The Annals of Applied Statistics
2025, Vol. 19, No. 3, 1826–1846
https://doi.org/10.1214/25-AOAS2011

AVERAGED PREDICTION MODELS (APM): IDENTIFYING CAUSAL EFFECTS IN CONTROLLED PRE-POST SETTINGS WITH APPLICATION TO GUN POLICY

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]*Statistics and Data Science Department, NORC at the University of Chicago,* [b]*hatfield-laura@norc.org*

To investigate causal impacts, many researchers use controlled pre-post designs that compare over-time differences between a population exposed to a policy change and an unexposed comparison group. However, researchers using these designs often disagree about the "correct" specification of the causal model, perhaps most notably in analyses to identify the effects of gun policies on crime. To help settle these model specification debates, we propose a general identification framework that unifies a variety of models researchers use in practice. In this framework, which nests "brand name" designs like difference-in-differences as special cases, we use models to predict untreated outcomes and then correct the treated group's predictions using the comparison group's observable prediction errors. Our point identifying assumption is that treated and comparison groups would have equal prediction errors (in expectation) under no treatment. To choose among candidate models, we propose a data-driven procedure based on models' robustness to violations of this point identifying assumption. Our selection procedure averages over candidate models, weighting by each model's posterior probability of being the most robust, given its differential average prediction errors in the pre-period. This approach offers a way out of debates over the "correct" model by choosing on robustness instead and has the desirable property of being feasible in the "locked box" of preintervention data only. We apply our methodology to the gun policy debate, focusing specifically on Missouri's 2007 repeal of its permit-to-purchase law, and provide an `R` package (`apm`) for implementation.

1. Introduction. The causal effects of public policies are the subject of intense debate. For instance, the question of whether guns make society more or less safe permeates the gun control debate, with opposite sides claiming that policies expanding firearms access either reduce or increase crime. To bring evidence to bear on this debate, we can contrast the change in a relevant outcome (such as gun homicides) after a policy intervention to the contemporaneous change in outcomes among an unexposed comparison population. These "controlled pre-post" designs identify policy effects if their assumptions hold. For example, difference-in-differences (DID) depends on parallel trends, sequential DID depends on parallel trends-in-trends and comparative interrupted time series (CITS) depends on assumptions about group differences in slopes and intercepts of a linear model.

To choose among 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)). Because reasonable people may disagree about plausibility and because it is impossible to prove *any* causal assumption, researchers tend to use methods that

Received September 2023; revised December 2024.
Key words and phrases. Difference-in-differences, predictive models, robustness, model selection, design sensitivity, Bayesian inference.

are popular in their disciplines. For instance, CITS is popular in education policy, while DID is popular in health policy (Fry and Hatfield (2021)).

Disagreement over causal assumptions and attendant modeling choices is not merely academic; it can impede progress on the policy front. On the firearm policy question, 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), page 151. More recent syntheses have reached similar conclusions (Morral et al. (2018), Smart et al. (2020)).

In this paper we consider the impact of a particular firearm policy change: Missouri's 2007 repeal of its permit-to-purchase law; see Figure 1 below. Previous authors have analyzed this policy change using models with different causal assumptions: Webster, Crifasi and Vernick (2014) fit a Poisson regression model with unit and time fixed effects, while Hasegawa, Webster and Small (2019) used a nonparametric DID estimator. How, in general, can we reconcile competing models, assumptions, and possible conclusions?

We provide a general identification framework that unifies controlled pre-post designs by characterizing them as a combination of a *prediction* step and a *correction* step. First, for an arbitrary model in a class of candidate models, we use observations in the preintervention period to train a model that *predicts* untreated outcomes in the postintervention period. Second, we use the comparison group's prediction errors in the postintervention period to *correct* the treated group's predictions; this step accounts for time-varying shocks that affect both groups. Our point identifying assumption is that prediction errors would be equal (in expectation) in treated and comparison groups, absent the policy change. We show that we can reproduce many familiar "brand name" designs by careful choice of prediction model. For instance, we get DID when we use the preintervention group mean as a (simple) prediction model.

To choose among a set of candidate models—unified under a single identification framework—we move away from the question of model "correctness" and focus on another

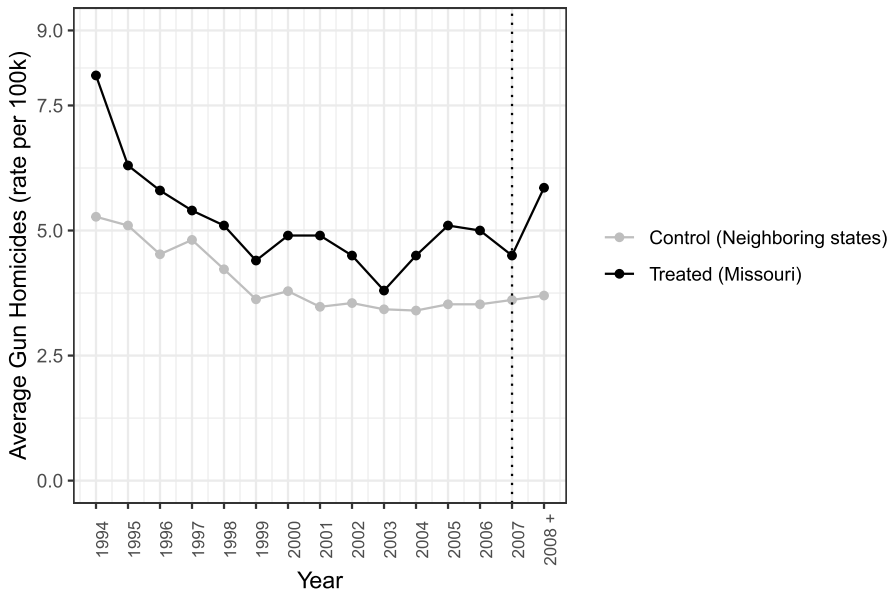

FIG. 1. *Average gun homicides* (*rate per* 100,000) *before and after the* 2007 *permit-to-purchase repeal in Missouri* (*treated state*) *and eight neighboring states without such a change* (*Arkansas*, *Illinois*, *Iowa*, *Kansas*, *Kentucky*, *Nebraska*, *Oklahoma and Tennessee*).

criterion: robustness. To *select* the most robust model given differential average prediction errors in the preperiod, we use Bayesian model averaging (BMA) in which we weight each model by its posterior probability of being the most robust, hence the name *averaged prediction models* (or *APM*). Using the corrected predictions from our averaged model, we estimate our causal target quantity, the average effect of treatment on the treated (the ATT). Since this model selection procedure based on robustness uses only preperiod observations, it has the benefit of inhibiting "fishing," whereby researchers select the model that yields the most desirable conclusion about the ATT in the postperiod.

Our conception of robustness builds on Manski and Pepper (2018) and Rambachan and Roth (2023) who formalize an idea implicit in preperiod parallel trends tests (Granger (1969), Angrist and Pischke (2008), Roth (2022), Egami and Yamauchi (2023)): Departures from the assumed causal model in the preperiod inform us about violations in the postperiod. Since the true relationship between untreated outcomes in the two periods is unknown, these sensitivity analyses take observed departures in the preperiod and assume a relationship to departures in the postperiod. Robustness is a (lack of) change in the estimand under departures from the point identifying assumption.

Other authors have also developed methods for causal inference from longitudinal data and applied them to study gun/policing policies and violence/crime outcomes. With a similar focus on prediction models, Antonelli and Beck (2023) use Bayesian spatiotemporal models to produce posterior predictive distributions for unit-specific treatment effects in a staggered adoption setting. Ben-Michael et al. (2023) use multitask Gaussian process models to draw causal inferences from panel data with one treated unit and count outcomes, contributing to the literature on synthetic controls. We build on these existing approaches via a Bayesian model selection procedure that is guided by an anticipated sensitivity analysis. Our methodology, therefore, joins recent research that applies ideas from design sensitivity (Rosenbaum (2004); see also Heller, Rosenbaum and Small (2009), Hsu, Small and Rosenbaum (2013), Small et al. (2013)) to settings other than those of matched observational studies (Huang, Soriano and Pimentel (2024)).

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

2. General identification strategy. Suppose a population-level data generating process with two groups, a treated group ($G = 1$) and comparison group ($G = 0$) as well $t = 1, \ldots, T$ periods of which T is the only posttreatment 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 $Z_t := G\mathbb{1}\{t = T\}$, where $\mathbb{1}\{\cdot\}$ is the indicator function that equals 1 if its argument is true and 0 if not. For the treated group, $Z_t = 0$ for all $t < T$ and $Z_T = 1$. For the comparison group, $Z_t = 0$ for all periods.

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

$$\text{ATT} := \mathrm{E}_{\mathcal{P}}\big[Y_T(1) - Y_T(0) | G = 1\big], \tag{1}$$

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

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

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

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

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

With Assumption 1 we can express the ATT as

$$\text{ATT} = \text{E}_{\mathcal{P}}\big[Y_T - Y_T(0) | G = 1\big], \tag{3}$$

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

Suppose we predict the untreated potential outcome in period t, $Y_t(0)$, via $f(\boldsymbol{X}_t)$, where f is a model belonging to class of candidate models $\mathcal{F}$ and $\boldsymbol{X}_t$ is the collection of predictors for untreated potential outcomes in period t.[1] The predictors of untreated potential outcomes in period t are quantities whose values are determined before (or are independent of) possible treatment onset between periods $t - 1$ and t. When we have only one postperiod, T, "before T" and "preperiod" are the same. For extensions to multiple posttreatment periods, we also limit a prediction model's inputs to quantities whose values are determined before the start of (or are independent of) treatment.

If the prediction function were perfect, we could identify the ATT without a comparison group. That is, the ATT could be identified as

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

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

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

ASSUMPTION 2 (Equal expected prediction errors).

$$\text{E}_{\mathcal{P}}\big[Y_T(0) - f(\boldsymbol{X}_T) | G = 1\big] = \text{E}_{\mathcal{P}}\big[Y_T(0) - f(\boldsymbol{X}_T) | G = 0\big]. \tag{4}$$

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

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

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

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

The proof, given in Section 1 of the Supplementary Material (Leavitt and Hatfield (2025)), is straightforward, by linearity of expectation and substitution of observed outcomes using Assumptions 1 and 2.

2.1. *Existing designs as special cases.* Under what circumstances would equal expected prediction errors hold? It turns out that several popular nonparametric identification assumptions and structural causal models imply Assumption 2. That is, we show that when these assumptions hold, Assumption 2 will also hold, for particular choices of prediction models. We consider two such situations below and detail several more in the Supplementary Material (Leavitt and Hatfield (2025)).

2.1.1. *Nonparametric identifying assumptions.* Identification for DID—a popular method for observational causal inference in a range of fields—may be shown using either nonparametric assumptions or structural models. We follow the spirit of Angrist and Pischke (2010, p. 14) who regard DID as a "design-based" method. Hence, we use a nonparametric identification assumption to show that it implies our identification assumption given a careful choice of prediction function.

We consider a simple setting in which there are two groups (treated and comparison) and two periods (preperiod $T-1$ and postperiod T). DID's crucial counterfactual assumption is that untreated potential outcomes would have evolved in parallel in the two groups,

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

In the Supplementary Material (Leavitt and Hatfield (2025), Section 2.1), we show that if parallel trends hold, Assumption 2 does also for prediction model $f(\boldsymbol{X}_T) = Y_{T-1}$.

Of course, this is only the simplest example of a DID strategy. We can extend to more complex DID settings with, for example, conditioning on covariates (in both the assumption of parallel trends and the prediction model). In addition, as shown in the Supplementary Material (Leavitt and Hatfield (2025)), we can use alternative assumptions such as those of sequential DID.

2.1.2. *Structural models.* Suppose that we have multiple outcome measurement occasions in the pre- and postintervention periods and multiple units in the treated and comparison groups. In this case, researchers often fit two-way fixed effects (TWFE) linear regression models, where "two-way" refers to unit and time fixed effects (de Chaisemartin and D'Haultfœuille (2023)). The models contain an interaction between an indicator of the postintervention period and treated group, the coefficient of which is interpreted as an estimator of the ATT. This approach can be justified by the equivalence of the TWFE estimator and the DID estimator (Angrist and Pischke (2008), Egami and Yamauchi (2023), Imai and Kim (2019), Kropko and Kubinec (2020), Sobel (2012), Wooldridge (2005)) in a particular setting, which leads to the popular impression that TWFE model identification is also by a parallel trends assumption. However, the equivalence does not extend to the more general setting. Imai and Kim (2021) show that the TWFE model's promise of simultaneous adjustment for unobserved unit and time confounders depends crucially on linearity and additivity.

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

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

in which $\mathrm{E}_{\mathcal{P}}[\epsilon_{u,t}|\alpha_u, G_u] = 0$, where $u = 1, \ldots, U$. We show in the Supplementary Material (Leavitt and Hatfield (2025), Section 2.2) that when this structural model holds, Assumption 2 also holds if the prediction model is $f(\boldsymbol{X}_T) = \arg\min_{\alpha_u} \sum_{t=1}^{T-1}(Y_{u,t} - \alpha_u)^2$. This prediction model is the population-level ordinary least squares (OLS) solution to the unit fixed effects model's objective function fit to data before period T, which is equivalent to the mean of a unit's outcomes in all pre-treatment periods.

