcome in any of the period $t = 1, \ldots, T$ periods and $Y_T(1)$ denotes the treated

potential outcome in the post-treatment period, T. Our causal target is the ATT, which is defined with respect to the population-level DGP. The ATT is defined as

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

where $\mathrm{E}_P[\cdot]$ denotes an expectation taken over a population-level joint cumulative distribution function (CDF).

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 any given time is the potential outcome corresponding to the treatment condition at that time. Assumption 1 rules out two problematic scenarios. First, it rules out treatment anticipation, in which the treated group manifests treated outcomes before treatment begins. Second, it rules out spillovers, in which, say, the untreated group manifests treated potential outcomes despite no direct exposure to treatment.

With Assumption 1, we express the ATT as

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

replacing the treated potential outcomes with the observed outcome since treated potential outcomes can be observed in the post-period. It remains to replace the untreated potential outcome with an observable quantity since the treated group's untreated potential outcome is unobservable in the post-period.

To do so, let $\boldsymbol{X}_t$ denote the collection of predictors for untreated potential outcomes in period t. These predictors can include prior untreated potential outcomes, time indices, as well as any additional covariates. To predict $Y_t(0)$ from $\boldsymbol{X}_t$, denote a prediction model f in a class of models, $\mathcal{F}$, by $f(\boldsymbol{x})$ for $\boldsymbol{x} \in \Re^K$, where K is the number of predictors. A prediction of $Y_t(0)$ in group $G = g$ replaces $\boldsymbol{x} \in \Re^K$ with $\boldsymbol{X}_t \mid G = g$, yielding $f(\boldsymbol{X}_t) \mid G = g$.

If a prediction function were accurate in expectation, then the ATT could be identified as

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

This identification assumption is the basis for single interrupted time series designs (Wagner et al., 2002; Bloom, 2003; Zhang and Penfold, 2013; McDowall, McCleary and Bartos, 2019; Shadish, Cook and Campbell, 2002). However, it is often implausible that a model f can anticipate unexpected shocks (i.e, shocks between periods T and $T-1$ not contained in information up to and including period $T-1$), a longstanding concern for research on the effects of gun policy legislation (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. It assumes that a model's prediction errors are equal in the treated and comparison groups (in expectation) or, said another way, that unexpected shocks shape both groups' outcomes equally, on average.

ASSUMPTION 2 (Equal expected prediction errors).

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

With this additional assumption, we can identify the ATT under an arbitrary model, $f(\boldsymbol{x})$, whereby an observable population-level quantity given by

$$\mathrm{E}_P\left[Y_T - f\left(\boldsymbol{X}_T\right) \mid G = 1\right] - \mathrm{E}_P\left[Y_T - f\left(\boldsymbol{X}_T\right) \mid G = 0\right] \tag{5}$$

is equal to the ATT in Eq. 1. To distinguish (5) under Assumption 2 from the target causal quantity, ATT in Eq. 1, we refer to the former as an *identified* estimand.

Theorem 1 formally establishes that (5) is equal to the ATT under Assumptions 1 and 2.

THEOREM 1 (Causal identification by equal expected prediction errors). *If Assumptions 1 and 2 hold, then* (5) *is equal to the ATT in Eq. 1, i.e.,*

$$E_P\left[Y_T - f\left(\boldsymbol{X}_T\right) \mid G = 1\right] - E_P\left[Y_T - f\left(\boldsymbol{X}_T\right) \mid G = 0\right] = ATT. \tag{6}$$

The proof rearranges (5) via the linearity of expectation and uses Eq. 2 to substitute potential for observed outcomes. The proof concludes by then invoking equal expected prediction errors in Eq. 4 to substitute for the newly formulated right-hand side of (5), which yields a quantity equal to the ATT in Eq. 1.

2.1. *Existing designs as special cases.* Under what circumstances would equal expected prediction errors hold? We offer two kinds of answers to this question. First, we show that some popular non-parametric identification assumptions imply that Assumption 2 holds. Second, we show that some familiar structural causal models imply that Assumption 2 holds. In each case, we specify a prediction function, use the nonparametric or structural identifying assumption of the model to plug in the prediction function's expected values, and then show that prediction errors are equal in the treated and comparison group (in expectation), thus satisfying Assumption 2.

2.1.1. *Difference-in-differences.* Difference-in-differences is a popular method for observational causal inference in several social science fields, including 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, 2023; 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. We show that this assumptions implies our identification assumption 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:

$$\begin{aligned} &\mathrm{E}_P\left[Y_T(0) \mid G = 1\right] - \mathrm{E}_P\left[Y_{T-1}(0) \mid G = 1\right] = \\ &\mathrm{E}_P\left[Y_T(0) \mid G = 0\right] - \mathrm{E}_P\left[Y_{T-1}(0) \mid G = 0\right]. \end{aligned} \tag{7}$$

Assumption 1 and Eq. 7 allow us to solve for $\mathrm{E}_P[Y_T(0) \mid G=1]$ and substitute into the expression for the ATT in Eq. 3 to obtain,

$$ATT = (\mathrm{E}_P[Y_T \mid G=1] - \mathrm{E}_P[Y_{T-1} \mid G=1]) - (\mathrm{E}_P[Y_T \mid G=0] - \mathrm{E}_P[Y_{T-1} \mid G=0]).$$

This is the identified DiD estimand under the non-parametric assumption of parallel trends.

As a corollary to Theorem 1, we show that there exists a prediction model such that, if parallel trends holds, Assumption 2 does also.

COROLLARY 1. *If Assumption 1 and Eq. 7 hold, then Assumption 2 also holds for the prediction model of*

$$f(\boldsymbol{X}_t) = Y_{t-1}. \tag{8}$$

The proof is in Appendix A.1. It is straightforward to extend Corollary 1 to conditional parallel trends by conditioning on the value of a covariate in both the statement of parallel trends in Eq. 7 and the prediction model in Eq. 8.

2.1.2. *Two-way fixed effects models.* In the case of multiple time periods (rather than only pre and post), 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). These model contains a term for the interaction of post-intervention time periods and treated group, the coefficient on which is interpreted as an estimator of the ATT. This approach is justified by the equivalence of the TWFE estimator to the canonical DID estimator in the two-group, two-period case (Angrist and Pischke, 2008; Egami and Yamauchi, 2022; Imai and Kim, 2019; Kropko and Kubinec, 2020; Sobel, 2012; Wooldridge, 2005). That leads the the popular impression that TWFE models are also justified by parallel trends. However, the equivalence does not extend to the more general setting. Imai and Kim (2021) showed 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 a TWFE identification strategy relies on the following structural model. In addition to a set of time periods, suppose a set of units, e.g., states, indexed by $u = 1, \ldots, U$. The TWFE model is

$$\mathrm{E}_p[Y_{u,t}(0)] = \alpha_u + \gamma_t, \tag{9}$$

for all $u = 1, \ldots, U$ and $t = 1, \ldots, T$. As the next corollary shows, this is also a special case of Assumption 2.

COROLLARY 2. *If Assumption 1 and Eq. 9 hold, then Assumption 2 also holds for*

$$\arg\min_{\alpha_u} \sum_{t=1}^{T-1} (Y_{u,t} - \alpha_u)^2. \tag{10}$$

The proof is in Appendix A.2. The prediction model in Eq. 10 is the population-level OLS solution to the unit fixed effects model's objective function fit to data before period t. This solution is equivalent to the mean of a unit's outcomes in all pre-treatment periods.

2.2. *Existing designs that are not special cases.* The designs considered so far use a pre-post contrast to account for time-invariant group differences and a treated-comparison contrast to account for common shocks. Likewise, our proposed framework is a prediction step that uses predictable features of each group's outcome trajectories to project into the future and a correction step that uses the comparison group to correct for unexpected shocks. Interrupted time series is different in that it uses only a pre-post contrast; there are no comparison units with which to perform our correction step. Likewise, synthetic controls use only a treated-comparison contrast, omitting the pre-post contrast. Appendix A.5 contains more detail on the question of synthetic controls.

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., Arkhangelsky et al., 2021), and we imagine this is a fruitful topic for further research.

Finally, we consider two further permutations of the relationships among standard assumptions and our proposed assumption. In Appendix A.3, we show how a standard structural model (interactive fixed effects) could hold but Assumption 2 would not. In Appendix A.4, we show that a standard nonparametric identifying assumption (parallel trends) could be violated but Assumption 2 could still hold.

