Submitted to the Annals of Applied Statistics

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

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

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

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

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

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

In this paper, we apply controlled pre-post designs to study how homicide rates changed after Missouri repealed its permit-to-purchase (PTP) law in 2007. The law, in place since 1921, had required people purchasing handguns from private sellers to obtain a license that verified the purchaser had passed a background check. We compare changes in Missouri's homicide rate to changes in 8 bordering states that did not repeal their PTP laws (see Figure 1).

To choose among the various controlled pre-post designs, conventional wisdom holds that we should choose the one that relies on the most plausible assumptions (Roth and Sant'Anna,

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

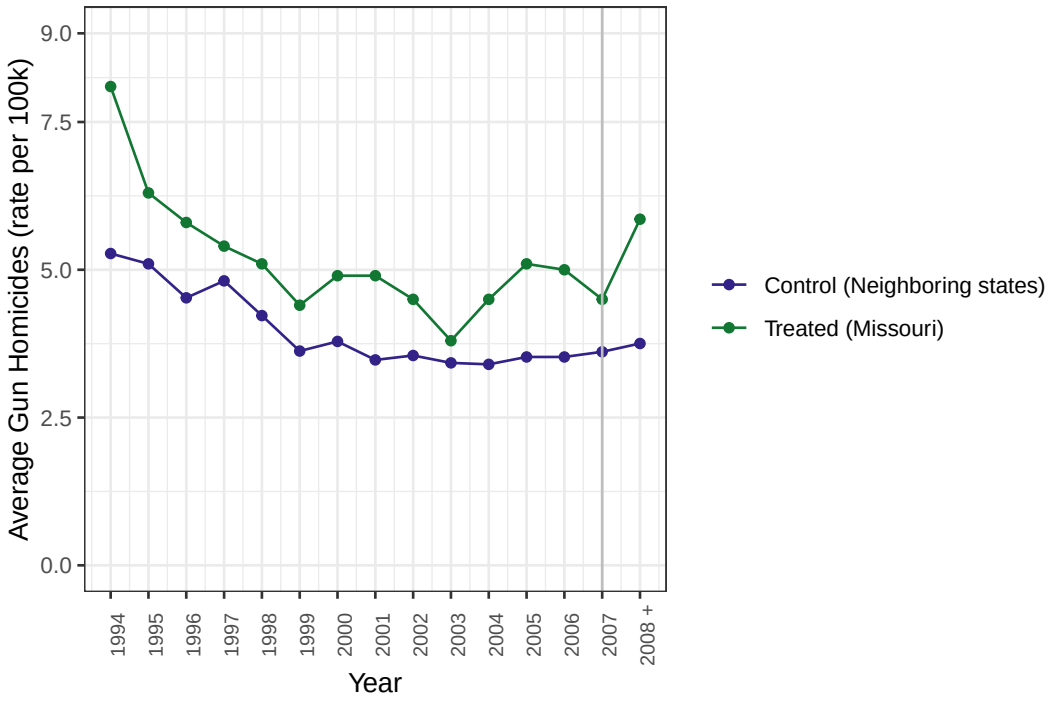

FIG 1. *Average gun homicides (rate per 100,000) before and after the 2007 permit-to-purchase repeal in Missouri (treated state) and control states (Arkansas, Illinois, Iowa, Kansas, Kentucky, Nebraska, Oklahoma, and Tennessee)*

2023; Lopez Bernal, Soumerai and Gasparrini, 2018; Ryan, Burgess and Dimick, 2015; Kahn-Lang and Lang, 2020). However, reasonable people may disagree about which model is most plausible or "correct". It is impossible to establish the correctness of *any* set of causal assumptions (and here, modeling assumptions amount to causal assumptions), so researchers tend to use methods that are popular in their disciplines. For instance, CITS dominates in education policy research, while DID is more popular in health policy research (Fry and Hatfield, 2021).

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

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

In this paper, we move away from the question of model plausibility or correctness and focus on another criterion: robustness. Our method proceeds in three steps: predict, correct, select. First, using data from the period before the policy change, we train a model that *predicts* untreated outcomes. Then, to account for unpredictable shocks, we use the comparison group's prediction errors to *correct* the treated group's predictions after the policy change. Third, using a validation set of pre-policy data, we *select* among competing models to maximize robustness. Finally, using the corrected predictions from our selected model, we identify

our causal target estimand by assuming that the expected prediction errors would be equal in treated and comparison groups, absent the policy change.