Again, this is a simple instance of a TWFE structural model. In the Supplementary Material (Leavitt and Hatfield (2025)), we also show that this idea extends to similar structural models that include unit- or group-specific time trends (see Section 2.4) or lagged dependent variables (see Section 2.5). Many researchers fit more complicated models and obtain estimators from them. We have not proved that these are also special cases of Assumption 2.

2.2. *Existing designs that are not special cases*. The designs considered above all use a pre-post contrast (to account for time-invariant group differences) and a treated-comparison contrast (to account for common shocks). Likewise, our proposed framework uses a prediction step (leveraging predictable features of each group's outcome trajectories) and a correction step (leveraging the comparison group to correct for unexpected shocks). By contrast, some designs lack an analog of either the prediction or correction steps. The ITS design uses only a pre-post contrast; there are no comparison units with which to perform our correction step. Synthetic control uses only a treated-comparison contrast, omitting the pre-post contrast. In the Supplementary Material (Leavitt and Hatfield (2025), Section 2.6) we provide more detail on the question of synthetic control, showing that it is not a special case of our framework.

2.3. *Staggered adoption*. Thus far, we have assumed that all treated units receive intervention at the same time. We now extend to staggered adoption settings, taking the perspective of Callaway and Sant'Anna (2021). That is, we consider each treatment adoption time as its own simple design, identify the treatment effect in each and weight these effects together in a sensible way.

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

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

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

$$\mathrm{ATT}(g) := \mathrm{E}_{\mathcal{P}}\big[Y_g(g) - Y_g(0)|G_g = 1\big].$$

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

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

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

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

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

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

The robustness of a model to violations of an identifying assumption is different from the plausibility that a model is "correct," that is, satisfies an identifying assumption. In the name of assessing plausibility that a model is "correct," researchers often study whether a version of an identification assumption holds in the preperiod. For example, in DID designs it is common to test for nonparallel trends in the preperiod, 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 preperiod would have continued into the postperiod in the absence of treatment. In other words, as Egami and Yamauchi (2023) explain, this approach replaces one unverifiable assumption about counterfactual outcomes with another. The framework of design sensitivity offers a practical way out of this bind: we study the robustness of our inference to violations of the point identifying assumption, grounded in empirical evidence about the potential magnitude of those violations.

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

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

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

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

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

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

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

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

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

The proof is in the Supplementary Material (Leavitt and Hatfield (2025), Section 1.2).

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

How is choosing the most robust model different from choosing the "correct" model? Suppose that $M = 0$ holds for one model, implying that Assumption 2 is satisfied, but that this model is nonetheless less robust (by our criterion) than another candidate model. Proposition 2 quantifies the consequences of this trade-off between "correctness" and robustness.

PROPOSITION 2. *Suppose $M = 0$ holds for f', which implies that Assumption 2 is satisfied for f', but that f' is less robust than f, as defined in Proposition 1. The difference between the ATT of the correct model and the population-level difference in expected prediction errors of the robust model is*

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

The proof is in the Supplementary Material (Leavitt and Hatfield (2025), Section 1.3).

Proposition 2 shows that when a model's differential prediction errors in the validation periods provide "misleading" information about (unobservable) differential prediction error in the postperiod, our conclusions will suffer. This is related to the idea that conclusions are more robust if point estimates are stable across competing models (Brown and Atal (2019),

O'Neill et al. (2016)). In our framework, two prediction models that yield identical point estimates for $M = 0$ can have quite different robustness for $M > 0$. However, if two models yield identical point estimates, then equation (10) is equal to 0. Therefore, stable point estimates across prediction models do not imply our conclusions are more robust, but they do mitigate a potential trade-off in which choosing the most robust model could come at the expense of choosing the "correct" model.

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

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

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

where $\boldsymbol{Y}_{i,<t}$ and $\boldsymbol{X}_{i,<t}$ are the outcomes and predictors from $t = 1, \ldots, t-1$ and $Y_{i,t}$ and $\boldsymbol{X}_{i,t}$ are outcomes and predictors in period t. The respective collections of $\boldsymbol{Y}_{i,<t}$ and $\boldsymbol{X}_{i,<t}$ over all $i = 1, \ldots, n$ units are $\boldsymbol{Y}_{<t} := \{\boldsymbol{Y}_{1,<t}, \ldots, \boldsymbol{Y}_{n,<t}\}$ and $\boldsymbol{X}_{<t} := \{\boldsymbol{X}_{1,<t}, \ldots, \boldsymbol{X}_{n,<t}\}$. We can collect the data in equation (11) over all $i = 1, \ldots, n$ units into $\boldsymbol{D}_t := \{\boldsymbol{D}_{1,t}, \ldots, \boldsymbol{D}_{n,t}\}$, over times into $\boldsymbol{D}_i := \{\boldsymbol{D}_{i,1}, \ldots, \boldsymbol{D}_{i,T}\}$ and over all units and all times into $\boldsymbol{D} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n\}$.

We assume the following condition on how these data are sampled.

ASSUMPTION 3. For all $i = 1, \ldots, n$, the sample data, $\{\boldsymbol{D}_i\}$, are independent and identically distributed (i.i.d.).

Assumption 3 states that i.i.d. sampling occurs at the cluster level, where the clusters are the individual units indexed by $i = 1, \ldots n$.

Next, we set up the prediction models in the OLS framework. Before we proceed, we need an additional assumption that places conditions on the population moments. This assumption applies to each corresponding matrix of predictors for each of the candidate models in $\mathcal{F}$.

ASSUMPTION 4 (Population moment conditions). For groups $G = 0$ and $G = 1$, $\mathrm{E}_{\mathcal{P}}[\boldsymbol{Y}_t | G = g] < \infty$ and $\mathrm{E}_{\mathcal{P}}[\|\boldsymbol{X}_t\|^2 | G = g] < \infty$ for all $t = 1, \ldots, T$, and $\mathrm{E}_{\mathcal{P}}[\boldsymbol{X}_{<t}\boldsymbol{X}_{<t}^{\top} | G = g]$ is positive definite for all $t = 2, \ldots, T$.

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

We write the model f for group g in period t as a function of both predictors and parameters, $f(\boldsymbol{X}_{i,t}; \boldsymbol{\beta}_{f,g,t})$, where $\boldsymbol{\beta}_{f,g,t} \in \Re^K$ (in which K is the dimension of $\boldsymbol{X}_{i,t}$ and $\Re$ is the set of real numbers). Note that $\boldsymbol{\beta}_{f,g,t}$ is simply a collection of linear projection coefficients

for a particular model f in group g based on the population-level OLS solution to the linear regression model's objective function in periods before t. That is, under Assumption 4,

$$\boldsymbol{\beta}_{f,g,t} = \left(\mathrm{E}_{\mathcal{P}}[\boldsymbol{X}_{<t}\boldsymbol{X}_{<t}^{\top}|G=g]\right)^{-1}\mathrm{E}_{\mathcal{P}}[\boldsymbol{X}_{<t}\boldsymbol{Y}_{<t}|G=g]. \tag{12}$$

The collection of estimated coefficients, $\hat{\boldsymbol{\beta}}_{f,g,t}$, is the sample analog of equation (12). We collect the estimated coefficients over groups into $\hat{\boldsymbol{\beta}}_{f,t} := \{\hat{\boldsymbol{\beta}}_{f,1,t}, \hat{\boldsymbol{\beta}}_{f,0,t}\}$. We denote the estimated coefficients collected over times by $\hat{\boldsymbol{\beta}}_{f}$ and the estimated coefficients collected over models by $\hat{\boldsymbol{\beta}}_{t}$. The collection of the estimated coefficients over all models and times is $\hat{\boldsymbol{\beta}}$ and the collection over all models and pretreatment validation times is $\hat{\boldsymbol{\beta}}_{\mathcal{V}}$.

With Assumptions 3 and 4 in hand, we write a point estimator of $\delta_{f,t}$ as

$$\begin{aligned}\hat{\delta}(\boldsymbol{D}_t, \hat{\boldsymbol{\beta}}_{f,t}) := &\left(\frac{1}{n_1}\right)\sum_{i=1}^{n}\mathbb{1}\{G_i=1\}Y_{i,t} - \left(\frac{1}{n_1}\right)\sum_{i=1}^{n}\mathbb{1}\{G_i=1\}\boldsymbol{X}_{i,t}\hat{\boldsymbol{\beta}}_{f,1,t} \\ &- \left[\left(\frac{1}{n_0}\right)\sum_{i=1}^{n}\mathbb{1}\{G_i=0\}Y_{i,t} - \left(\frac{1}{n_0}\right)\sum_{i=1}^{n}\mathbb{1}\{G_i=0\}\boldsymbol{X}_{i,t}\hat{\boldsymbol{\beta}}_{f,0,t}\right],\end{aligned} \tag{13}$$

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

$$\hat{\Delta}(\boldsymbol{D}, \hat{\boldsymbol{\beta}}_f, M) := \hat{\delta}(\boldsymbol{D}_T, \hat{\boldsymbol{\beta}}_{f,T}) \pm M \max_{v\in\mathcal{V}}|\hat{\delta}(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}_{f,v})|. \tag{14}$$

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

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

The usual approach of splitting data into testing and training subsets is not feasible. We cannot split the data "vertically" (i.e., in time) because our estimators and model selection criterion use the same data *by construction*: terms in equation (14) use data from pretreatment validation periods $\mathcal{V}$. Nor can we rely on splitting the data "horizontally:" many applications (including the one we consider here) have only a single or a few treated units, so we cannot afford to split the units.

Therefore, we propose to use a BMA estimator, which averages the estimates across models, weighting each by the model's posterior probability of being the most robust in the population. We write this estimator as

$$\widehat{\mathrm{E}}_{\mathcal{F}|\boldsymbol{D}}[\hat{\Delta}(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M)] := \sum_{f\in\mathcal{F}}\hat{\Delta}(\boldsymbol{D}, \hat{\boldsymbol{\beta}}_f, M)\hat{p}_f, \tag{15}$$

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

How do we generate these posterior probabilities? We extend what Gelman and Hill (2006, p. 140) refer to as their "informal Bayesian approach." This has been employed by many researchers (e.g., King, Tomz and Wittenberg (2000), Tomz, Wittenberg and King (2003)), including in ITS designs (Miratrix (2022)). The idea is to generate samples from the "informal" posterior of all the coefficients across all prediction models and pretreatment validation

periods. For this distribution we use a multivariate Normal with a mean equal to the estimated coefficients $\hat{\boldsymbol{\beta}}_{\mathcal{V}}$ (collected over all validation times and models in $\mathcal{F}$) and a variance given by their estimated (robust) variance-covariance matrix clustered at the individual level (Liang and Zeger (1986), Arellano (1987)), $\widehat{\boldsymbol{\Sigma}}_{\mathcal{V}}$. This approach is equivalent to the posterior distribution of the models' parameters if the prior were flat. To estimate the variance-covariance of all the parameters across all the model and time periods simultaneously, we use seemingly unrelated regression tools pioneered by Zellner (1962, 1963) (see also Mize, Doan and Long (2019)) detailed in the Supplementary Material (Leavitt and Hatfield (2025), Section 3).

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

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

$$\hat{p}_f := \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\Big\{ f = \arg\min_{f \in \mathcal{F}} \max_{v \in \mathcal{V}} |\delta(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}_{f,v}^{(s)})| \Big\}, \tag{16}$$

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

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

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

The proof is given in the Supplementary Material (Leavitt and Hatfield (2025), Section 1.4). As we show next, this lemma implies that our BMA estimator is consistent.

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

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

The proof is also given in the Supplementary Material (Leavitt and Hatfield (2025), Section 1.5). Proposition 3 shows that the BMA estimator converges in probability to the same limit as that of an estimator in which the optimal model in the population is known before observing data. We provide a conceptual diagram of the overall estimation process in Figure 1 of the Supplementary Material (Leavitt and Hatfield (2025), Section 4).