2.3. *Staggered adoption.* We have thus far assumed that all treated units receive the intervention at the same time. Now we extend to the setting in which treatment is implemented at different times for different treated units (i.e., staggered adoption). We take the perspective of Callaway and Sant'Anna (2021) and others in which we consider each treatment adoption time as its own mini-design. Then we can estimate the treatment effect for each of these treatment timing groups and weight the estimates together in a sensible way.

Instead of a treated and control group, $G = 1$ and $G = 0$, suppose a greater number of groups defined by the respective periods of treatment onset, $G = g > 1$, in which there exists one group that is never treated. By convention, let $g = \infty$ denote the never-treated group and let $G_g = \mathbb{1}\{G = g\}$ be an indicator for membership in treatment timing group g. Define $Y_t(0)$ as the potential outcome at time t if in the never-treated group and let $Y_t(g)$ be the potential outcome at time t if in treatment timing group g. Then we can re-state consistency (Assumption 1) as

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

and our target estimand as the average treatment effect on the treated for each treatment time g,

$$\text{ATT}(g) \coloneqq \text{E}_P [Y_g(g) - Y_g(0) \mid G_g = 1] .$$

This 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,

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

This simplifies Callaway and Sant'Anna (2021)'s approach in the two ways. First, those authors also consider using not-yet-treated comparison units, but we omit this for simplicity. 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 $\text{ATT}(g)$ reduces to the simple case of treated versus comparison units. Callaway and Sant'Anna (2021), in their Section 3, provide several ideas for weighting the resulting collection of estimates together.

3. Selecting models for robustness. Next, we turn to choosing among prediction models. We first define robustness (the complement of sensitivity) and discuss the difference between robustness and plausibility.

3.1. *Design sensitivity.* Design sensitivity was originally developed in the context of matched observational studies in which the crucial assumption is that the matched design is equivalent to a block randomized experiment (Rosenbaum, 2004). If this assumption were exactly met, then standard point estimators would converge in probability to a constant as the size of the study increases indefinitely. However, for any magnitude of violation of the assumption, there is a *range* of point estimates consistent with the data and therefore a limiting interval to which the point estimates converge in probability (Rosenbaum, 2005, 2012). For a given magnitude of violation, a sensitive design has a wider limiting interval than a more robust one. In our framework, we can think of design sensitivity in terms of design choices that shape the width of this limiting interval. Our focus is on the specific choice of prediction model.

The robustness of a design to violations of a point identifying assumption is different from the *plausibility* of an identifying assumption. In the name of assessing plausibility, 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 and rigorous way out of this bind in that pre-period data can inform how robust our inferences would be under increasingly severe violations of the point identifying assumption.

3.2. *Robustness criterion.* We propose a data-driven procedure that chooses a prediction model based on design sensitivity. This builds on Rambachan and Roth (2023) who, following Manski and Pepper (2018), set-identify the ATT by bounding the possible violations of parallel trends. They posited that the violation lies in a set defined by the observed pre-period differential trends. From this set restriction, they obtained sensitivity bounds on the ATT.

We likewise suppose that violations of our identifying assumption lie in a set defined by the (observable) pre-period differential prediction errors. Denote the population-level differential prediction errors in period t under model specification $f \in \mathcal{F}$ by

$$\delta_{f,t} := \mathrm{E}_P\left[Y_t(0) - f\left(\boldsymbol{X}_t\right) \mid G = 1\right] - \mathrm{E}_P\left[Y_t(0) - f\left(\boldsymbol{X}_t\right) \mid G = 0\right]. \tag{11}$$

The point identification of Assumption 2 can be written as $\delta_{f,T} = 0$. For set identification, we would instead suppose $\delta_{f,T} \in \Delta_{f,T}$, where $\Delta_{f,T}$ is a compact set for all $f \in \mathcal{F}$. This implies the following sensitivity bounds on the ATT:

$$\underline{\mathrm{ATT}} = \mathrm{ATT} + \min_{\delta \in \Delta_{f,T}} \delta \tag{12}$$

$$\overline{\mathrm{ATT}} = \mathrm{ATT} + \max_{\delta \in \Delta_{f,T}} \delta\,, \tag{13}$$

where δ is an element in $\Delta_{f,T}$. This leads to our definition of sensitivity: $\overline{\mathrm{ATT}} - \underline{\mathrm{ATT}}$. A smaller difference between these bounds implies less sensitivity (i.e., greater robustness).

To define a relevant set $\Delta_{f,T}$, we follow Rambachan and Roth (2023) and include values of δ up to M times the largest absolute differential prediction error in a set of pre-treatment

validation periods $v \in \mathcal{V} \subseteq \{2, \ldots, T-1\}$. That is,

$$\Delta_T = \left\{ \delta : |\delta| \leq M \max_{v \in \mathcal{V}} |\delta_{f,v}| \right\}. \tag{14}$$

For each validation period, $v \in \mathcal{V}$, $\delta_{f,v}$ is differential prediction error from prediction model f fit to data from $t < v$.

We can imagine alternatives to this set restriction that entail different relationships between pre- and post-periods. For instance, we could remove the absolute value in Eq. 14 to create an asymmetric set restriction. If we think more recent validation periods are more informative, we might use the set restriction of $\Delta_{f,T} = \{\delta : |\delta| \leq M|\delta_{f,V}|\}$ for $V = \max \mathcal{V}$, 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 define $\Delta_{f,T} = \left\{\delta : |\delta| \leq M/|\mathcal{V}| \sum_{v \in \mathcal{V}} |\delta_{f,v}|\right\}$. We proceed with the set restriction in Eq. 14, 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 trends to choose 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. 14, model f is more robust than f' if and only if* $\max_{v \in \mathcal{V}} |\delta_{f,v}| \leq \max_{v \in \mathcal{V}} |\delta_{f',v}|$.

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 $f \in \mathcal{F}$ with the smallest maximum absolute difference in prediction errors. By being able to define robustness in terms of observable pre-period quantities, we can choose among candidate models by maximizing robustness. If we had a different set restriction than Eq. 14, such as the most recent or mean over all validation periods, we could still choose the prediction model that minimizes the width of the implied sensitivity bounds.

To reiterate, our argument does not speak to the plausibility of point identification in Assumption 2 or, equivalently, that $M = 0$ in Eq. 14. This distinction between robustness and plausibility is closely connected to the conditions under which choosing the most robust model can backfire. For example, 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 plausibility and robustness.

— namely, when a model's differential prediction errors in the pre-period provides "misleading" information about its (unobservable) differential prediction error in the post-period. That is, Assumption 2 may hold exactly for some model that nonetheless is less robust (by our criterion) than some other candidate models. Proposition 2 quantifies the consequences of this trade-off.

PROPOSITION 2. *Let f' be a prediction model for which Assumption 2 holds and let f be one for which it does not. However, suppose f' has a greater maximum absolute differential prediction in the pre-period compared to f. By choosing the more robust model f over the "correct" model f', the difference between the ATT and the $\delta_{f,T}$ is*

$$E_P\left[f\left(\boldsymbol{X}_T\right) - f'\left(\boldsymbol{X}_T\right) \mid G = 0\right] - E_P\left[f\left(\boldsymbol{X}_T\right) - f'\left(\boldsymbol{X}_T;\right) \mid G = 1\right]. \tag{15}$$

Proposition 2 lends clarity to a common argument in controlled pre-post and related designs — namely, that a design is more robust if point estimates are stable across competing

models (Brown and Atal, 2019; O'Neill et al., 2016). In terms of our conception of robustness, two prediction models that yield identical point estimates under $M = 0$ can have vastly different degrees of robustness under $M > 0$. However, note that if two different models yield identical point estimates in expectation, then (15) is equal to 0. Therefore, stable point estimates across different models does not make a design more robust, per se, but rather mitigates the possibility that choosing the most robuts model (under our robustness criterion) will backfire. The value of Proposition 2 is that it clarifies the valuable role for the stability of point estimates across models above and beyond the notion of robustness.