This identification strategy accommodates a wide variety of prediction models and expands the model space of controlled pre-post comparisons without requiring bespoke identifying assumptions for each model. Moreover, we show that many familiar "brand name" designs are special cases. For instance, difference-in-differences is a special case when the prediction model is simply the pre-period group mean. Most important, our procedure recasts the choice among causal assumptions (competing on plausibility) as a choice among prediction models (competing on robustness).

What do we mean by robustness? We build on recent work by Manski and Pepper (2018); Rambachan and Roth (2023) who formalize an idea implicit in tests of parallel trends in the pre-period (Granger, 1969; Angrist and Pischke, 2008; Roth, 2022; Egami and Yamauchi, 2022). The idea is that departures from the assumed causal model in the pre-period inform us about violations in the post-period. Since the true relationship between untreated outcomes in the two periods is unknown, sensitivity analyses begin with departures observed in the pre-period, apply these to the post-period, then dial up the magnitude to observe the impact of increasingly severe violations. A sensitive procedure's conclusions will be undermined before a robust one's.

In the rest of this paper, we elaborate on our approach to controlled pre-post designs applied to gun policy evaluation. Section 2 details our general identification strategy and establishes that two popular designs (DID and two-way fixed effects (TWFE) models) are special cases of our framework. Then in Section 3, we introduce a sensitivity analysis framework that motivates our model selection procedure. We provide a data-driven algorithm to select a model that maximizes robustness in Section 4 We implement our methods to estimate the effect of Missouri's PTP repeal on homicide in Section 5. Finally, we conclude in Section 6 and point to open questions for future research.

2. General identification strategy. First, some notation. Suppose individual units, indexed by $i = 1, \ldots, n$, belong to either a treated ($G_i = 1$) or comparison ($G_i = 0$) group. Let there be $t = 1, \ldots, T$ periods in which T is the only post-period. That is, between periods $T-1$ and T, all members of the treated group are exposed to treatment and all members of the comparison group are not. Let the treatment indicator be $D_{i,t} = G_i \mathbb{1}\{t = T\}$, where $\mathbb{1}\{\cdot\}$ is the indicator function that equals 1 if its argument is true and 0 if not. For units in the treated group, $D_{i,t} = 0$ in all $t < T$, and $D_{i,T} = 1$ in period T. For units from the comparison group, $D_{i,t} = 0$ in all periods.

We use potential outcomes to define our causal target. Each unit has potential outcomes under both treatment $Y_{i,t}(1)$ and no treatment $Y_{i,t}(0)$. Our causal estimand is a function of these treated and untreated potential outcomes. Specifically, we target the average effect of treatment on the treated (ATT):

$$\text{ATT} := \mathrm{E}\left[Y_{i,T}(1) - Y_{i,T}(0) \mid G_i = 1\right]. \tag{1}$$

To put this in terms we can observe (estimate), we make assumptions about how potential outcomes relate to observable quantities. Each observed outcome, $Y_{i,t}$, is a realization of either the treated or untreated potential outcomes, and we assume consistency in these realizations.

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

$$Y_{i,t} = D_{i,t} Y_{i,t}(1) + (1 - D_{i,t})\, Y_{i,t}(0).$$

Assumption 1 ensures that each unit's observed outcome is the potential outcome corresponding to the observed treatment condition applied to that unit at that time. This rules out two problematic scenarios. First, it rules out treatment anticipation, in which treated units begin manifesting treated outcomes before treatment begins. Second, it rules out spillovers, in which untreated units manifest treated potential outcomes despite not being directly exposed to treatment.

With Assumption 1 and the linearity of expectations, we re-write the ATT as

$$\text{ATT} = \text{E}\left[Y_{i,T} \mid G_i = 1\right] - \text{E}\left[Y_{i,T}(0) \mid G_i = 1\right], \tag{2}$$

where we have replaced treated potential outcomes with observed outcomes in the first term, since treated units' treated potential outcomes are observed in the post-period. It remains to replace the second term with something observable, since we cannot observe treated units' untreated potential outcomes in the post-period.