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

$$\widehat{\mathrm{Var}}_{\mathcal{F}|\boldsymbol{D}}\big[\hat{\Delta}(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M)\big] := \sum_{f \in \mathcal{F}} \big(\hat{\Delta}(\boldsymbol{D}, \hat{\boldsymbol{\beta}}_f, M) - \widehat{\mathrm{E}}_{\mathcal{F}|\boldsymbol{D}}\big[\hat{\Delta}(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M)\big]\big)^2 \hat{p}_f. \tag{17}$$

Then, let $r = 1, \ldots, R$ index resamples of the data, and denote the variance of our estimator over R resamples, holding fixed the observed posterior, as

$$\widehat{\mathrm{Var}}_{\boldsymbol{D}^{(r)}|\mathcal{F}}\big[\widehat{\mathrm{E}}_{\mathcal{F}|\boldsymbol{D}}[\hat{\Delta}(\boldsymbol{D}^{(r)}, \widehat{\boldsymbol{\beta}}^{(r)}, M)]\big] := \frac{1}{R}\sum_{r=1}^{R}\Bigg(\widehat{\mathrm{E}}_{\mathcal{F}|\boldsymbol{D}}[\hat{\Delta}(\boldsymbol{D}^{(r)}, \widehat{\boldsymbol{\beta}}^{(r)}, M)] - \frac{1}{R}\sum_{r=1}^{R}\widehat{\mathrm{E}}_{\mathcal{F}|\boldsymbol{D}}[\hat{\Delta}(\boldsymbol{D}^{(r)}, \widehat{\boldsymbol{\beta}}^{(r)}, M)]\Bigg)^2. \tag{18}$$

In practice, we draw R resamples of the data via the fractional weighted bootstrap (Xu et al. (2020)). The overall variance estimator, accounting for both sampling and model uncertainty, of the BMA estimator in equation (15) is the sum of equations (17) and (18). Confidence intervals can then be constructed by drawing on a Normal approximation. We demonstrate the coverage properties of this approach's 95% confidence intervals through simulations presented in the Supplementary Material (Leavitt and Hatfield (2025), Section 6). We also observe the conservatism that Antonelli, Papadogeorgou and Dominici (2022, p. 103) note.

5. The effect of gun laws on violent crime. We now return to our analysis of Missouri's 2007 repeal of its permit-to-purchase law. 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. Our data comprise state-year observations of the homicide rate in Missouri and each of its eight neighboring comparison states from 1994 to 2016. For simplicity, we recode the data so that there is a single posttreatment period (denoted by "2008+" in Figure 1) in which each state's outcome in 2008 is the average of that state's outcomes over all posttreatment periods (2008–2016). To estimate the repeal's impact on gun homicides, we form a set of candidate prediction models drawn from the gun policy literature. Researchers agree on a basic model with unit fixed effects (as in Webster, Crifasi and Vernick (2014)) but disagree on other model components. Based on our survey of the literature, we divide the relevant model components into three categories:

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

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

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

Obviously, this framework leaves out some modeling variations. For example, some studies in the gun policy literature employ random effects (Crifasi et al. (2018)) and two-stage models (Rubin and Dezhbakhsh (2003)). However, given the prominence of these three model

TABLE 1
Model components used to create a set of candidate prediction models

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

components as well as unit fixed effects and linear models, we believe the resulting set of candidate models is reasonably broad and relevant to the gun policy literature.

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

To select among the 18 prediction models, we estimate the differences in average prediction errors between treated and comparison groups. For each year prior to the law's passage in 2007, we train our prediction models on the previous years. For example, in 2006, we train a model on data from 1994 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 five or more years of training data, even in the first validation year (1999, for which we train the model on data from 1994–1998).

Figure 2 shows the absolute differential average prediction errors for all 18 models over all validation years, with the maximum for each model highlighted in black. The LDV, that is, AR(1), model with unit fixed effects fit to the log of the outcome (row 1, column 5) minimizes our sensitivity criterion on the sample data. The baseline mean model with unit fixed effects

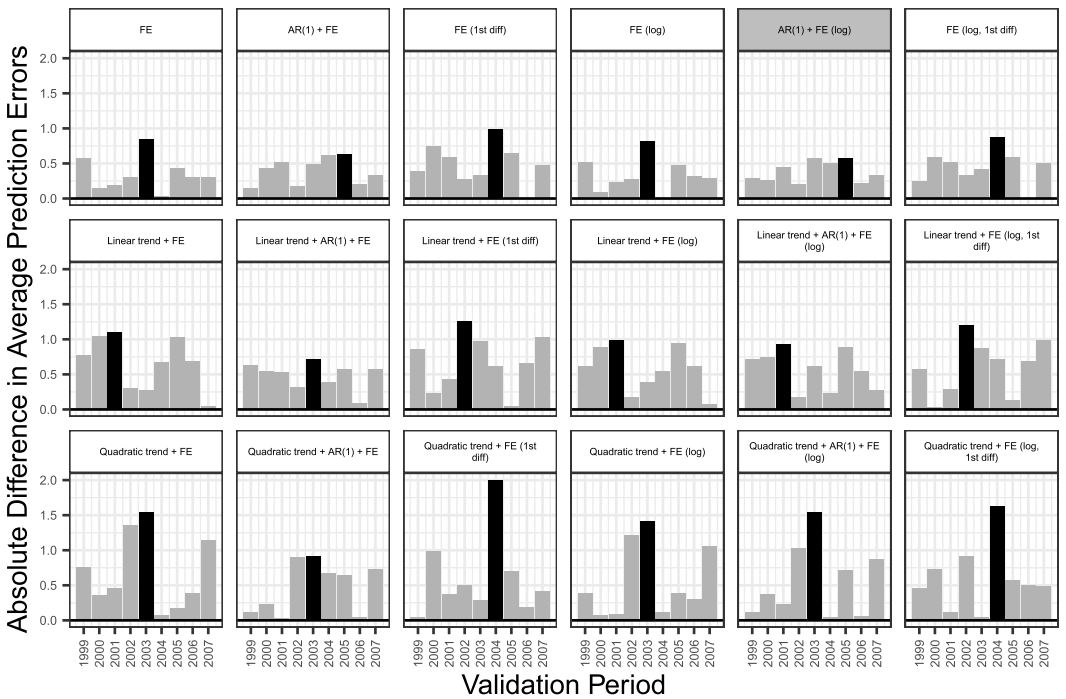

FIG. 2. *Absolute difference in average prediction errors for all candidate models. The maximum for each model is highlighted in black. The optimal model is highlighted in gray.*

(row 1, column 1), which is arguably the closest correspondent to the model of choice in Hasegawa, Webster and Small (2019), is the fifth-best model.

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 preperiod (2007) is the linear time trend model with unit fixed effects (row 2, column 1). By contrast, the prediction model with the smallest absolute difference in average prediction errors, averaged over all validation periods, is the baseline mean model with unit fixed effects fit to the log of the outcome (row 1, column 4). These different loss functions for choosing the optimal model can be justified by an appropriate sensitivity analysis model. Given the sensitivity analysis in equation (9), which aligns with the sensitivity analysis proposed in recent research (Rambachan and Roth (2023)), the aforementioned LDV model with unit fixed effects fit to the outcome's log scale is optimal.

Figure 3 shows the relationship between models' point estimates and their robustness. As this figure illustrates, the potential trade-offs between models' "correctness" and robustness are not especially severe. The standard deviation of point estimates across models is 0.38. In addition, the most and least robust models yield relatively similar point estimates of 1.14 and 1.88, respectively. In contrast to other application in the gun policy literature (National Research Council of the National Academies (2005), Morral et al. (2018), Smart et al. (2020)), this similarity of point estimates across models appears atypical.

Although point estimates may be similar across models, these models can differ in terms of robustness. Nevertheless, much value remains in the similarity of point estimates across models. As Proposition 2 shows, if point identification of $M = 0$ happens to be true under one model that is not the most robust, then the point estimate under the most robust model will not be too misleading insofar as the estimates under both models are similar.

Turning to estimation and inference, this empirical setting requires careful attention to the sources of randomness. In the setting of gun policy research, an influential article by Manski and Pepper (2018) argues that it is often difficult to conceive the units of analysis as randomly sampled from a target population of interest: "Random sampling assumptions, however, are not natural when considering states or countries as units of observations" (Manski and Pepper (2018), p. 235). Instead, in the setting of most gun policy research, as Manski and Pepper

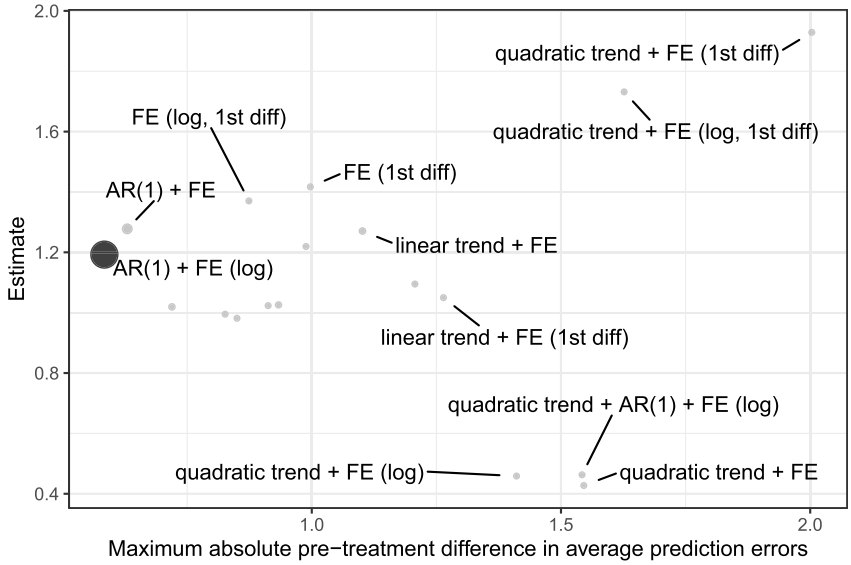

FIG. 3. *Estimates under each model (y axis) and corresponding maximum absolute differential prediction errors in the preperiod (x-axis). Point size is proportional to the model's posterior weight.*

(2018) argue, uncertainty is driven by a fundamental ambiguity over whether counterfactual point identification assumptions hold—that is, what Rambachan and Roth ((2023), p. 2556) call "identification uncertainty."

In a setting characterized by only identification rather than sampling uncertainty, Rambachan and Roth ((2023), p. 2563) argue that a natural starting point for controlled pre-post designs is one of set identification with $M = 1$. In this set identification framework (as opposed to point identification in which $M = 0$), researchers can then gradually increase M in a subsequent sensitivity analysis. A crucial feature of this inferential setting is the absence of uncertainty over which model is truly optimal.

In this setting, one could deterministically select the truly optimal model. Then, given the selection of this optimal model, it would be straightforward to calculate bounds on the ATT under $M = 1$ and to assess the sensitivity of these bounds under increasing values of M. Under this approach the bounds of the ATT (with $M = 1$) under the most robust model is $[0.61, 1.78]$. The changepoint value of M, that is, the smallest value of M at which the estimated lower and upper bounds of the ATT bracket 0, is 2.04.

The analysis above supposes the setting that Manski and Pepper (2018) argue is most sensible for our application. However, if we suppose that states are independent and identically distributed draws from a target population, then the estimation and inferential procedure in Section 4 is appropriate. The BMA point estimate (under $M = 0$) of 1.2 is nearly identical to the point estimate under the optimal model (1.19) in the realized sample. The reason for this similarity is because the optimal model in the sample (LDV with unit fixed effects fit to the outcome's log scale) receives high posterior probability of being the population's optimal model (approximately 0.97). The model with the second greatest posterior probability of approximately 0.03 is the next best model in the sample data (the LDV model with unit fixed effects fit without the log transformation of the outcome). Figure 4 below shows the full posterior distribution given the sample data, where the x-axis includes only the models in the support of the observed posterior.

Our proposed variance estimation procedure, which we would expect not to perform at its best in small samples, yields an estimated standard error (accounting for both model and sampling uncertainty) of 0.14 and corresponding 95% confidence interval of $[0.93, 1.46]$. That is, we conclude that the repeal of Missouri's permit-to-purchase law increased the state's

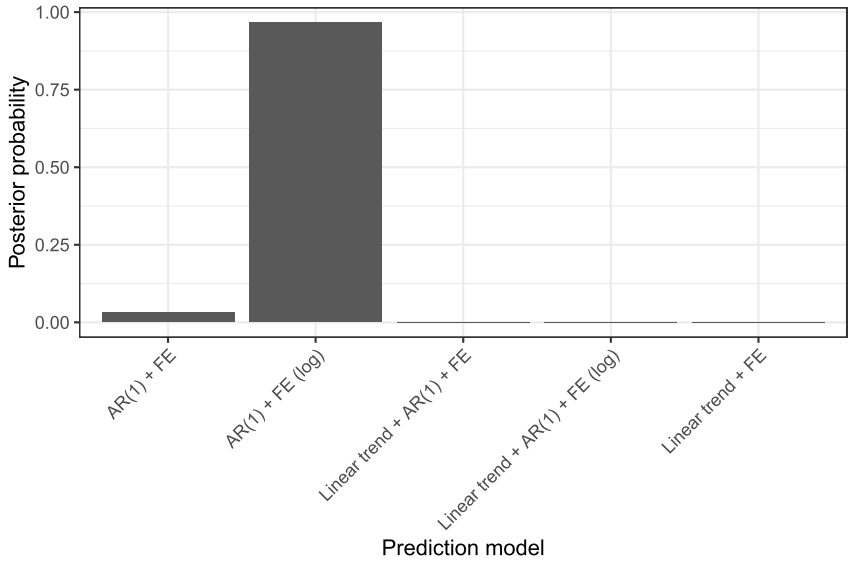

FIG. 4. *Posterior plausibility that each candidate model is optimal.*

gun homicide rate by somewhere between 0.93 to 1.46 per 100,000 people. The observed homicide rate in 2007 (just before the repeal) in Missouri was 4.5, so the point estimate of roughly 1.2 represents a 27% increase.