Nevertheless, researchers often interpret differential pre-period trends as informative about the plausibility of an identification assumption. Thus, in practice, choosing the most robust model is unlikely to present a trade-off between plausibility and robustness. Insofar as there is a trade-off between plausibility and robustness, our framework enables researchers to choose among equally plausible models, which is important in the gun policy literature in which a wide range of models are plausible. Nevertheless, a general decision framework that incorporates both plausibility and robustness of candidate models is a topic ripe for future research.

4. Model selection, estimation, and inference. Thus far, we have considered population quantities only. For many related approaches, estimation and inference from sample data can proceed by plugging in sample analogues of population quantities. However, in these approaches the prediction model is conceived as fixed before sample data are realized. Our approach, by contrast, incorporates a data-driven model selection procedure, which requires greater care in how to conduct estimation and inference.

While our argument is general, we root what follows in ordinary least squares (OLS) linear regression because of its importance and accessibility to both researchers and practitioners. The framework of OLS linear regression is sufficiently rich to capture the range of actual models researchers employ, especially with respect to the gun policy literature. Nevertheless, it is straightforward to extend our methodology to additional approaches, such as logistic, Poisson, transformed-outcome and isotonic regression (see, e.g., Guo and Basse, 2023).

In the setting of OLS regression, now write the specification of a prediction function by $f(\boldsymbol{x};\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \Re^K$, where K is the number of predictors (i.e., the dimensionality of $\boldsymbol{X}_t$). The prediction function itself under specification f in period t is $f(\boldsymbol{x};\boldsymbol{\beta}_{f,g,t})$. The vector of parameters, $\boldsymbol{\beta}_{f,g,t}$, is the solution to model specification f's objective function based on group g's prior data, $\boldsymbol{Y}_{<t} \,|\, G = g$ and $\boldsymbol{X}_{<t} \,|\, G = g$, which denote group g's collections from periods $t = 1, \ldots, t-1$ of all outcomes and predictors, respectively. A prediction of $Y_t(0)$ in group $G = g$ replaces $\boldsymbol{x} \in \Re^K$ with $\boldsymbol{X}_t \,|\, G = g$, yielding $f(\boldsymbol{X}_t;\boldsymbol{\beta}_{f,g,t}) \,|\, G = g$. For utmost clarity, we should also index the predictors, $\boldsymbol{X}_t$ or $\boldsymbol{X}_{<t}$, by the corresponding model specification, f. For example, $\boldsymbol{X}_t$ will be different for a model with unit fixed effects compared to a model without them. We do not do so for notational simplicity; the corresponding model for any $\boldsymbol{X}_t$ or $\boldsymbol{X}_{<t}$ should be clear from context.

We also assume the following population moment conditions on predictors and outcomes.

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

The first two conditions in Assumption 3 are standard. 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.

Suppose a sample of n units from the population-level distribution, where the sample data for an individual unit, i, in period t are denoted by

$$\boldsymbol{D}_{i,t} = \{Y_{i,t}, \boldsymbol{Y}_{i,<t}, \boldsymbol{X}_{i,t}, \boldsymbol{X}_{i,<t}, G_i\}\,. \tag{16}$$

Then denote the collection of $\boldsymbol{D}_{1,t}, \ldots, \boldsymbol{D}_{n,t}$ for any given t by $\boldsymbol{D}_t$ and the collection of $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_T$ by $\boldsymbol{D}$. Likewise, let $\hat{\boldsymbol{\beta}}_{f,t}$ denote the estimated coefficients in both treated and comparison groups under model f in period t, $\{\hat{\boldsymbol{\beta}}_{f,1,t}, \hat{\boldsymbol{\beta}}_{f,0,t}\}$. Analogously, $\hat{\boldsymbol{\beta}}_f$ denotes the collection of $\{\hat{\boldsymbol{\beta}}_{f,t}\}_{t=1}^T$ and $\hat{\boldsymbol{\beta}}$ denotes $\{\hat{\boldsymbol{\beta}}_f\}_{f\in\mathcal{F}}$.

Going forward, we assume independent and identically distributed (i.i.d.) sampling of data.

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.).

We now denote a point estimator of $\delta_{f,t}$ for any $t = 1, \ldots, T$ under any $f \in \mathcal{F}$ by

$$\begin{aligned}\hat{\delta}\left(\boldsymbol{D}_t, \hat{\boldsymbol{\beta}}_{f,t}\right) = &\left(\frac{1}{n_1}\right)\sum_{i=1}^{n} \mathbb{1}\{G_i = 1\} Y_{i,t}(1) - \left(\frac{1}{n_1}\right)\sum_{i=1}^{n} \mathbb{1}\{G_i = 1\} \boldsymbol{X}_{i,t}\hat{\boldsymbol{\beta}}_{f,1,t} \\ &- \left[\left(\frac{1}{n_0}\right)\sum_{i=1}^{n} \mathbb{1}\{G_i = 0\} Y_{i,t}(0) - \left(\frac{1}{n_0}\right)\sum_{i=1}^{n} \mathbb{1}\{G_i = 0\} \boldsymbol{X}_{i,t}\hat{\boldsymbol{\beta}}_{f,0,t}\right].\end{aligned} \tag{17}$$

The estimator of lower and upper bounds in period T for any $M \geq 0$, where $M = 0$ is equivalent to $\hat{\delta}\left(\boldsymbol{D}_T, \hat{\boldsymbol{\beta}}_{f,T}\right)$, is

$$\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}_f, M\right) = \hat{\delta}\left(\boldsymbol{D}_t, \hat{\boldsymbol{\beta}}_{f,t}\right) \pm M \max_{v\in\mathcal{V}} \hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}_{f,v}\right). \tag{18}$$

A seemingly natural approach to estimation and inference would be as follows. For each prediction model $f \in \mathcal{F}$ and each validation period $v \in \mathcal{V}$, estimate the differential average by $\hat{\delta}\left(\boldsymbol{D}_t, \hat{\boldsymbol{\beta}}_{f,t}\right)$. Then, to minimize estimated sensitivity, one could choose the model with the smallest worst-case absolute difference in prediction errors over the validation periods. Then one could use that model to estimate ATT and its bounds when $M > 0$.

The issue with this approach is that, in principle, the model specification should be fixed before observing data. However, in the procedure above, the optimal model may change depending upon which sample we draw from the population-level distribution. The usual approach to this problem of splitting data into testing and training subsets is not viable in our application. 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 Equation Eq. 18 use data from pre-treatment validation periods $\mathcal{V}$. Nor can we rely on being able to split the data "horizontally": many applications (including the one we consider here) have only a single or a few treated units, so we often cannot afford to split the units into training and test sets.

Ultimately, without access to the population-level distribution, the optimal model is unknown and cannot be fixed before observing sample data. Given this uncertainty over which model is optimal, we propose a Bayesian model averaging (BMA) estimator, which averages estimates under each model based on the plausibility (expressed as a probability measure) that each model is optimal given the sample data. In general, such Bayesian procedures are a well-known approach to conducting estimation and inference without conditioning on a "winning" model and thereby ignoring the fundamental uncertainty over which model is optimal (Piironen and Vehtari, 2017; Madigan and Raftery, 1994; Draper, 1995; Moulton, 1991; Raftery, Madigan and Hoeting, 1997).

We write our BMA estimator in period t as

$$\widehat{\mathrm{E}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M\right)\right] = \sum_{f\in\mathcal{F}} \hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}_f, M\right)\hat{p}_f, \tag{19}$$

where $\hat{p}_f$ is the plausibility (expressed as a probability measure) that model specification $f \in \mathcal{F}$ is the truly optimal model in the population given the sample data. To estimate the plausibility that any model $f \in \mathcal{F}$ is optimal, we extend the quasi-Bayesian procedure of Gelman and Hill (2006), employed by many researchers (e.g., King, Tomz and Wittenberg, 2000; Tomz, Wittenberg and King, 2003), but most notably by Miratrix (2022) in the closely related ITS design. The key idea of this procedure, described more rigorously below, is to generate draws from the "quasi-posterior" of all parameters of all prediction model specifications.

To implement this quasi-Bayesian procedure, we draw parameter values across all models and validation periods from a multivariate Normal distribution centered at the estimated parameters with variance-covariance matrix equal to the parameter estimator's variance-covariance matrix. To compute the joint covariance matrix for parameters across all models for all validation periods, we draw on tools from seemingly unrelated regressions (SUR), pioneered by Zellner (1962, 1963). We denote this estimated variance-covariance matrix by $\widehat{\boldsymbol{\Sigma}}$ and provide its details in Appendix A. Then for each parameter draw from this multivariate Gaussian distribution, $\mathcal{N}(\hat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$, we predict outcomes with each model, calculate absolute differential average prediction errors, and select the best model. Doing this many times, generates a distribution for the best model, where 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. The draws from the multivariate Gaussian are effectively samples from the posterior distribution of the models' coefficients if the prior were flat, hence the "quasi-Bayesian" moniker.

To formally characterize this procedure, denote B draws of parameters from $\mathcal{N}(\hat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$ by

$$\left\{\hat{\boldsymbol{\beta}}^{(b)}\right\}_{b=1}^{B}.$$

Then, for all $f \in \mathcal{F}$, write $\hat{p}_f$ as

$$\hat{p}_f \coloneqq \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{f = \arg\min_{f\in\mathcal{F}} \max_{v\in\mathcal{V}} \left|\delta\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}_{f,v}^{(b)}\right)\right|\right\}, \tag{20}$$

which is the proportion of B draws in which f is the optimal model. In Proposition 1 below, we show that $\hat{p}_f$ of the truly optimal model, $f^\dagger$, converges in probability to 1. In other words, in a sufficiently large sample, the probability is high that the posterior probability of the optimal model, $f^\dagger$, will be close to 1.

LEMMA 1. *Let $f^\dagger \in \mathcal{F}$ denote the truly optimal model in the population. Under Assumptions 1 and 3, as well as iid sampling,*

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

One reason why Proposition 1 is important is because it implies that our BMA estimator in (19) is consistent for the population-level bounds under the optimal model. In other words, as the sample size increases indefinitely, the BMA estimator converges to the probability limit of an estimator that selects the truly optimal model before observing data.

PROPOSITION 3. *Under Assumptions 1 and 3, as well as iid sampling,*

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

In order for the probability limit in Proposition 3 to be equal to the ATT, Assumption 2 of equal expected prediction errors under model $f^\dagger$ would need to be true. Knowing whether this assumption is true under model $f^\dagger$ or any other is impossible. Hence, as Proposition 2 shows, our BMA estimator may be biased for the ATT. We do know, however, that $f^\dagger$ is the most robust model in that it implies the tightest bounds on the ATT for any $M > 0$ under the set restriction in Eq. 14.

For inference, we build on the approach from Antonelli, Papadogeorgou and Dominici (2022), which accounts for uncertainty in which model is optimal given sample data and uncertainty in which sample data are realized. As Antonelli, Papadogeorgou and Dominici (2022) explain, uncertainty in the optimal model depends on which sample data are realized; however, drawing from a posterior of parameters over separate bootstraps of sample data is computationally intractable. Hence, following Antonelli, Papadogeorgou and Dominici (2022) we instead take the sum of our BMA estimator's variance over two sources of uncertainty: (a) repeated draws from the observed posterior, holding the sample data fixed and (b) resamples of data, holding the observed posterior fixed. In a slightly different setting, Antonelli, Papadogeorgou and Dominici (2022) show that this variance estimator tends to be conservative. In our setting, we show via simulations that this approach to inference performs well in terms of accurately approximating the true variance of the estimator and yielding confidence intervals with coverage close to nominal levels.

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

$$\widehat{\text{Var}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M\right)\right] \coloneqq \sum_{f \in \mathcal{F}} \left(\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}_f, M\right) - \widehat{\text{E}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M\right)\right]\right)^2 \hat{p}_f. \tag{21}$$

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

$$\begin{aligned}\widehat{\text{Var}}_{\boldsymbol{D}^{(r)}\,|\,\mathcal{F}}\left[\widehat{\text{E}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}(\boldsymbol{D}^{(r)}, \widehat{\boldsymbol{\beta}}^{(r)}, M)\right]\right] \coloneqq \frac{1}{R}\sum_{r=1}^{R}\Bigg(&\widehat{\text{E}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}(\boldsymbol{D}^{(r)}, \widehat{\boldsymbol{\beta}}^{(r)}, M)\right] \\ &- \frac{1}{R}\sum_{r=1}^{R}\text{E}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}(\boldsymbol{D}^{(r)}, \widehat{\boldsymbol{\beta}}^{(r)}, M)\right]\Bigg)^2.\end{aligned} \tag{22}$$

The sum of (21) and (22) is the variance estimator accounting for both sampling and model uncertainty of the BMA estimator in (19). Confidence intervals can then be constructed by drawing on a Normal approximation. To construct confidence intervals via the percentiles intervals over the empirical distribution accounting for uncertainty in the model and data would require generating the posterior over each resample of the data. Our `R` package enables users to do this, but doing so will be computationally expensive.

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. Many researchers agree on a basic model with two-way fixed effects for units and time (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, quadratic, or cubic time trends. In these models, we omit the time fixed effects to avoid over-parameterizing the models.

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). Therefore, generalized linear models such as Poisson and negative binomial regression are often used instead. However, these non-linear models have their own challenges with interpreting the coefficients on interaction terms (Karaca-Mandic, Norton and Dowd, 2012; Ai and Norton, 2003; Puhani, 2012). We use only linear models, but we 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, random effects (Crifasi et al., 2018) and two-stage models (Rubin and Dezhbakhsh, 2003). However, given the prominence of these three model components and the prominence of unit fixed effects and linear models, we believe the resulting set of candidate models is sufficiently 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 24 candidate models. We fit each prediction model to each individual unit; thus all models effectively include unit fixed effects via the data subsetting approach (Kropko and Kubinec, 2020).

Time trend	Lagged dependent variables	Outcome transformations
None	None	None
t	Y_{t-1}	$\log(Y_t)$
t^2		$Y_t - Y_{t-1}$
t^3		

TABLE 1

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

To select among the 24 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, in which we train the model on 1994-1998 data).

Figure 2 shows the absolute differential average prediction errors for all 24 models, with the maximum for each model highlighted in black. The LDV model on the outcome's original

scale without unit-specific time trends (row 3, column 1) minimizes our sensitivity criterion on the sample data. The baseline mean model (row 1, column 1), which most closely corresponds to the model of choice in Webster, Crifasi and Vernick (2014) and Hasegawa, Webster and Small (2019), is the fourth-best model.

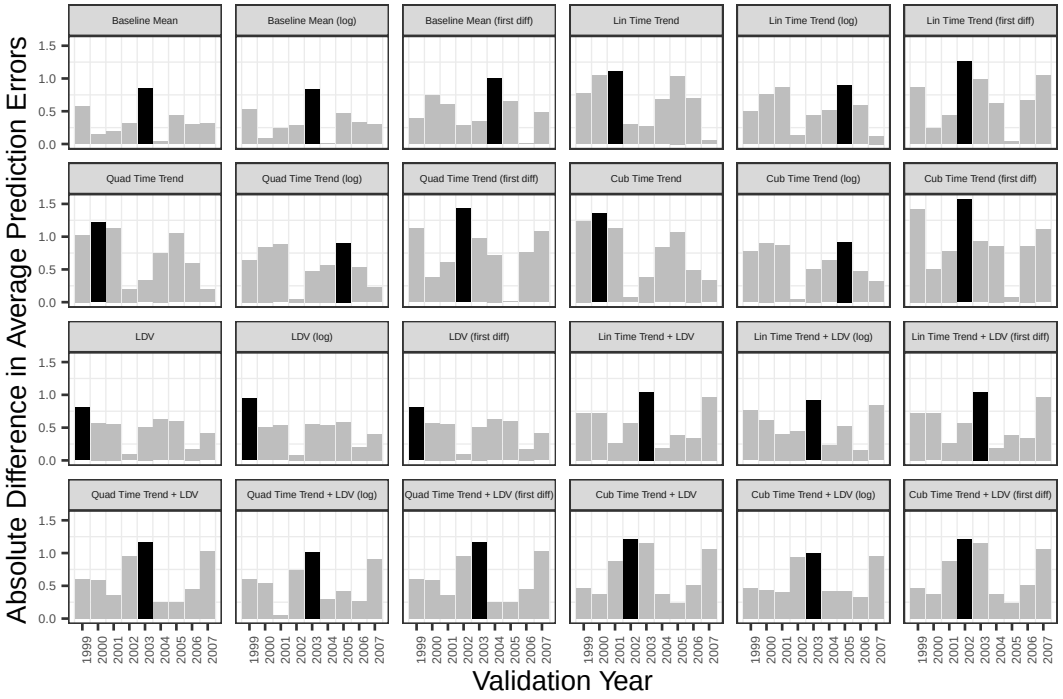

FIG 2. *Absolute difference in average prediction errors for all candidate models. The maximum for each model is highlighted in black.*

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 1, column 4). 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 2). These different loss functions for choosing the optimal model can be justified by an appropriate sensitivity analysis model. Given the sensitivity analysis in Eq. 14, which aligns with the sensitivity analysis proposed in recent research (Rambachan and Roth, 2023), the aforementioned LDV model is optimal. The bootstrap procedure repeatedly samples (with replacement) states from the comparison group, fits the optimal model to predict outcomes in the year 2008, computes the prediction error, and corrects the treated state's outcome prediction using that error to obtain samples from the distribution of estimated ATTs.