For context, Webster, Crifasi and Vernick (2014) state that they estimate an increase of 1.09 per 100,000 (+23%); Hasegawa, Webster and Small (2019) state that they estimate an increase of 1.2 per 100,000 (+24%) using standard DID methods and increases between 0.9 and 1.3 per 100,000 (+17% to +27%) with their bracketing approach. Using synthetic control methods, other authors estimate that Connecticut's *adoption* of a permit-to-purchase handgun law decreased firearm homicides by 40% (Rudolph et al. (2015)). Thus, our estimate is on the higher end of estimates of the effect for Missouri's policy change specifically, and there is some evidence that the effects of implementation and repeal of these kinds of laws may be asymmetric.

The changepoint value of M for the BMA estimator is 2.04—slightly smaller (but only by a rounding margin) than the changepoint value of M obtained without accounting for model uncertainty under the most robust model. The changepoint value of M at which the lower bound estimator's 95% confidence interval no longer excludes 0 is approximately 1.25. This smaller value is expected in a study with a limited sample size.

6. Conclusion and open questions. In this paper we introduce a new method for causal inference in controlled pre-post settings, *averaged prediction models* (or *APM*). We began by introducing a general identification framework for a broad class of prediction models in which one *predicts* untreated potential outcomes and *corrects* these predictions using the observable prediction error in the comparison group. We have shown that several popular designs are special cases of our general identification framework. Then to choose among the set of candidate prediction models, we propose a BMA procedure based on each model's robustness given preperiod data.

We applied these ideas to reconcile disparate models and assumptions from gun policy evaluations. Specifically, we studied the repeal of Missouri's permit-to-purchase law in 2007 using models drawn from the literature. Rather than make claims that any one underlying causal model is "correct," we selected the optimal model based on robustness. We found that a lagged dependent variable model with unit fixed effects, fit to the outcome's log scale, minimized our robustness criterion in our sample, making this model the most likely to be the truly optimal model in the population (although other models are plausible as well). Our overall point estimate, averaging over the posterior probability that each model is optimal in the population, was an increase of 1.2 homicides per 100,000 people.

Our sensitivity bounds would include 0 for $M \geq 2.04$. That is, the violation of Assumption 2 would have to be at least 2.04 times greater than a weighted combination of each model's worst violation in the *nine* validation years. By contrast, in the absence of our Bayesian model selection procedure, the value of M that leads the sensitivity bounds to include 0—when conditioning on any single candidate model—could be much smaller, as low as 0.28, with an unweighted average (across all models) of 1.12.

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 preperiod has similar conceptual problems to testing for differential pretrends in DID. This is why we use a sensitivity perspective to choose a prediction model based on robustness.

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

Third, our prediction models use only variables that are measured prior to (or are independent of) treatment. For some data-generating models, such as interactive fixed effects, the correction step will not de-bias 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' postperiod outcomes, while other comparison units' postperiod outcomes are used to correct for unexpected common shocks.

Fourth, by switching to a robustness criterion for model selection, we induce a possible "correctness" vs. robustness trade-off (Proposition 2). Rather than claim that we can choose the "correct" model, we choose a model that maximizes our robustness criterion. A model for which our identifying assumption (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 appropriately account for uncertainty in both the model and data, may not sufficiently do so in all scenarios. For example, bootstrap methods perform poorly when there are few clusters, as in our analysis with only one treated unit and eight comparison units (Bertrand, Duflo and Mullainathan (2004), MacKinnon and Webb (2020), Conley and Taber (2011), Rokicki et al. (2018)). However, we still believe that our proposal for formally accounting for the model selection procedure is an improvement over the status quo in which model selection is usually hidden from view and outside the bounds of inference entirely. Postselection inference is an active area of research and, as a recent review article noted, "has a long and rich history, and the literature has grown beyond what can reasonably be synthesized in our review" (Kuchibhotla, Kolassa and Kuffner (2022), p. 506). Future research should explore the application of these simultaneous inference and conditional selective inference methods to problems like ours in which sample splitting is not feasible.

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

Second, many researchers already interpret robustness in terms of "correctness." In DID, for instance, researchers interpret parallel trends in the preperiod as evidence for the plausibility of parallel trends from the pre- to postperiods. Yet violations of preperiod parallel trends can still be consistent with the identifying assumption (Kahn-Lang and Lang (2020), Roth and Sant'Anna (2023), Egami and Yamauchi (2023)). Therefore, our proposal offers a more transparent version of this practice, recasting the evaluation of preperiod violations in terms of a sensitivity analysis rather than as a test of untestable assumptions.

Third, we show that our identification framework unifies a wide variety of prediction models researchers employ in practice. We also show that some familiar designs are special cases of our general identifying 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, we provide an `R` package (`apm`), which implements our estimation and inference procedures. The package includes functions that create a variety of prediction models and fit

them to the preintervention data. It then extracts the differential prediction errors and computes a model-averaged estimate and standard error using this paper's Bayesian and bootstrap procedures, which account for both sampling and model uncertainty.

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

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 or the Agency for Healthcare Research and Quality.

SUPPLEMENTARY MATERIAL

Supplement to "Averaged Prediction Models (APM): Identifying causal effects in controlled pre-post settings with application to gun policy" (DOI: 10.1214/25-AOAS2011SUPP; .pdf). Supplementary material for this article include: *Supplementary PDF*. Contains complete mathematical proofs of all theoretical results stated in the main text. Also includes derivations showing how existing designs are special cases (or not) of our general framework, a detailed description of the joint variance-covariance matrix used for Bayesian model averaging, a conceptual diagram of the estimation procedure, a full list of prediction models used in the applied analysis, and additional simulation results evaluating the performance of our procedure across various sample sizes. *Replication Archive*. Includes the full dataset and `R` scripts required to replicate all empirical results, figures, and simulation studies. The archive also documents how the analysis dataset was constructed from the raw data. A `README` file is provided to guide users through the replication process.

REFERENCES

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 Univ. 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. *J. Econ. Perspect.* **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. *J. Roy. Statist. Soc. Ser. A* **186** 772–787. MR4754074 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. MR4408573 https://doi.org/10.1111/biom.13417

ARELLANO, M. (1987). Computing robust standard errors for within-group estimators. *Oxf. Bull. Econ. Stat.* **49** 431–434.

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. *Ann. Appl. Stat.* **17** 985–1016. MR4582700 https://doi.org/10.1214/22-aoas1654

BERTRAND, M., DUFLO, E. and MULLAINATHAN, S. (2004). How much should we trust differences-in-differences estimates? *Q. J. Econ.* **119** 249–275.

BLACK, D. A. and NAGIN, D. S. (1998). Do right-to-carry laws deter violent crime? *J. Leg. Stud.* **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. *Eval. Rev.* **27** 3–49.

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 Soc. Rev.* **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 Econ.* **28** 280–298.

CALLAWAY, B. and SANT'ANNA, P. H. C. (2021). Difference-in-differences with multiple time periods. *J. Econometrics* **225** 200–230. MR4328640 https://doi.org/10.1016/j.jeconom.2020.12.001

CONLEY, T. G. and TABER, C. R. (2011). Inference with 'difference in differences' with a small number of policy changes. *Rev. Econ. Stat.* **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. *J. Urban Health* **95** 383–390.

DE CHAISEMARTIN, C. and D'HAULTFŒUILLE, X. (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *Econom. J.* **26** C1–C30. MR4643826 https://doi.org/10.1093/ectj/utac017

DRAPER, D. (1995). Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. Ser. B, Methodol.* **57** 45–97. MR1325378

DUWE, G., KOVANDZIC, T. and MOODY, C. E. (2002). The impact of right-to-carry concealed firearm laws on mass public shootings. *Homicide Stud.* **6** 271–296.

EGAMI, N. and YAMAUCHI, S. (2023). Using multiple pre-treatment periods to improve difference-in-differences and staggered adoption designs. *Polit. Anal.* **31** 195–212.

FRENCH, B. and HEAGERTY, P. J. (2008). Analysis of longitudinal data to evaluate a policy change. *Stat. Med.* **27** 5005–5025. MR2528779 https://doi.org/10.1002/sim.3340

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

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

GRANGER, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37** 424–438.

GUO, K. and BASSE, G. (2023). The generalized Oaxaca-Blinder estimator. *J. Amer. Statist. Assoc.* **118** 524–536. MR4571139 https://doi.org/10.1080/01621459.2021.1941053

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.

HELLER, R., ROSENBAUM, P. R. and SMALL, D. S. (2009). Split samples and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **104** 1090–1101. MR2750238 https://doi.org/10.1198/jasa.2009.tm08338

HSU, J. Y., SMALL, D. S. and ROSENBAUM, P. R. (2013). Effect modification and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **108** 135–148. MR3174608 https://doi.org/10.1080/01621459.2012.742018

HUANG, M., SORIANO, D. and PIMENTEL, S. D. (2024). Design sensitivity and its implications for weighted observational studies. arXiv Preprint. Available at https://arxiv.org/pdf/2307.00093.

IMAI, K. and KIM, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *Amer. J. Polit. Sci.* **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. *Polit. Anal.* **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. *J. Bus. Econom. Statist.* **38** 613–620. MR4115421 https://doi.org/10.1080/07350015.2018.1546591

KING, G., TOMZ, M. and WITTENBERG, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *Amer. J. Polit. Sci.* **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. *Annu. Rev. Stat. Appl.* **9** 505–527. MR4394918 https://doi.org/10.1146/annurev-statistics-100421-044639

LEAVITT, T. and HATFIELD, L. A (2025). Supplement to "Averaged Prediction Models (APM): Identifying Causal Effects in Controlled Pre-Post Settings with Application to Gun Policy." https://doi.org/10.1214/25-AOAS2011SUPP

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430 https://doi.org/10.1093/biomet/73.1.13

LOPEZ BERNAL, J., SOUMERAI, S. and GASPARRINI, A. (2018). A methodological framework for model selection in interrupted time series studies. *J. Clin. Epidemiol.* **103** 82–91.

MACKINNON, J. G. and WEBB, M. D. (2020). Randomization inference for difference-in-differences with few treated clusters. *J. Econometrics* **218** 435–450. MR4149234 https://doi.org/10.1016/j.jeconom.2020.04.024

MADIGAN, D. and RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *J. Amer. Statist. Assoc.* **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. *Rev. Econ. Stat.* **100** 232–244.

MCDOWALL, D., MCCLEARY, R. and BARTOS, B. J. (2019). *Interrupted Time Series Analysis*. Oxford Univ. Press, Oxford. MR4283510 https://doi.org/10.1093/oso/9780190943943.001.0001

MIRATRIX, L. W. (2022). Using simulation to analyze interrupted time series designs. *Eval. Rev.* **46** 750–778.

MIZE, T. D., DOAN, L. and LONG, J. S. (2019). A general framework for comparing predictions and marginal effects across models. *Sociol. Method.* **49** 152–189.

MOODY, C. E. (2001). Testing for the effects of concealed weapons laws: Specification errors and robustness. *J. Law Econ.* **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. *Rev. Econ. 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. et al. (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. *J. Econometrics* **49** 169–193.

NICKELL, S. (1981). Biases in dynamic models with fixed effects. *Econometrica* **49** 1417–1426. MR0636160 https://doi.org/10.2307/1911408

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

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 Serv. Outcomes Res. Methodol.* **16** 1–21.

PIIRONEN, J. and VEHTARI, A. (2017). Comparison of Bayesian predictive methods for model selection. *Stat. Comput.* **27** 711–735. MR3613594 https://doi.org/10.1007/s11222-016-9649-y

PLASSMANN, F. and TIDEMAN, T. N. (2001). Does the right to carry concealed handguns deter countable crimes? Only a count analysis can say. *J. Law Econ.* **44** 771–798.

RAFTERY, A. E., MADIGAN, D. and HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* **92** 179–191. MR1436107 https://doi.org/10.2307/2291462

RAMBACHAN, A. and ROTH, J. (2023). A more credible approach to parallel trends. *Rev. Econ. Stud.* **90** 2555–2591. MR4636242 https://doi.org/10.1093/restud/rdad018

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. *Med. Care* **56** 97–105.

ROSENBAUM, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91** 153–164. MR2050466 https://doi.org/10.1093/biomet/91.1.153

ROSENBAUM, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist.* **59** 147–152. MR2133562 https://doi.org/10.1198/000313005X42831

ROSENBAUM, P. R. (2012). An exact adaptive test with superior design sensitivity in an observational study of treatments for ovarian cancer. *Ann. Appl. Stat.* **6** 83–105. MR2951530 https://doi.org/10.1214/11-AOAS508

ROTH, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *Amer. Econ. Rev. 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. MR4566237 https://doi.org/10.3982/ecta19402

RUBIN, P. H. and DEZHBAKHSH, H. (2003). The effect of concealed handgun laws on crime: Beyond the dummy variables. *Int. Rev. Law Econ.* **23** 199–216.

RUDOLPH, K. E., STUART, E. A., VERNICK, J. S. and WEBSTER, D. W. (2015). Association between Connecticut's permit-to-purchase handgun law and homicides. *Amer. J. Publ. Health* **105** 49–54.