This procedure yields a point estimate for the ATT of 1.16 with 95% confidence interval $[0.95, 1.37]$. That is, we conclude that the repeal of Missouri's permit-to-purchase law increased the state's gun homicide rate by 1.16 per 100,000 population. For context, the observed homicide rate in 2007 (just before the repeal) in Missouri was 4.5, so this represents a 26% increase.

Figure 3 shows the predictions and their errors for all validation periods with the optimal prediction model. In it, we see that the maximum absolute differential prediction error occurs

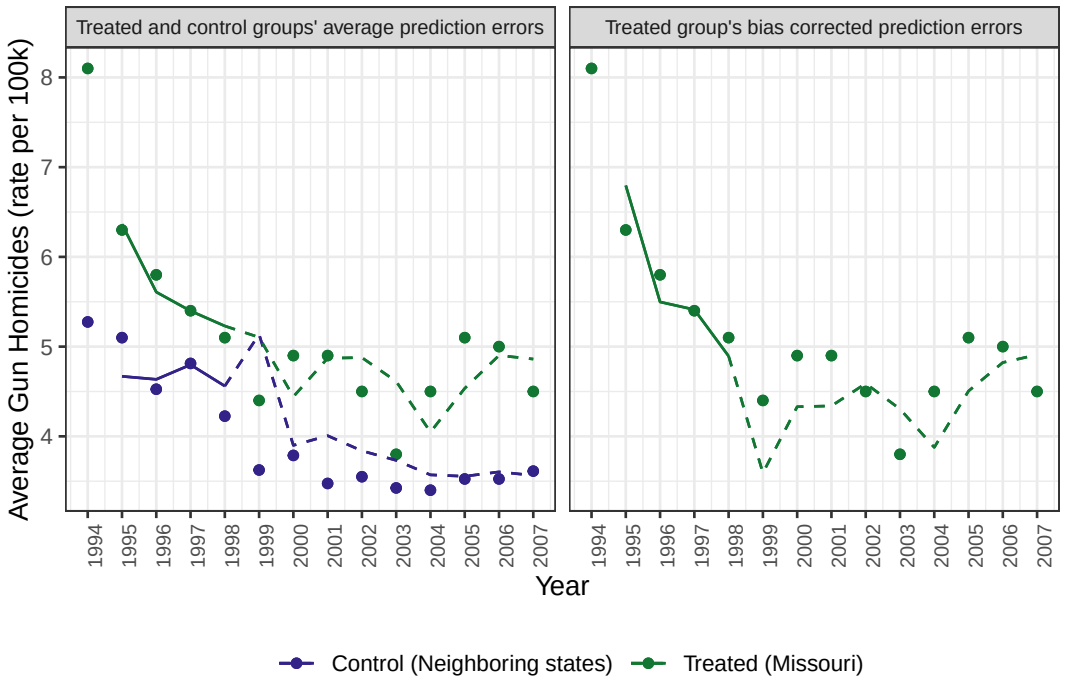

FIG 3. *Average outcomes and the optimal model's average predictions for all pre-treatment validation periods in treated and control states. The points are observed outcomes, the solid lines are model-fitted values in the training periods and the dashed lines are modeled predictions in the validation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

in 1999 and is approximately 0.81. That year's is driven largely by the prediction errors in three control states, Arkansas, Oklahoma and Tennessee. Why was this model's differential prediction error especially big in those states that year? The chosen LDV model had negative coefficients of -0.06, -1.49 and -1.35 on the lagged outcome in those states. Hence, each state's drop in homicide rate in 1998 leads the model to predict *increased* homicide rates in 1999. Instead, the homicide rate dropped further in 1999. The training period in these three states has never shown two consecutive decreases, so the LDV model did not anticipate this pattern. However, in general, a model that predicts that the homicide rates will bump up and down each year is pretty accurate. Models that have both an overall time trend and a lagged dependent variable, by contrast, tend to fail at the moment when the overall trend flattens out, which occurs around 2002/2003 (when all of those models have their worst differential prediction errors).

The LDV model yields a point estimate of 1.16 under Assumption 2. This point estimate is not unusual among the point estimates from all 24 models, which have a standard deviation of 0.16. Hence, one might naively conclude that the stakes of the model selection procedure under each model's point identification assumption are not particularly high.

Figure 4 shows the ATT point estimates and the maximum differential pre-treatment prediction errors of all 24 models. In it, we see that models yielding similar point estimates can differ dramatically in terms of robustness. For instance, the most and least robust models yield similar point estimates of 1.16 and 1.13, respectively.

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 observed prediction error in the comparison group, and *select* the optimal prediction model using a robustness criterion. Our causal identification of

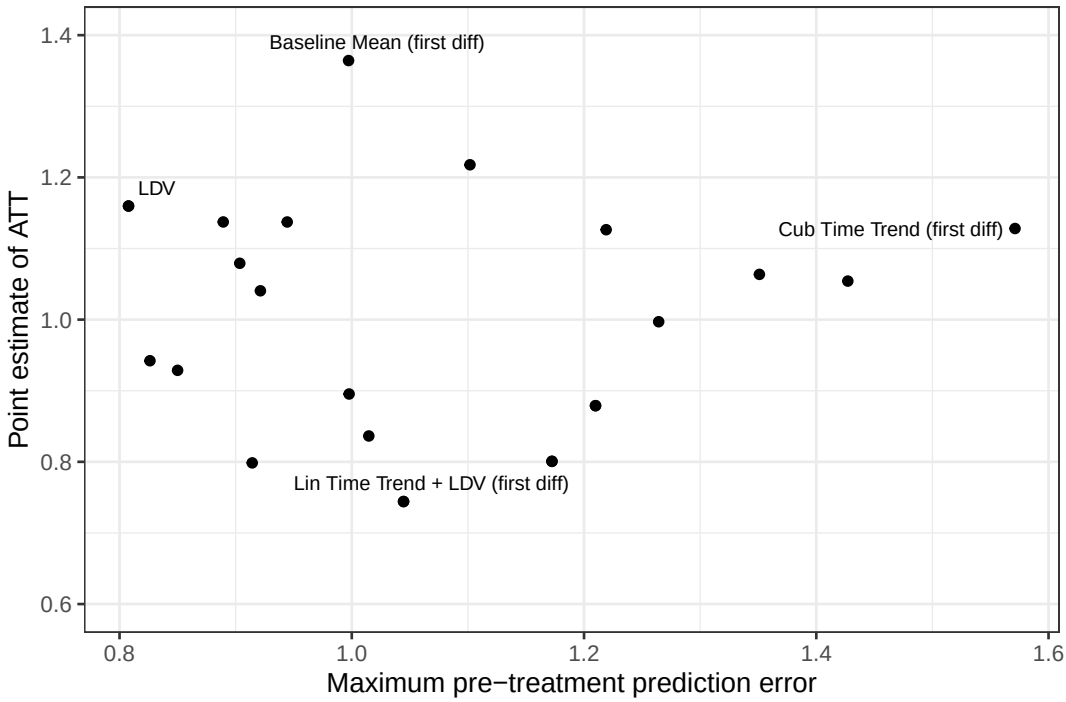

FIG 4. *Point estimates of ATT from each model (y axis) and corresponding maximum absolute differential prediction errors in the pre-period (x axis).*

the ATT based on these predictions assumes equal prediction errors (in expectation) in the treated and comparison groups.

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 about the plausibility of any underlying causal models, we selected the optimal model based on robustness. We found that a lagged dependent variable model minimized our robustness criterion and produced a point estimate of an increase of 1.16 homicides per 100,000 population. This was similar to point estimates under alternative models, which ranged from 0.75 to 1.4.

Considering design sensitivity allows us to compare the robustness of the estimates from the competing models. Our chosen model's sensitivity bounds would include 0 for $M \geq 1.4$. That is, the violation of Assumption 2 would have to be more than 1.4 times greater than the worst violation in the 9 validation years. By contrast, for the least robust model, the value of M that causes the sensitivity bounds to include 0 is much smaller, only 0.7.

We have showed that several popular designs are special cases of our general identification framework.

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, but we believe this limitation is compatible with applied models in common use.

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 make an explicit bias-robustness trade-off (Proposition 2). Rather than claim that we can choose the "correct" model, we choose a model that maximizes our robustness criterion. Even 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. The proposed quasi-Bayesian procedure accounts for dependence among parameters within candidate models, but not *across* candidate models. Moreover, bootstrap methods are known to 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). 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 an optimal 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 plausibility, which involves assumptions about unknowable counterfactual outcomes and therefore introduces the temptation to claim that the model with the most favorable results is the most plausible.

Second, many researchers already interpret robustness in terms of bias. 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. Because our proposed model selection procedure is grounded in a model-free, causal identification framework, 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 inferential procedure would need to be substantially updated to accommodate non-parametric 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, $\mathrm{E}\left[Y_T(0) \mid G=1\right]$ and $\mathrm{E}\left[Y_T(0) \mid G=0\right]$.
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, $f\left(\boldsymbol{x}\right)$.[1]
   b) Next, take expectation (with respect to the identification assumptions) of the prediction models in each group, $\mathrm{E}\left[f\left(\boldsymbol{X}_T\right) \mid G=1\right]$ and $\mathrm{E}\left[f\left(\boldsymbol{X}_T\right) \mid G=0\right]$
   c) Finally, compute the differential prediction error (in expectation),
$$\mathrm{E}\left[Y_T(0) - f\left(\boldsymbol{X}_T\right) \mid G=1\right] - \mathrm{E}\left[Y_T(0) - f\left(\boldsymbol{X}_T\right) \mid G=0\right]$$
3. Show that this is equal to 0, thereby implying Assumption 2 and, consequently, the identified estimand in Eq. Eq. 6.

A.1. Difference-in-Differences. First, use parallel trends in Eq. 7 to write the treated and comparison groups untreated potential outcomes (in expectation) in the post-treatment period as

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

Next, using Assumption 1, take expectation of the prediction model in Eq. 8 in each group

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

Finally, we substitute to compute the differential prediction error (in expectation),

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

which is zero by Eq. 7, so Assumption 2 also holds.

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

A.2. Two-way Fixed Effects. First, the structural model in Eq. 9 yields the following untreated potential outcomes (in expectation) in the post-period:

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

The prediction model in Eq. 10 is simply each unit's average outcome in the pre-period,

$$f\left(\boldsymbol{X}_T\right) = \arg\min_{\alpha_u} \sum_{t=1}^{T-1} \left(Y_{u,t} - \alpha_u\right)^2 = \frac{1}{(T-1)} \sum_{t=1}^{T-1} Y_{u,t}, \tag{23}$$

so substituting $Y_{u,t}(0)$ for the observed outcomes (by Assumption 1) and taking expectation with respect to the structural model in Eq. 9 yields

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

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

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

It is apparent that this is 0, thereby implying Assumption 2.

Thus, the popular TWFE structural model implies our identification condition when the prediction function is that of the canonical DID's or 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 canonical DID's prediction function does not imply equal expected prediction errors, but use of the OLS analogue 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.

A.3. Standard structural model holds but EEPE does not. Suppose untreated potential outcomes are generated by an interactive fixed effects structural model,

$$Y_{u,t}(0) = \alpha_u + \gamma_t + \nu_u F_t + \epsilon_{u,t}, \tag{24}$$

where ν_u is an unobserved, unit-specific "loading" of the unobserved common factor, F_t, and $\mathrm{E}[\epsilon_{u,t}] = 0$ 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

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

Consider the prediction function in (10), which is simply each unit's average outcome prior to t:

$$\alpha_u^* = \arg\min_{\alpha_u} \sum_{s=1}^{t-1} (Y_{u,s} - \alpha_u)^2 = \frac{1}{(t-1)} \sum_{s=1}^{t-1} Y_{u,s}. \tag{25}$$

With this prediction function, the prediction in period T is

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

which, by the consistency assumption, is

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

The IFE model in Eq. 24 implies that the expectations of the predictions in the treated and control groups are

$$\begin{aligned}
\mathrm{E}[f(\boldsymbol{X}_T) \mid G_u = 1] &= \frac{1}{(T-1)} \left[\sum_{s=1}^{T-1} (\mathrm{E}[\alpha_u \mid G_u = 1] + \gamma_s + \mathrm{E}[\nu_u F_s \mid G_u = 1]) \right] \\
&= \frac{1}{(T-1)} \left[\sum_{s=1}^{T-1} \mathrm{E}[\alpha_u \mid G_u = 1] + \sum_{s=1}^{T-1} \gamma_s + \sum_{s=1}^{T-1} \mathrm{E}[\nu_u F_s \mid G_u = 1] \right] \\
&= \mathrm{E}[\alpha_u \mid G_u = 1] + \frac{1}{(T-1)} \sum_{s=1}^{T-1} \gamma_s + \frac{1}{(T-1)} \sum_{s=1}^{T-1} \mathrm{E}[\nu_u F_s \mid G_u = 1] \\
\mathrm{E}[f(\boldsymbol{X}_T) \mid G_u = 0] &= \mathrm{E}[\alpha_u \mid G_u = 0] + \frac{1}{(T-1)} \sum_{s=1}^{T-1} \gamma_s + \frac{1}{(T-1)} \sum_{s=1}^{T-1} \mathrm{E}[\nu_u F_s \mid G_u = 0]
\end{aligned}$$

and the expected prediction errors in each group are

$$\begin{aligned}
\mathrm{E}\left[Y_{u,T}(0) \mid G_u = 1\right] - \mathrm{E}[f(\boldsymbol{X}_T) \mid G_u = 1] &= \mathrm{E}[\alpha_u \mid G_u = 1] + \gamma_T + \mathrm{E}[\nu_u F_T \mid G_u = 1] \\
&\quad - \left(\mathrm{E}[\alpha_u \mid G_u = 1] + \frac{1}{(T-1)} \sum_{s=1}^{T-1} \gamma_s + \frac{1}{(T-1)} \sum_{s=1}^{T-1} \mathrm{E}[\nu_u F_s \mid G_u = 1] \right) \\
&= \gamma_T - \sum_{s=1}^{T-1} \gamma_s + \mathrm{E}[\nu_u F_T \mid G_u = 1] - \sum_{s=1}^{T-1} \mathrm{E}[\nu_u F_s \mid G_u = 1] \\
\mathrm{E}\left[Y_{u,T}(0) \mid G_u = 0\right] - \mathrm{E}[f(\boldsymbol{X}_T) \mid G_u = 0] &= \mathrm{E}[\alpha_u \mid G_u = 0] + \gamma_T + \mathrm{E}[\nu_u F_T \mid G_u = 0]
\end{aligned}$$

$$- \left(\mathrm{E}\left[\alpha_u \mid G_u = 0\right] + \frac{1}{(T-1)} \sum_{s=1}^{T-1} \gamma_s + \frac{1}{(T-1)} \sum_{s=1}^{T-1} \mathrm{E}\left[\nu_u F_s \mid G_u = 0\right] \right)$$

$$= \gamma_T - \sum_{s=1}^{T-1} \gamma_s + \mathrm{E}\left[\nu_u F_T \mid G_u = 0\right] - \sum_{s=1}^{T-1} \mathrm{E}\left[\nu_u F_s \mid G_u = 0\right].$$

Taking the difference in expected prediction errors yields

$$\mathrm{E}\left[\nu_u F_T \mid G_u = 1\right] - \mathrm{E}\left[\nu_u F_T \mid G_u = 0\right] - \frac{1}{T-1} \left(\sum_{s=1}^{T-1} \mathrm{E}\left[\nu_u F_s \mid G_u = 1\right] - \mathrm{E}\left[\nu_u F_s \mid G_u = 0\right] \right),$$

which is not necessarily equal to 0.