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 Serv. Res.* **50** 1211–1235.

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.

SMALL, D. S., CHENG, J., HALLORAN, M. E. and ROSENBAUM, P. R. (2013). Case definition and design sensitivity. *J. Amer. Statist. Assoc.* **108** 1457–1468. MR3174721 https://doi.org/10.1080/01621459.2013.820660

SMART, R., MORRAL, A. R., SMUCKER, S., CHERNEY, S., SCHELL, T. L., PETERSON, S., AHLUWALIA, S. C., CEFALU, M., XENAKIS, L. et al. (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. *J. Amer. Statist. Assoc.* **107** 521–529. MR2980064 https://doi.org/10.1080/01621459.2011.646917

TOMZ, M., WITTENBERG, J. and KING, G. (2003). Clarify: Software for interpreting and presenting statistical results. *J. Stat. Softw.* **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. *J. Clin. Pharm. Ther.* **27** 299–309.

WEBSTER, D., CRIFASI, C. K. and VERNICK, J. S. (2014). Effects of the repeal of Missouri's handgun purchaser licensing law on homicides. *J. Urban Health* **91** 293–302.

WOLFERS, J. (2006). Did unilateral divorce laws raise divorce rates? A reconciliation and new results. *Amer. Econ. Rev.* **96** 1802–1820.

WOOLDRIDGE, J. M. (2005). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Rev. Econ. Stat.* **87** 385–390.

XU, L., GOTWALT, C., HONG, Y., KING, C. B. and MEEKER, W. Q. (2020). Applications of the fractional-random-weight bootstrap. *Amer. Statist.* **74** 345–358. MR4168255 https://doi.org/10.1080/00031305.2020.1731599

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57** 348–368. MR0139235

ZELLNER, A. (1963). Estimators for seemingly unrelated equations: Some exact finite sample results. *J. Amer. Statist. Assoc.* **58** 977–992. MR0157439

ZHANG, F. and PENFOLD, R. B. (2013). Use of interrupted time series analysis in evaluating health care quality improvements. *Acad. Pediatr.* **13** S38–S44.

Submitted to the Annals of Applied Statistics

AVERAGED PREDICTION MODELS (APM): IDENTIFYING CAUSAL EFFECTS IN CONTROLLED PRE-POST SETTINGS WITH APPLICATION TO GUN POLICY

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]*Statistics and Data Science Department, NORC at the University of Chicago,* [b]*hatfield-laura@norc.org*

SUPPLEMENTARY MATERIAL

1. Proofs.

1.1. *Proof of Theorem 1.*

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

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

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

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

Equal-expected-prediction-errors in Assumption 2 of the manuscript then implies that

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

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

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

thereby completing the proof. □

1.2. *Proof of Proposition 1.*

PROOF. The proof is immediate from the the ATT's lower and upper bounds in Eq. 9 of the manuscript: The difference between the upper and lower bounds of the ATT is

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

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

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

□

1.3. *Proof of Proposition 2.*

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

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

Then taking the difference between

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

the population-level difference in expected prediction errors under the robust model, and the ATT in Eq. 3 yields

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

thereby completing the proof. □

1.4. *Proof of Lemma 1.*

PROOF. The proof proceeds in the following steps.

1. First, it shows that draws of coefficients from the multivariate Normal centered at the estimated coefficients (for all validation periods) with variance-covariance matrix equal to Eq. 22 of the manuscript are arbitrarily close to (i.e., within a distance of $\varepsilon > 0$ from) the population-level coefficients with probability limiting to 1.
2. Then the proof shows that the difference in average prediction errors, calculated over random draws from the aforementioned multivariate Normal, will be arbitrarily close to the population-level difference in expected prediction errors with probability limiting to 1.
3. Finally, the proof concludes by showing that one event, *the difference in average prediction errors' being within a distance of ε from the population-level difference in expected prediction errors*, implies another event, *the most robust model in the population minimizes the maximum absolute difference in average prediction errors over draws of coefficients from the multivariate Normal*. Since Step 2 establishes that the former event occurs with probability limiting to 1, Step 3 implies, by logical implication, that the latter event must also occur with probability limiting to 1. Therefore, in our procedure, the posterior probability of the truly most robust model will converge in probability to 1.

To carry out the proof via the steps above, first let $\hat{\boldsymbol{\beta}}^*_{\mathcal{V}}$ denote a draw from $\mathcal{N}\left(\hat{\boldsymbol{\beta}}_{\mathcal{V}}, \widehat{\boldsymbol{\Sigma}_{\mathcal{V}}}\right)$ conditional on sample data and let $\Pr^*$ denote conditional probability given sample data. Then note that the weak law of large numbers (WLLN) implies that $\hat{\boldsymbol{\beta}}_{\mathcal{V}} \xrightarrow{p} \boldsymbol{\beta}_{\mathcal{V}}$, where $\boldsymbol{\beta}_{\mathcal{V}}$ is the population-level coefficients for the validation periods, and $\widehat{\boldsymbol{\Sigma}_{\mathcal{V}}} \xrightarrow{p} 0$ as $n \to \infty$. The continuous mapping theorem (CMT), implies that $\mathcal{N}\left(\hat{\boldsymbol{\beta}}_{\mathcal{V}}, \widehat{\boldsymbol{\Sigma}_{\mathcal{V}}}\right)$ converges in probability to

a constant, whereby the probability that any draw, $\hat{\boldsymbol{\beta}}^*_{\mathcal{V}}$, is equal to $\boldsymbol{\beta}_{\mathcal{V}}$ 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 $\varepsilon > 0$,

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

Turning to Step 2, to show convergence of the difference in average prediction errors to the population-level difference in expected prediction errors, we first establish convergence in probability of the regression prediction in a sample to its population-level analog. To do so, first write the average of the squared differences in predictions between $\hat{\boldsymbol{\beta}}^*_{f,v}$ and the population-level $\boldsymbol{\beta}_{f,v}$ for any (f, v) as

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

The Cauchy-Schwarz inequality implies that

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

The WLLN implies that the second factor, $\frac{1}{n_g}\sum_{i=1}^{n} \mathbb{1}\{G_i = g\}\boldsymbol{X}_{i,v}\boldsymbol{X}_{i,v}^\top$, limits in probability to $\mathrm{E}_{\mathcal{P}}\left[\boldsymbol{X}_v\boldsymbol{X}_v^\top \mid G = g\right]$, where the regularity conditions in Assumption 4 of the manuscript imply that $\mathrm{E}_{\mathcal{P}}\left[\boldsymbol{X}_v\boldsymbol{X}_v^\top \mid G = g\right] < \infty$. Consequently, since $\|\hat{\boldsymbol{\beta}}^*_{f,v} - \boldsymbol{\beta}_{f,v}\|^2 \xrightarrow{p} 0$, the CMT implies that

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

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

The CMT then implies that

$$\hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}^*_{f,v}\right) \xrightarrow{p^*} \delta_{f,v}, \tag{5}$$

where $\delta_{f,v}$ denotes the observable population-level differential prediction errors in validation period v under model specification $f \in \mathcal{F}$, as in Eq. 8 of the manuscript. That is, Eq. 5 states that, for all $\varepsilon > 0$,

$$\overset{*}{\Pr}\left(\left|\hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}^*_{f,v}\right) - \delta_{f,v}\right| \leq \varepsilon\right) \xrightarrow{p} 1, \tag{6}$$

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

Now turning to Step 3, define $\overline{v}_f$ as the validation period with the greatest population-level absolute difference in expected prediction errors under model f. Since Eq. 6 holds for all $\varepsilon > 0$, we can pick an $\varepsilon > 0$ that satisfies two conditions in Eq. 7 and Eq. 8: For all $f \in \mathcal{F}$,

$$|\delta_{(f,\overline{v}_f)}| - \varepsilon > |\delta_{(f,v)}| + \varepsilon \text{ for all } v \in \{\mathcal{V} \setminus \overline{v}_f\} \tag{7}$$

and

$$|\delta_{(f^\dagger,\overline{v}_{f^\dagger})}| - \varepsilon < |\delta_{(f,\overline{v}_f)}| + \varepsilon \text{ for all } f \in \{\mathcal{F} \setminus f^\dagger\}. \tag{8}$$

With $\varepsilon > 0$ satisfying Eq. 7 and Eq. 8, it follows that the event contained within the probability limit statement in Eq. 6, i.e.,

$$\left|\hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}^*_{f,v}\right) - \delta_{f,v}\right| \leq \varepsilon \text{ for all } (f, v) \in \mathcal{F} \times \mathcal{V}, \tag{9}$$

implies the event that

$$f^\dagger = \arg\min_{f \in \mathcal{F}} \max_{v \in \mathcal{V}} \hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}^*_{f,v}\right). \tag{10}$$

Hence, Eq. 6 implies that

$$\overset{*}{\Pr}\left(f^\dagger = \arg\min_{f \in \mathcal{F}} \max_{v \in \mathcal{V}} \hat{\delta}\left(\boldsymbol{D}_v, \hat{\boldsymbol{\beta}}^*_{f,v}\right)\right) \xrightarrow{p} 1, \tag{11}$$

thereby completing the proof. □

1.5. *Proof of Proposition 3.*

PROOF. First, note that

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

for all $(f,t) \in \mathcal{F} \times \mathcal{T}$, where $\mathcal{T} \coloneqq \{1,\ldots,T\}$. Then, by reasoning analogous to that in Eq. 7 of Lemma 1's proof and the CMT, it follows that

$$\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}_f, M\right) \xrightarrow{p} \delta_{f,T} \pm M \max_{v \in \mathcal{V}} |\delta_{f,v}|$$

for all $f \in \mathcal{F}$. Finally, Lemma 1 and the law of total probability imply that $\hat{p}_f \xrightarrow{p} 0$ as $n \to \infty$ for all $f \in \left\{\mathcal{F} \setminus f^\dagger\right\}$, which, along with another application of the CMT, implies that

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

thereby completing the proof. □

2. Existing models as special cases (or not). Our proofs each follow the steps sketched out below.

1. Use the design's identification assumptions to re-express the treated and comparison groups' untreated potential outcomes (in expectation) in the post-period, $\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right]$ and $\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right]$.
2. Write the prediction errors in treated and comparison groups (in expectation):
   a) First, use Assumption 1 of the manuscript to substitute untreated potential outcomes for any observed outcomes in the argument $\boldsymbol{X}_t$ to the prediction model, f.[1]
   b) Next, take expectation (with respect to the identification assumptions) of the prediction models in each group, $\mathrm{E}_{\mathcal{P}}\left[f\left(\boldsymbol{X}_T\right) \mid G=1\right]$ and $\mathrm{E}_{\mathcal{P}}\left[f\left(\boldsymbol{X}_T\right) \mid G=0\right]$.
   c) Finally, compute the differential prediction error (in expectation),
$$\mathrm{E}_{\mathcal{P}}\left[Y_T(0)-f\left(\boldsymbol{X}_T\right) \mid G=1\right]-\mathrm{E}_{\mathcal{P}}\left[Y_T(0)-f\left(\boldsymbol{X}_T\right) \mid G=0\right].$$
3. Show that this is equal to 0, thereby implying Assumption 2 in the manuscript and, consequently, the identified estimand in Eq. 5 of the manuscript.

2.1. *Difference-in-Differences.* If the prediction function is

$$f(\boldsymbol{X}_t)=Y_{t-1}, \tag{12}$$

then Assumption 2 of the manuscript will be true whenever parallel trends holds.

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

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

Next, using Assumption 1 of the manuscript, the expectations of the prediction model in Eq. 19 of the manuscript in each group are

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

Hence, the differential prediction error (in expectation) is

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

which is equal to 0 by parallel trends in Eq. 6 of the manuscript. Hence, Assumption 2 of the manuscript also holds.

2.2. *Two-way Fixed Effects.* If the prediction function is

$$f(\boldsymbol{X}_t)=\underset{\alpha_u}{\arg\min}\sum_{l=1}^{t-1}\left(Y_{u,l}-\alpha_u\right)^2, \tag{13}$$

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

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

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

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

The prediction model in Eq. 13 is simply each unit's average outcome in the pre-period,

$$f\left(\boldsymbol{X}_T\right) = \arg\min_{\alpha_u} \sum_{t=1}^{T-1} \left(Y_{u,t} - \alpha_u\right)^2 = \frac{1}{(T-1)} \sum_{t=1}^{T-1} Y_{u,t}, \tag{14}$$

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

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

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

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

which is equal to 0, thereby implying Assumption 2 of the manuscript.

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

On the other hand, other structural models require more careful thought about the appropriate prediction function. For example, with a unit- or group-specific linear time trend model, use of the prediction function in Eq. 13 would not imply equal expected prediction errors. However, using the OLS analog of the same linear time trend model would. 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 our 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.