This same issue holds for unit- or group-specific linear time trend model, which is given by

$$Y_{u,t}(0) = \xi_u t + \gamma_t + \epsilon_{u,t} \tag{26}$$

where ξ_u is the linear time slope of the u^{th} unit and $\mathrm{E}[\epsilon_{u,t}] = 0$ for all $u = 1, \ldots, U$ and $t = 1, \ldots, T$.

Now suppose we use the same prediction model, implying, along with the consistency assumption, that the expected predictions in period T for treated and control groups are

$$\begin{aligned}
\mathrm{E}\left[f(\boldsymbol{X}_T) \mid G_u = 1\right] &= \frac{1}{(T-1)} \sum_{s=1}^{T-1} \mathrm{E}\left[Y_{u,s}(0)\right] \\
&= \frac{1}{(T-1)} \sum_{s=1}^{T-1} \mathrm{E}\left[\xi_u s \mid G_u = 1\right] + \frac{1}{(T-1)} \sum_{s=1}^{T-1} \gamma_s \\
\mathrm{E}\left[f(\boldsymbol{X}_T) \mid G_u = 0\right] &= \frac{1}{(T-1)} \sum_{s=1}^{T-1} \mathrm{E}\left[Y_{u,s}(0)\right] \\
&= \frac{1}{(T-1)} \sum_{s=1}^{T-1} \mathrm{E}\left[\xi_u s \mid G_u = 0\right] + \frac{1}{(T-1)} \sum_{s=1}^{T-1} \gamma_s.
\end{aligned}$$

Consequently, the expected prediction errors are

$$\mathrm{E}\left[Y_{u,T}(0) \mid G_u = 1\right] - \mathrm{E}\left[f(\boldsymbol{X}_T) \mid G_u = 1\right] = \mathrm{E}\left[\xi_u T \mid G_u = 1\right] - \frac{1}{(T-1)} \sum_{s=1}^{T-1} \mathrm{E}\left[\xi_u s \mid G_u = 1\right] + \gamma_T - \frac{1}{(T-1)} \sum_{s=1}^{T-1} \gamma_s$$

$$\mathrm{E}\left[Y_{u,T}(0) \mid G_u = 0\right] - \mathrm{E}\left[f(\boldsymbol{X}_T) \mid G_u = 0\right] = \mathrm{E}\left[\xi_u T \mid G_u = 0\right] - \frac{1}{(T-1)} \sum_{s=1}^{T-1} \mathrm{E}\left[\xi_u s \mid G_u = 0\right] + \gamma_T - \frac{1}{(T-1)} \sum_{s=1}^{T-1} \gamma_s.$$

Taking the difference in expected prediction errors yields

$$\mathrm{E}\left[\xi_u T \mid G_u = 1\right] - \mathrm{E}\left[\xi_u T \mid G_u = 0\right] - \left(\frac{1}{(T-1)} \sum_{s=1}^{T-1} \mathrm{E}\left[\xi_u s \mid G_u = 1\right] - \mathrm{E}\left[\xi_u s \mid G_u = 0\right] \right),$$

which is also not necessarily equal to 0.

A.4. Standard non-parametric assumption fails but EEPE holds. A kind of converse to the question above is, as a reviewer wondered, "whether it is possible for nonparametric identification to not hold, but the proposed assumption to hold." Indeed, we can construct a

case where some standard non-parametric assumption is violated, but equal prediction errors (in expectation) nonetheless holds for some choice of prediction model.

Suppose that the true underlying causal structural model is a simple linear model with differential slopes in the treated and comparison groups. This structural model is the one that underlies the comparative interrupted time series (CITS) model, which, when specified using the model of Bloom and Riccio (2005), is equivalent to a flexible difference-in-differences specification that uses time fixed effects and group-specific linear trends (Fry and Hatfield, 2021). Consider the non-parametric identifying assumption that underlies difference-in-differences, that of parallel trends. When the true underlying structural model is CITS, the parallel trends assumption will not hold because of differential trends in the two groups. However, the assumption of equal expected prediction errors can still hold for a prediction function that incorporates differential trends in the two groups into its predictions.

A.5. Proof that synthetic controls are not a special case. Suppose that we are studying a single treated unit (denote it $i = 1$ without loss of generality). The synthetic control weights, w_i, solve a regularized minimization of the mean squared difference between the treated unit's outcomes and the weighted control outcomes at each pre-period time,

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

(This is slightly simplified because it omits the penalty term.) The synthetic control estimator, as originally proposed by Abadie (2005), is simply

$$Y_{1,T} - \frac{1}{N-1}\sum_{i=2}^{N} Y^*_{i,t},$$

where $Y^*_{i,t} = w_i Y_{i,t}$ for $i = 2, \ldots, N$ 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 synthetic control estimator. Thus, whenever the synthetic control assumption is met, Assumption 2 will be also (since the estimators are identical).

First, suppose that the prediction function is $f_{g,t}(\boldsymbol{X}_{i,t}) = 0$ and we had weighted the comparison units by w_i as a "pre-processing" step. Then the estimator implied by Eq. 6 would be

$$ATT = \mathrm{E}\left[Y_{1,T} - 0\right] - \mathrm{E}\left[Y^*_{i,T} - 0 \mid G_i = 0\right] = Y_{1,T} - \frac{1}{N-1}\sum_{i=2}^{N} Y^*_{i,t}\,,$$

which is the synthetic control estimator.

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,

$$\begin{cases} \frac{1}{N-1}\sum_{i=2}^{N-1} Y^*_{i,T} & \text{if } G_i = 1 \\ Y_{i,T} & \text{if } G_i = 0 \end{cases}$$

Then the estimator implied by Eq. 6 would be

$$ATT = \mathrm{E}\left[Y_{1,T} - \frac{1}{N-1}\sum_{i=2}^{N-1} Y_{i,T}^{*}\right] - \mathrm{E}\left[Y_{i,T} - Y_{i,T} \mid G_i = 0\right]$$
$$= Y_{1,T} - \frac{1}{N-1}\sum_{i=2}^{N-1} Y_{i,T}^{*}\,,$$

which is again the synthetic control estimator.

Both prediction functions are trivial because they do no actual prediction: one is a scalar (0) and the other is based on current outcomes, not past. 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.

APPENDIX B: LIMITING BEHAVIOR OF THE PROCEDURE FOR SELECTING AN OPTIMAL PREDICTION MODEL

To account for variances and covariances of estimated coefficients across models and periods, let $\widehat{\boldsymbol{\beta}}$ be the collection of all estimated coefficients across all $|\mathcal{F} \times \mathcal{V}|$ models and validation periods. The estimated joint variance-covariance matrix for all coefficients across all models and validation periods is

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_{(f_1,v_1),(f_1,v_1)} & \cdots & \widehat{\boldsymbol{\Sigma}}_{(f_1,v_1),(f_{|\mathcal{F}|},v_{|\mathcal{V}|})} \\ \vdots & \ddots & \vdots \\ \widehat{\boldsymbol{\Sigma}}_{(f_{|\mathcal{F}|},v_{|\mathcal{V}|}),(f_1,v_1)} & \cdots & \widehat{\boldsymbol{\Sigma}}_{(f_{|\mathcal{F}|},v_{|\mathcal{V}|}),(f_{|\mathcal{F}|},v_{|\mathcal{V}|})} \end{bmatrix} \tag{27}$$

where

$$\widehat{\boldsymbol{\Sigma}}_{(f,v),(f',v')} = \left(\boldsymbol{X}_{f,v}^{\top}\boldsymbol{X}_{f,v}\right)^{-1}\boldsymbol{X}_{f,v}^{\top}\boldsymbol{E}_{f,v}\boldsymbol{E}_{f',v'}\boldsymbol{X}_{f',v'}\left(\boldsymbol{X}_{f',v'}^{\top}\boldsymbol{X}_{f',v'}\right)^{-1} \tag{28}$$

and $\boldsymbol{E}_{f,v}$ 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 (28) as

$$\widehat{\boldsymbol{\Sigma}}_{(f,v),(f',v')} = (1/n)\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{f,v,i}^{\top}\boldsymbol{X}_{f,v,i}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}e_{f,v,i}e_{f',v',i}\boldsymbol{X}_{f,v,i}^{\top}\boldsymbol{X}_{f',v',i}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{f',v',i}^{\top}\boldsymbol{X}_{f',v',i}\right)^{-1} \tag{29}$$

B.1. Proof of Lemma 1.

PROOF. First note that the WLLN implies that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ and $\hat{\boldsymbol{\Sigma}} \xrightarrow{p} 0$, which, by the 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 $\varepsilon > 0$,

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

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 as

$$\frac{1}{n_g}\sum_{i=1}^{n}\mathbb{1}\left\{G_i=g\right\}\left[\boldsymbol{X}_{i,t}\left(\hat{\boldsymbol{\beta}}^{*}_{f,v}-\boldsymbol{\beta}_{f,v}\right)\right]^2. \tag{30}$$

The Cauchy-Schwarz inequality implies that

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

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 $\mathrm{E}_P\left[\boldsymbol{X}_v^2\right]$, where the regularity condition that $\mathrm{E}_P\left[\|\boldsymbol{X}_v\|^2\right]<\infty$ implies that $\mathrm{E}_P\left[\boldsymbol{X}_v^2\right]<\infty$. Consequently, since $\|(\hat{\boldsymbol{\beta}}^{*}_{f,v}-\boldsymbol{\beta}_{f,v})\|^2 \xrightarrow{p} 0$, the CMT implies that

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

Since the upper-bound of (30) converges in probability to 0, so, too, must (30) itself.

The CMT then implies that

$$\hat{\delta}\left(\boldsymbol{D}_v,\hat{\boldsymbol{\beta}}^{*}_{f,v}\right) \xrightarrow{p^*} \mathrm{E}_P\left[\delta\left(\boldsymbol{D}_v,\boldsymbol{\beta}_{f,v}\right)\right],$$

i.e., for all $\varepsilon>0$,

$$\overset{*}{\Pr}\left(\left|\hat{\delta}\left(\boldsymbol{D}_v,\hat{\boldsymbol{\beta}}^{*}_{f,v}\right)-\mathrm{E}_P\left[\delta\left(\boldsymbol{D}_v,\boldsymbol{\beta}_{f,v}\right)\right]\right|\leq\varepsilon\right) \xrightarrow{p} 1, \tag{31}$$

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

For all $f\in\mathcal{F}$ and $v\in\mathcal{V}$, denote $\mathrm{E}_P\left[\delta\left(\boldsymbol{D}_v,\boldsymbol{\beta}_{f,v}\right)\right]$ by $\delta_{(f,v)}$. Then let $\varepsilon>0$ be such that, for all $f\in\mathcal{F}$,

$$\delta_{(f,\overline{v}(f))}-\varepsilon>\delta_{(f,\overline{v}(f))}+\varepsilon \tag{32}$$

for all $v\in\{\mathcal{V}\setminus\overline{v}(f)\}$ and

$$\delta_{(f^\dagger,\overline{v}(f^\dagger))}-\varepsilon<\delta_{(f,\overline{v}(f))}+\varepsilon \tag{33}$$

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

With $\varepsilon>0$ satisfying (32) and (33), it follows that the event

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

implies the event that

$$f^\dagger=\arg\min_{f\in\mathcal{F}}\max_{v\in\mathcal{V}}\hat{\delta}\left(\boldsymbol{D}_v,\hat{\boldsymbol{\beta}}^{*}_{f,v}\right). \tag{34}$$

Hence, (31) implies that

$$\overset{*}{\Pr}\left(f^\dagger=\arg\min_{f\in\mathcal{F}}\max_{v\in\mathcal{V}}\hat{\delta}\left(\boldsymbol{D}_v,\hat{\boldsymbol{\beta}}^{*}_{f,v}\right)\right) \xrightarrow{p} 1, \tag{35}$$

thereby completing the proof.

□

B.2. 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 $(f,t) \in \mathcal{V} \times \mathcal{T}$, the CMT implies that

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

thereby completing the proof. □

APPENDIX C: MODEL IMPLEMENTATION IN THE APPLIED ANALYSIS

Baseline Mean	$Y_{it} \sim \beta_0$
Baseline Mean (log)	$\log(Y_{it}) \sim \beta_0$
Baseline Mean (first diff)	$Y_{it} - Y_{i,t-1} \sim \beta_0$
Lin Time Trend	$Y_{it} \sim \beta_0 + \beta_1 t$
Lin Time Trend (log)	$\log(Y_{it}) \sim \beta_0 + \beta_1 t$
Lin Time Trend (first diff)	$Y_{it} - Y_{i,t-1} \sim \beta_0 + \beta_1 t$
Quad Time Trend	$Y_{it} \sim \beta_0 + \beta_1 t^2$
Quad Time Trend (log)	$\log(Y_{it}) \sim \beta_0 + \beta_1 t^2$
Quad Time Trend (first diff)	$Y_{it} - Y_{i,t-1} \sim \beta_0 + \beta_1 t^2$
Cub Time Trend	$Y_{it} \sim \beta_0 + \beta_1 t^3$
Cub Time Trend (log)	$\log(Y_{it}) \sim \beta_0 + \beta_1 t^3$
Cub Time Trend (first diff)	$Y_{it} - Y_{i,t-1} \sim \beta_0 + \beta_1 t^3$
LDV	$Y_{it} \sim \beta_0 + \beta_2 Y_{i,t-1}$
LDV (log)	$\log(Y_{it}) \sim \beta_0 + \beta_2 Y_{i,t-1}$
LDV (first diff)	$Y_{it} - Y_{i,t-1} \sim \beta_0 + \beta_2 Y_{i,t-1}$
Lin Time Trend + LDV	$Y_{it} \sim \beta_0 + \beta_1 t + \beta_2 Y_{i,t-1}$
Lin Time Trend + LDV (log)	$\log(Y_{it}) \sim \beta_0 + \beta_1 t + \beta_2 Y_{i,t-1}$
Lin Time Trend + LDV (first diff)	$Y_{it} - Y_{i,t-1} \sim \beta_0 + \beta_1 t + \beta_2 Y_{i,t-1}$
Quad Time Trend + LDV	$Y_{it} \sim \beta_0 + \beta_1 t^2 + \beta_2 Y_{i,t-1}$
Quad Time Trend + LDV (log)	$\log(Y_{it}) \sim \beta_0 + \beta_1 t^2 + \beta_2 Y_{i,t-1}$
Quad Time Trend + LDV (first diff)	$Y_{it} - Y_{i,t-1} \sim \beta_0 + \beta_1 t^2 + \beta_2 Y_{i,t-1}$
Cub Time Trend + LDV	$Y_{it} \sim \beta_0 + \beta_1 t^3 + \beta_2 Y_{i,t-1}$
Cub Time Trend + LDV (log)	$\log(Y_{it}) \sim \beta_0 + \beta_1 t^3 + \beta_2 Y_{i,t-1}$
Cub Time Trend + LDV (first diff)	$Y_{it} - Y_{i,t-1} \sim \beta_0 + \beta_1 t^3 + \beta_2 Y_{i,t-1}$

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. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies* **72** 1–19.

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.

AI, C. and NORTON, E. C. (2003). Interaction Terms in Logit and Probit Models. *Economics Letters* **80** 123-129.

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.

ARKHANGELSKY, D., ATHEY, S., HIRSHBERG, D. A., IMBENS, G. W. and WAGER, S. (2021). Synthetic Difference In Differences. *American Economic Review* **111** 4088–4118.

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.

BLOOM, H. S. and RICCIO, J. A. (2005). 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** 19–51.

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.

KARACA-MANDIC, P., NORTON, E. C. and DOWD, B. (2012). Interaction Terms in Nonlinear Models. *Health Services Research* **47** 255-274.

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. (2023). A Practical Guide to Counterfactual Estimators for Causal Inference with TimeSeries CrossSectional Data. *American Journal of Political Science*.

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.

PUHANI, P. A. (2012). The Treatment Effect, the Cross Difference, and the Interaction Term in Nonlinear "Difference-in-Differences" Models. *Economics Letters* **115** 85–87.

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.