2.3. *Sequential DID.* Sequential DID relies on a parallel-trends-in-trends assumption in which each group's average outcome in period $T-1$ plus the group's change in average outcomes from periods $T-2$ to $T-1$ is equal to each group's expected untreated potential outcome (Mora and Reggio, 2012, 2019; Egami and Yamauchi, 2023; Lee, 2016; Olden and Møen, 2022). We formally write parallel trends-in-trends as

$$\textbf{Parallel trends-in-trends} := \tag{15}$$
$$\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] - \left(\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right]\right) =$$
$$\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=1\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=1\right] - \left(\mathrm{E}_{\mathcal{P}}\left[Y_{T-1}(0) \mid G=0\right] - \mathrm{E}_{\mathcal{P}}\left[Y_{T-2}(0) \mid G=0\right]\right).$$

This can be generalized to K time-wise differences (see, e.g., Lee, 2016), but for simplicity, we focus on $K=2$.

If the prediction function is

$$f(\boldsymbol{X}_t) = Y_{t-1} + (Y_{t-1} - Y_{t-2}) \text{ for } t = 3, \ldots, T, \tag{16}$$

then Assumption 2 of the manuscript will be true whenever Eq. 15 holds.

First, parallel trends-in-trends in Eq. 15 implies that

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

Then the prediction function in Eq. 16 and Assumption 1 of the manuscript imply that

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

which implies that the expected prediction errors in each group are

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

Therefore, the difference in expected prediction errors is

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

which parallel trends-in-trends in Eq. 15 implies is equal to 0, thereby completing the proof.

2.4. *Unit- or group-specific time trends.* In contrast to methods that assume similar time dynamics in treated and comparison groups, comparative interrupted time series (CITS) methods explicitly model differential time trends in the two groups. A fully linear implementation measures changes in intercepts and slopes across the two groups, but a more flexible version of CITS measures period-by-period differences from an extrapolated linear trend in each individual or group (Riccio and Bloom, 2002; Bloom and Riccio, 2005).

Like two-way fixed effects (TWFE), this method assumes a parametric structural model for the untreated potential outcomes

$$Y_t(0) = \xi_u t + \gamma_t + \epsilon_{u,t}, \tag{17}$$

where ξ_u is the linear time slope of the u^{th} unit and $\mathrm{E}[\epsilon_{u,t} \mid \xi_u, G_u] = 0$ for all $u = 1, \ldots, U$ and $t = 1, \ldots, T$. With this model, we can show that there exists a prediction function such that when Eq. 17 holds, equal expected prediction errors holds also.

If the prediction function is

$$f(\boldsymbol{X}_t) = \hat{\xi}_u t \text{ where } \hat{\xi}_u = \arg\min_{\xi_u} \sum_{l=1}^{t-1} (Y_{u,l} - \xi_u l)^2, \tag{18}$$

then Assumption 2 of the manuscript will hold whenever the structural model in Eq. 17 is true.

First, the structural model in Eq. 17 implies that

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

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

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

which, from the linear time trend model in Eq. 17, can be expressed as

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

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

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

Then, due to the structural model in Eq. 17 and since all $t = 1, \ldots, T$ are fixed constants, it follows that

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

Finally, the model in Eq. 17 implies that the difference in expected prediction errors is equal to 0:

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

$$= \mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u = 1\right] + \gamma_T - \mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u = 1\right] - \left(\frac{\sum_{t=1}^{T-1} t\gamma_t}{\sum_{t=1}^{T-1} t^2} \right) T$$

$$- \left(\mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u = 0\right] + \gamma_T - \mathrm{E}_{\mathcal{P}}\left[\xi_u T \mid G_u = 0\right] - \left(\frac{\sum_{t=1}^{T-1} t\gamma_t}{\sum_{t=1}^{T-1} t^2} \right) T \right)$$

$$= 0,$$

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

A complementary question is, as a reviewer wondered, "whether it is possible for non-parametric identification to not hold, but the proposed assumption to hold." Indeed, we can use the structural model above to illustrate such a case. Suppose the true structural model is CITS as implemented by Bloom and Riccio (2005), which is equivalent to a flexible difference-in-differences (DID) specification that uses time fixed effects and group-specific linear trends (Fry and Hatfield, 2021). The nonparametric parallel trends assumption in Eq. 6 of the manuscript, clearly will not hold because the truth is differential trends in the two groups. However, Assumption 2 of the manuscript will still hold for the prediction function that incorporates differential trends in the two groups into its predictions.

2.5. *Lagged dependent variable model.* Lagged dependent variable (LDV) models incorporate across-time dependence of outcomes within units. In political science, authors have debated the merits of autoregressive distributed lag (ADL) models that include lags of both outcomes and treatments (Beck and Katz, 2011). The relationship between lags of treatment and lags of the outcome is complicated by a classic observation about the bias of unit fixed effects in autoregressive models (Nickell, 1981). Thus, one recent comparison across model specifications argued that LDV models should use first differences (Griffin et al., 2021). Other authors have argued for a specification that includes the full vector of pre-treatment outcomes (like a regression analog of synthetic controls) (O'Neill et al., 2016). Other authors have emphasized the causal assumptions, including whether past treatments can affect current outcomes and whether past outcomes can affect current treatment (Imai and Kim, 2019), the problem of conditioning on post-treatment outcomes (Blackwell and Glynn, 2018), and the relationship between the causal assumptions of DID and methods that, like LDV, condition on past outcomes (Ding and Li, 2019).

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

$$Y_t(0) = \gamma_t + \lambda Y_{t-1}(0) + \epsilon_t, \tag{19}$$

where λ is a parameter that controls the strength of the dependence and $\mathrm{E}_{\mathcal{P}}[\epsilon_t] = 0$ for $t = 1, \ldots, T$.[2] Notice that this resembles the two-way fixed effects model of Eq. 7 of the manuscript, but instead of unit-level (time-invariant) fixed effects, it includes unit-level dependence on past outcomes. Then we assume a form of exogeneity conditional on past outcomes,

$$\textbf{Exogeneity conditional on past outcomes} := \mathrm{E}_{\mathcal{P}}\left[\epsilon_t \mid Y_{t-1}, Y_{t-2}, \ldots, Y_1, G\right] = 0\,. \tag{20}$$

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

There exists a prediction function such that when Eqs. 19 and Eq. 20 both hold, so does equal expected prediction errors.

If the prediction function is

$$f(\boldsymbol{X}_t) = \hat{\lambda} Y_{t-1} \text{ where } \hat{\lambda} = \arg\min_{\lambda} \sum_{l=2}^{t-1} \left(\tilde{Y}_l - \lambda \tilde{Y}_{l-1} \right)^2, \tag{21}$$

where $\tilde{Y}_t \coloneqq Y_t - \mathrm{E}_{\mathcal{P}}[Y_t]$ for all $t = 1, \ldots, T$, then Assumption 2 of the manuscript will hold whenever the outcome model in Eq. 19 and exogeneity in Eq. 20 are true.

First, note that the structural model in Eq. 19 implies that

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

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

$$\hat{\lambda} = \frac{\sum_{t=2}^{T-1} \tilde{Y}_{t-1} \tilde{Y}_t}{\sum_{t=2}^{T-1} \tilde{Y}_{t-1}^2}. \tag{22}$$

Given the equivalent representation of the LDV model in Eq. 19 in which outcomes, predictors and the error term are centered by their means across units for each time period (Kropko and Kubinec, 2020), the solution to the empirical risk minimization problem in Eq. 22 can be expressed as

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

It follows that the prediction in period T is

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

which exogeneity in Eq. 20 then implies has expectations in treated and control groups equal to

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

The LDV model in Eq. 19 further implies that the expected prediction errors in treated and comparison groups are

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

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

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

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

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

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

$$Y_{1,T} - \frac{1}{U-1}\sum_{u=2}^{U} w_u Y_{u,T}, \tag{23}$$

where $u = 2, \ldots, U$ are the comparison units.

What is the identifying assumption of synthetic controls? As far as we can tell, synthetic controls began with an estimation method and then invoked a structural model (interactive fixed effects) that would justify that estimator. However, for our purposes, write the the difference in conditional expectations of the first two terms of Eq. 23 as

$$\mathrm{E}_{\mathcal{P}}\left[Y_{1,T} \mid G_1 = 1\right] - \mathrm{E}_{\mathcal{P}}\left[\frac{1}{U-1}\sum_{u=2}^{U} w_u Y_{u,T} \mid G_u = 0\right], \tag{24}$$

which would be equal to the ATT in Eq. 1 of the manuscript whenever

$$\mathrm{E}_{\mathcal{P}}\left[Y_T(0) \mid G = 1\right] = \mathrm{E}_{\mathcal{P}}\left[\frac{1}{U-1}\sum_{u=2}^{U} w_u Y_{u,T} \mid G_u = 0\right]. \tag{25}$$

The identification condition in Eq. 25 illustrates that synthetic controls is outside the scope of our "predict, correct" paradigm. Neither the treated nor comparison group draws on past data *within* groups to predict future untreated outcomes. This makes sense because the synthetic control method involves only a treated-vs-comparison contrast, not a pre-vs-post contrast. The pre-period's only contribution in synthetic controls is to inform the weights. When we try to fit synthetic controls into our "predict, correct" paradigm, we find that it involves only the correction step without the prediction step.

Nevertheless, synthetic control weights may still be useful if we believe that weighting by similarity on pre-period outcomes helps us select a more suitable comparison group. We can weight as a pre-processing step, then apply our methods to the weighted combination of comparison units. Others have combined DID and synthetic controls (e.g., Arkhangelsky et al., 2021), and we envision this to be a fruitful topic for further research.

2.7. *Interactive fixed effects.* We now use an interactive fixed effects (IFE) structural model to demonstrate an example (inspired by a reviewer) in which a parametric structural model holds, but, given a specific prediction function, equal expected prediction errors in Assumption 2 of the manuscript does not. Here we show that the prediction model (corresponding to TWFE) in Eq. 20 of the manuscript does not imply equal expected prediction errors when the structural model is that of interactive fixed effects — an unsurprising result given that the interactive fixed effect model implies that time shocks differ between treated and comparison groups. That said, our argument below does not rule out the possibility that another prediction function could be found that does imply equal expected prediction errors (perhaps drawing upon Liu, Wang and Xu, 2024); however, it is unclear whether such an

appropriate prediction function would conduct the "predict" step from pre-period data within groups, in accordance with controlled pre-post designs to which our argument pertains.

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

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

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

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

Consider the prediction function in Eq. 20 of the manuscript, which is simply each unit's average outcome prior to t:

$$\arg\min_{\alpha_u} \sum_{l=1}^{t-1} \left(Y_{u,l} - \alpha_u\right)^2 = \frac{1}{(t-1)} \sum_{l=1}^{t-1} Y_{u,l}. \tag{27}$$

With this prediction function, the prediction in period T is

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

which, by the consistency assumption in Eq. 2 of the manuscript, is

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

The IFE model in Eq. 26 implies that the expectations of the predictions in the treated and control groups are

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

and

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

The IFE model in Eq. 26 also implies the expected prediction errors in each group are

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

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

and

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

Taking the difference in expected prediction errors yields

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

which is not necessarily equal to 0.

3. Joint Variance-Covariance matrix for Bayesian model selection. The estimated (cluster-robust) variance-covariance matrix for all coefficients across all models and all validation periods is

$$\widehat{\boldsymbol{\Sigma}}_{\mathcal{V}} \coloneqq \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_{(f_1,v_1),(f_1,v_1)} & \cdots & \widehat{\boldsymbol{\Sigma}}_{(f_1,v_1),(f_{|\mathcal{F}|},V)} \\ \vdots & \ddots & \vdots \\ \widehat{\boldsymbol{\Sigma}}_{(f_{|\mathcal{F}|},V),(f_1,v_1)} & \cdots & \widehat{\boldsymbol{\Sigma}}_{(f_{|\mathcal{F}|},v_{|\mathcal{V}|}),(f_{|\mathcal{F}|},V)} \end{bmatrix}, \tag{28}$$

where $\widehat{\boldsymbol{\Sigma}}_{(f,v),(f',v')}$ is the cluster-robust variance-covariance matrix between any two model-year pairs from $\mathcal{F} \times \mathcal{V}$ and the elements in the set of candidate models, $\mathcal{F}$, are denoted by $f_1, f_2, \ldots, f_{|\mathcal{F}|}$.

In accordance with the usual sandwich formula, $\widehat{\boldsymbol{\Sigma}}_{(f,v),(f',v')}$ can be decomposed into its "bread" and "meat" components. The "bread" matrix for any $(f, v) \in \mathcal{F} \times \mathcal{V}$ is

$$\boldsymbol{B}_{(f,v)} \coloneqq \left(\boldsymbol{X}_{f,<v}^{\top} \boldsymbol{X}_{f,<v} \right)^{-1} \tag{29}$$

in which $\boldsymbol{X}_{f,<v}$ is the $n(v-1) \times K_f$ model matrix for f in periods before v, where K_f is model f's number of coefficients. For the "meat" component, first let $\boldsymbol{e}_{f,i,<v}$ denote the $(v-1) \times 1$ vector of unit i's prediction errors (residuals) under model f for periods before v. Also let $\boldsymbol{X}_{f,i,<v}$ be the $(v-1) \times K_f$ model matrix under model f for unit i in periods before v. Now we can write the "meat" matrix between any two model-year pairs in $\mathcal{F} \times \mathcal{V}$ (clustered at the unit level) as

$$\mathbf{M}_{(f,v),(f',v')} \coloneqq \sum_{i=1}^{n} \left(\boldsymbol{X}_{f,i,<v}^{\top} \boldsymbol{e}_{f,i,<v} \right) \left(\boldsymbol{e}_{f',i,<v'}^{\top} \boldsymbol{X}_{f',i,<v'} \right). \tag{30}$$

Putting together the "breads" and the "meat" for any two elements from $\mathcal{F} \times \mathcal{V}$ and then multiplying by the usual small sample adjustment factor (originally derived in Hansen, 2007) results in

$$\widehat{\boldsymbol{\Sigma}}_{(f,v),(f',v')} \coloneqq \left(\frac{n}{n-1} \right) \boldsymbol{B}_{(f,v)} \mathbf{M}_{(f,v),(f',v')} \boldsymbol{B}_{(f',v')}. \tag{31}$$

This estimated (cluster robust) variance-covariance for any two elements from $\mathcal{F} \times \mathcal{V}$ in Eq. 31 can be equivalently expressed as

$$\left(\frac{1}{n-1} \right) \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_{f,i,<v}^{\top} \boldsymbol{X}_{f,i,<v} \right)^{-1} \left(\frac{1}{n} \mathbf{M}_{(f,v),(f',v')} \right) \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_{f',i,<v'}^{\top} \boldsymbol{X}_{f',i,<v'} \right)^{-1},$$

from which it is straightforward to see that Eq. 31 converges in probability to 0 as n increases indefinitely, as does the overall (cluster robust) variance-covariance in Eq. 28.

4. Conceptual diagram of estimation process. Fig. 1 provides a conceptual diagram of the overall estimation process. All of the mathematical quantities in Fig. 1 are defined in the manuscript. However, to reiterate, the index $s = 1, \ldots, S$ runs over the posterior draws from $\mathcal{N}\left(\hat{\boldsymbol{\beta}}_{\mathcal{V}}, \widehat{\boldsymbol{\Sigma}_{\mathcal{V}}}\right)$. In addition, $|\hat{\delta}_f^{\ddagger(s)}|$ denotes the largest absolute differential prediction error for model f over all validation periods, $\mathcal{V}$, where $V \coloneqq \max \mathcal{V}$, under the sth draw from $\mathcal{N}\left(\hat{\boldsymbol{\beta}}_{\mathcal{V}}, \widehat{\boldsymbol{\Sigma}_{\mathcal{V}}}\right)$. The optimal model under the sth draw is denoted by $f^{\dagger(s)}$. The elements in the set of candidate models, $\mathcal{F}$, are denoted by $f_1, f_2, \ldots, f_{|\mathcal{F}|}$. All other quantities — namely, $\hat{\Delta}(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M)$, $\widehat{\mathrm{E}}_{\mathcal{F}\,|\,\boldsymbol{D}}\left[\hat{\Delta}\left(\boldsymbol{D}, \hat{\boldsymbol{\beta}}, M\right)\right]$ and $\hat{p}_f$ — are as defined in Eqs. 14, 15 and 16 of the manuscript.

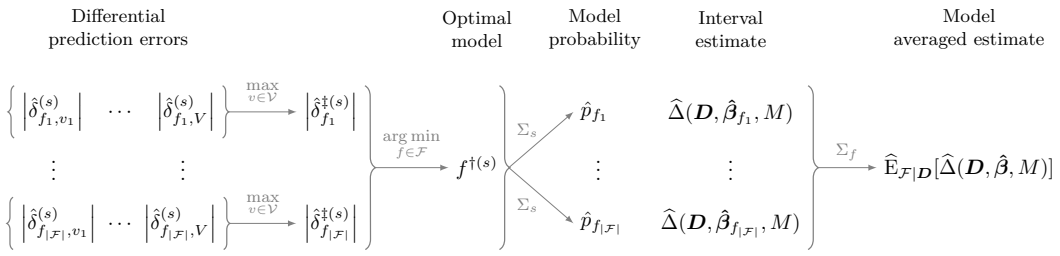

FIG 1. *Averaged Prediction Models (APM) estimation process*

5. Model implementation in the applied analysis. Table 1 below lists all model specifications in our applied analysis.

TABLE 1

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

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

6. Simulation Studies. We conduct several simulation studies to assess the performance of our Bayesian model averaged (BMA) estimator. We use the same simulation setup as Schell, Griffin and Morral (2018). This setup is especially compelling in a setting related to gun policy, and for that reason has been adopted in closely related simulation studies (Antonelli and Beck, 2023).

The simulation setup from Schell, Griffin and Morral (2018) consists of crude death rates in all 50 states in each year from 1979 to 2014. We focus on years 1994 to 2008 and suppose that 2008 is the only post-treatment year. Akin to our application in Sec. 5 of the manuscript, we let the years 1994 to 1998 serve as training years and let 1999 to 2007 serve as validation years. We randomly select 5 states to serve as "treated," which begins in 2008. The remaining 45 states are the "comparison" states.

We consider a class of 5 candidate models: (1) baseline mean, $Y_t \sim \beta_0$, (2) LDV, i.e., AR(1), $Y_t \sim \beta_0 + \beta_2 Y_{t-1}$, (3) baseline mean (first diff), $Y_t - Y_{t-1} \sim \beta_0$, (4) linear trend, $Y_t \sim \beta_0 + \beta_1 t$ and (5) linear time trend (first diff), $Y_t - Y_{t-1} \sim \beta_0 + \beta_1 t$. Performing our procedure on the population of 5 treated stated and 45 comparison states shows that model (2), the AR(1) model, minimizes our sensitivity criterion (i.e., the worst-case absolute prediction error in the pre-treatment validation periods).

To conduct our simulations, we treat the 5 treated states and 45 control states as the population of interest and consider properties of our BMA estimator over 1,000 random draws with replacement of states from this population. For each draw, we sample with replacement a fixed number from the distribution of treated states and a fixed number from the distribution of control states. This sampling corresponds to the standard assumption of independent and identically distributed (i.i.d.) random sampling of units (in this case states) within groups.

Over each realization of sample data, we record the posterior probability that each of the 5 candidate models is most robust. We also record the BMA estimates of the ATT and its lower and upper bounds (with $M = 1$). For each realization of sample data, we also record the estimated variance for the ATT estimator, as well as for the estimators of the ATT's lower and upper bounds. For each of these three targets, we construct 95% confidence intervals via a Normal approximation in which the lower bound is the BMA estimate minus 1.96 multiplied by the square root of the estimated variance. The upper bound of the 95% confidence interval is constructed analogously. A confidence interval covers the target if it brackets the population-level quantity for period T under the truly optimal model in the population. The bias of the BMA estimator also refers to this target.

We conduct our simulations under an increasing number of sampled units, holding the ratio of treated to control units fixed. We begin with 1 treated unit and 8 controls, which mirrors the setting of our application in Sec. 5 of the manuscript. We then increase the number of treated units to 3, 15, 35, 50, 500 and 2,000 with 24, 120, 280, 400, 4,000 and 16,000 control units, respectively.

Fig. 2 below shows the proportion of the 1,000 simulations in which the optimal model in the population, the AR(1) model, has the greatest posterior probability.

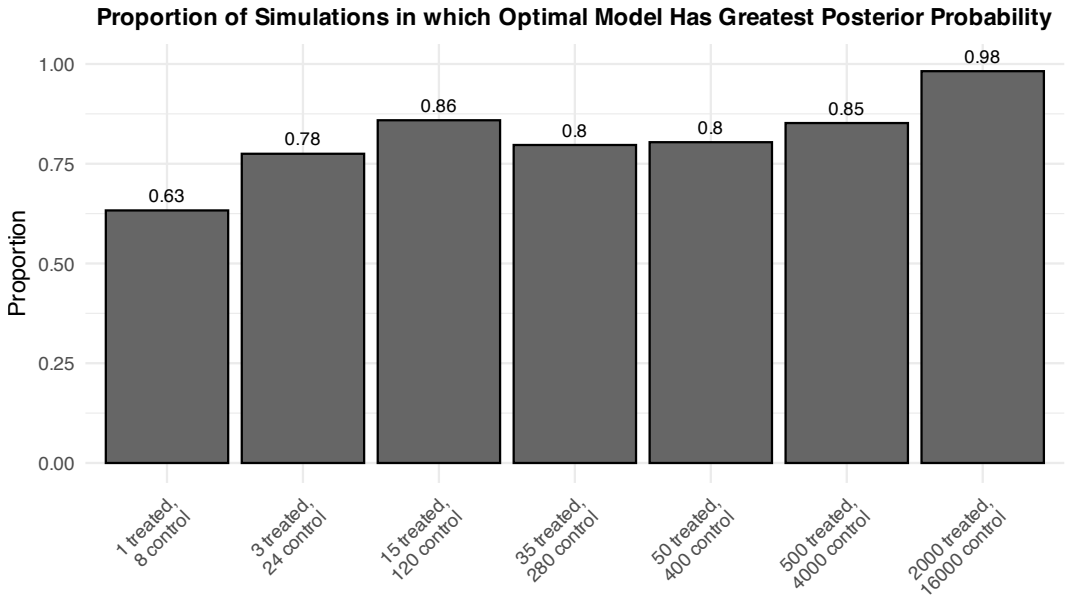

FIG 2. *The proportion of 1,000 simulation replications in which the optimal model in the population receives the highest posterior probability. Each bar corresponds to a different sample size configuration, and values are labeled above each bar.*

The truly optimal model receives the greatest posterior probability in a majority of the 1,000 simulations, even when the sample size is only 9 states. For the largest sample size, 2,000 treated states and 16,000 control states, the truly optimal model receives the greatest posterior probability in effectively all of the 1,000 simulations.

Fig. 3, which paints a similar picture, illustrates the expected posterior probability (over all 1,000 simulations) for each sample size.

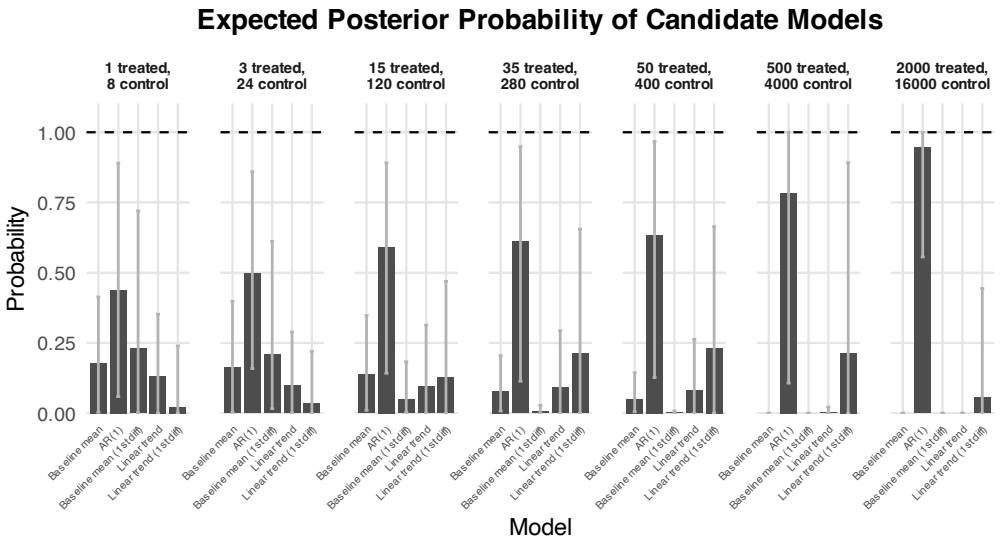

FIG 3. *Expected posterior probabilities of candidate models across simulation settings. Each panel corresponds to a different sample size configuration. Bars show the mean posterior probability assigned to each model across 1,000 simulations, and error bars indicate the* 2.5*th and* 97.5*th percentiles of posterior probabilities across simulations. Horizontal dashed lines mark the maximum possible posterior probability (*1*).*

As Fig. 3 shows, the truly optimal model tends to receive the highest posterior probability across simulation settings. While there is some variability in this probability — particularly in smaller samples — the distinction between the optimal model and others becomes more pronounced as sample size increases. In the largest sample, the optimal model's expected posterior probability reaches 0.95, with 95% simulation intervals that no longer overlap with those of competing models. Hence, Fig. 3 demonstrates the intuition of Lemma 1 in which the truly optimal model's posterior probability converges in probability to 1 as the sample size increases.

Nevertheless, in this simulation setting, the high variance in the optimal model's posterior probability — including at the largest sample size — suggests that the rate of convergence described in Lemma 1 may be slow. At the largest sample size, the 95% simulation interval for the posterior probability ranges from 0.62 to 1. Hence, quite large samples may be required for the posterior probability to concentrate tightly near 1 with high probability.

We now report the absolute percent bias of the BMA estimator (Fig. 4) followed by the ratio of the expected estimated variance to the variance across simulations (Fig. 5).

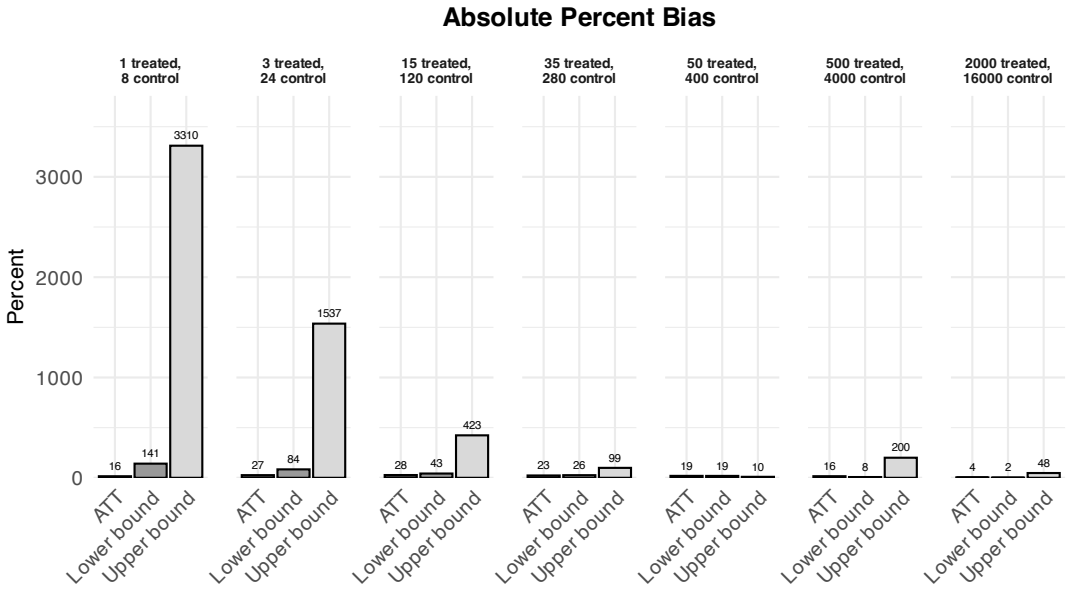

FIG 4. *The absolute percent bias of the estimator of the ATT, lower bound, and upper bound, relative to their respective population-level values for period T under the truly optimal model in the population. Each bias value reflects the absolute deviation of the estimator from its target, expressed as a percentage of the absolute value of the population ATT. Numeric values are shown above each bar, and each panel corresponds to a different sample size configuration.*

As we would expect, bias is substantial in small samples. However, as sample size increases, the bias of the estimators for both the ATT and its bounds decreases considerably. In larger samples, the estimators for all three targets exhibit relatively low bias, indicating more reliable inference.

The target ATT and its lower and upper bounds are approximately -0.4, -0.84, and 0.03, respectively. Because the upper bound is close to zero, even a small absolute bias appears large when expressed as a percentage of the target. Consequently, the absolute percent bias for the upper bound remains greater than the absolute percent bias for the ATT and lower bound in large samples. This difference in percent bias exists even though the absolute bias is nearly identical across all three targets in large samples.

Fig. 5 shifts focus from the bias of the BMA estimator for the ATT and its bounds to the accuracy of the procedure used to estimate the BMA estimator's variance. The figure presents the ratio of the expected estimated variance to the actual variances (across simulations) of the BMA estimators for the ATT, its lower bound, and its upper bound.

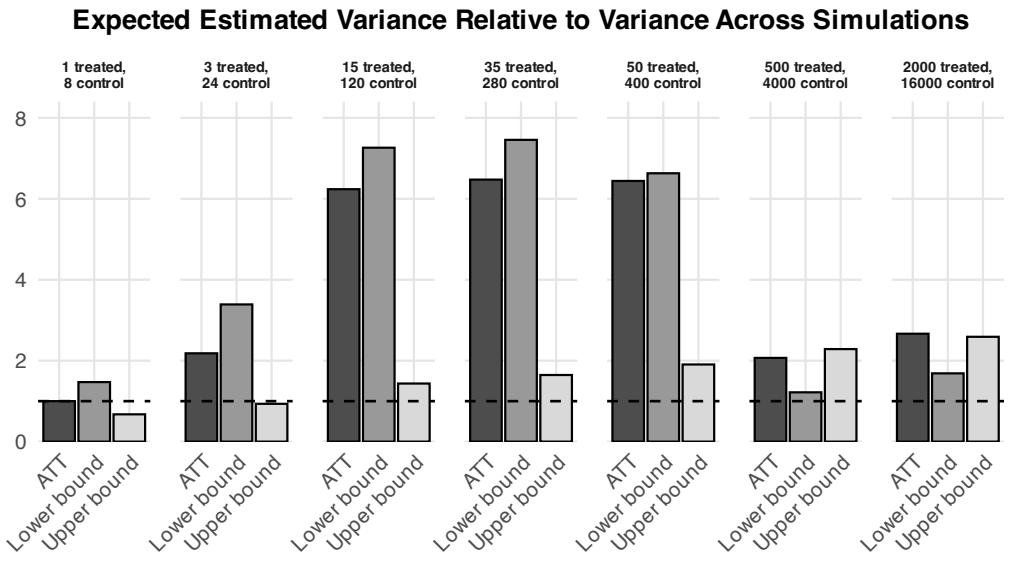

FIG 5. *The ratio of the expected estimated variance to the empirical variance across simulation replications for the ATT, its lower bound, and its upper bound. Each panel corresponds to a different sample size configuration. The dashed horizontal line at* 1 *denotes exact agreement between the expected estimated variance and the variance across* $1,000$ *simulations. Each panel corresponds to a different sample size configuration.*

In small to moderate sample sizes, the estimated variances systematically overstate the actual variability, often by substantial margins. As sample size increases, the expected estimated variances become more closely aligned with the empirical variances across simulations, but the estimated variances still exceed the actual variances, in expectation. For example, in the largest sample size (2,000 treated, 16,000 control), the expected estimated variances remain greater than the actual variances across simulations. This overestimation is consistent with the conservatism noted by Antonelli, Papadogeorgou and Dominici (2022, p. 103).

Finally, Fig. 6 below illustrates the coverage of 95% confidence bounds for the ATT, its lower bound, and its upper bound across a range of sample size configurations. In the smallest samples, coverage is substantially below the nominal level for all components, particularly for the upper bound. As sample size increases, coverage improves for all three quantities, and by approximately 15 treated and 120 control units, all 95% confidence intervals are close to nominal performance.

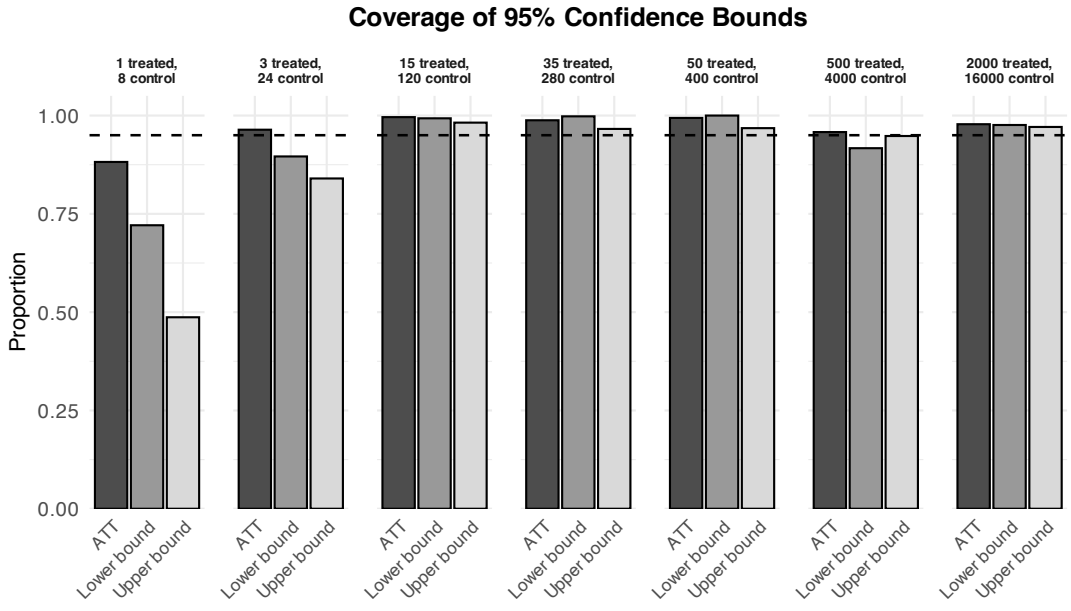

FIG 6. *The coverage of the 95% confidence interval for the ATT, its lower bound, and its upper bound across 1,000 simulation replications. Each panel corresponds to a different sample size configuration. The dashed horizontal lines mark the nominal 95% coverage level. Bars represent the proportion of simulations in which the true value lies within the estimated confidence bounds for each quantity.*

As Fig. 6 shows, the 95% confidence bounds achieve coverage at or above the nominal level across sample sizes ranging from 15 treated and 120 control units to 50 treated and 400 control units. However, coverage dips slightly below the nominal 95% level at a sample size of 500 treated and 4,000 control units, before recovering and exceeding the nominal level in the largest sample size of 2,000 treated and 16,000 control units. This pattern underscores that coverage need not increase monotonically with sample size, but may dip before stabilizing at or above the nominal level.

The explanation for this pattern is straightforward. The population-level values of the ATTs lower bound under the two leading models are 0.84 for the AR(1) model and 1.18 for the linear trend (first difference) model. In simulations where the truly optimal model, AR(1), receives low posterior probability (e.g., below 0.25), the BMA estimate of the ATTs lower bound is pulled downward due to greater weight being placed on the linear trend (first difference) model, which has a more negative lower bound.

This downward bias of the BMA estimator when the optimal model's posterior weight is low holds across all sample sizes. However, at smaller sample sizes — such as 50 treated and 400 control units — this bias is offset by the estimator's larger overall variance, which results in wider confidence intervals and helps maintain nominal coverage. As the sample size increases (e.g., to 500 treated and 4,000 control units), the variance of the estimator decreases, primarily due to reduced sampling variability rather than decreased uncertainty about which model is optimal. As a result, there are still enough cases in which the optimal model receives too little posterior weight. In those cases, the BMA estimator remains biased downward, but the narrower confidence intervals are no longer wide enough to compensate, leading to undercoverage.

Restoring coverage to at least the nominal level requires a larger sample size — such as 2,000 treated and 16,000 control units — so that uncertainty about which model is optimal is sufficiently reduced. At this larger sample size, the same issue seen at 500 treated and 4,000

control units still holds under simulations in which the optimal model receives low posterior probability: The BMA estimate is pulled downward, and the resulting confidence interval remains too narrow to include the true lower bound under the optimal model. However, because cases in which the optimal model receives a low posterior weight are rare at this large sample size, the overall coverage is no longer below the nominal level.

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.

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.

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.

BECK, N. and KATZ, J. N. (2011). Modeling Dynamics in Time-Series-Cross-Section Political Economy Data. *Annual Review of Political Science* **14** 331-352.

BLACKWELL, M. and GLYNN, A. N. (2018). How to Make Causal Inferences with Time-Series Cross- Sectional Data under Selection on Observables. *The American Political Science Review* **112** 1067–1082.

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.

DING, P. and LI, F. (2019). A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment. *Political Analysis* **27** 605–615.

EGAMI, N. and YAMAUCHI, S. (2023). Using Multiple Pre-treatment Periods to Improve Difference-in-Differences and Staggered Adoption Designs. *Political Analysis* **31** 195-212.

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.

GRIFFIN, B. A., SCHULER, M. S., STUART, E. A., PATRICK, S., MCNEER, E., SMART, R., POWELL, D., STEIN, B. D., SCHELL, T. L. and PACULA, R. L. (2021). Moving Beyond the Classic Difference-in-Differences Model: A Simulation Study Comparing Statistical Methods for Estimating Effectiveness of State-level Policies. *BMC Medical Research Methodology* **21**.

HANSEN, C. B. (2007). Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T Is Large. *Journal of Econometrics* **141** 597-620.

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* **17** 2139-2164.

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.

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

LEE, M.-J. (2016). Generalized Difference in Differences With Panel Data and Least Squares Estimator. *Sociological Methods & Research* **45** 134–157.

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

MORA, R. and REGGIO, I. (2012). Treatment Effect Identification Using Alternative Parallel Assumptions Working Paper, Economic Series (48) No. 12–33, Universidad Carlos III, Getafe, Spain.

MORA, R. and REGGIO, I. (2019). Alternative Diff-in-Diffs Estimators with Several Pretreatment Periods. *Econometric Reviews* **38** 465–486.

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

OLDEN, A. and MØEN, J. (2022). The Triple Difference Estimator. *The Econometrics Journal* **25** 531-553.

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.

RICCIO, J. A. and BLOOM, H. S. (2002). Extending the Reach of Randomized Social Experiments: New Directions in Evaluations of American Welfare-to-Work and Employment Initiatives. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **165** 13–30.

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

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.

SCHELL, T. L., GRIFFIN, B. A. and MORRAL, A. R. (2018). *Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths: A Simulation Study*. RAND Corporation, Santa Monica, CA.