We first define a prediction model that *predicts* untreated potential outcomes using observable quantities measured (or fixed) prior to the period t for which we are predicting. These observable quantities could include, for example, the time index, previous outcomes, and previous covariates.[1] We collect these quantities into a vector $\boldsymbol{X}_{i,t}$. The model $\widehat{Y}_{i,t} = f\left(\boldsymbol{X}_{i,t}, \boldsymbol{\beta}\right)$ predicts the untreated outcome for unit i in period t using the observables and parameters $\boldsymbol{\beta}$.

A simple identification assumption would be that this model correctly predicts untreated outcomes, on average. This identification assumption is the basis for interrupted time series (ITS) designs (Bloom, 2003) and their recent advancements (Miratrix, 2022). If it held, we could write the ATT as $\text{E}\left[Y_{i,T} \mid G_i = 1\right] - \text{E}\left[\widehat{Y}_{i,T} \mid G_i = 1\right]$.

However, predictions are vulnerable to unexpected shocks, so we use the comparison group to tell us about what our prediction model misses. This idea underlies our identification assumption that expected prediction errors are equal in the treated and comparison groups.

ASSUMPTION 2 (Equal expected prediction errors).

$$\text{E}\left[Y_{i,T}(0) \mid G_i = 1\right] - \text{E}\left[\widehat{Y}_{i,T} \mid G_i = 1\right] = \text{E}\left[Y_{i,T}(0) \mid G_i = 0\right] - \text{E}\left[\widehat{Y}_{i,T} \mid G_i = 0\right]. \tag{3}$$

With this assumption, we can solve for the needed counterfactual and substitute it into the ATT.

THEOREM 1 (Causal identification by equal expected prediction errors). *Suppose Assumptions 1 and 2 hold. Then the ATT is*

$$\begin{aligned} ATT = &E\left[Y_{i,T} \mid G_i = 1\right] - E\left[\widehat{Y}_{i,T} \mid G_i = 1\right] \\ &- \left(E\left[Y_{i,T} \mid G_i = 0\right] - E\left[\widehat{Y}_{i,T} \mid G_i = 0\right]\right). \end{aligned} \tag{4}$$

Now we have the ATT in terms of observable quantities only, so we say that it is "identified." To distinguish the estimand in Eq. (4) from the target causal quantity in Eq. (1), we call the former an *identified* estimand.

[1] When we have only one post-period $t = T$, "previous" and "pre-period" are the same. For extensions to multiple post-treatment outcomes, we must also limit the inputs to the prediction model to variables that are fixed or observed prior to the start of treatment.

2.1. *Existing designs as special cases.* We next show that some familiar designs are special cases of this general identification strategy. For each, we show that we can define a prediction model such that when the design's identification assumptions hold, Assumption 2 does also.

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

We begin with a basic implementation in which there are two groups (treated and comparison) and two periods (pre-period $T-1$ and post-period T). Then we assume untreated potential outcomes would have evolved in parallel in the two groups:

$$\begin{aligned} &\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 1\right] - \mathrm{E}\left[Y_{i,T-1}(0) \mid G_i = 1\right] = \\ &\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 0\right] - \mathrm{E}\left[Y_{i,T-1}(0) \mid G_i = 0\right]. \end{aligned} \tag{5}$$

Assumption 1 and Eq. (5) allow us to solve for $\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 1\right]$ and substitute into the expression for the ATT in Eq. (2) to obtain,

$$\begin{aligned} ATT = &\left(\mathrm{E}\left[Y_{i,T} \mid G_i = 1\right] - \mathrm{E}\left[Y_{i,T-1} \mid G_i = 1\right]\right) - \\ &\left(\mathrm{E}\left[Y_{i,T} \mid G_i = 0\right] - \mathrm{E}\left[Y_{i,T-1} \mid G_i = 0\right]\right). \end{aligned}$$

This is our identified estimand under the difference-in-difference assumption of parallel trends. As a corollary to Theorem 1, we now show that there exists a prediction model such that, if parallel trends holds, equal expected prediction errors does also.

COROLLARY 1. *If Assumption 1 and Eq. (5) hold, then Assumption 2 also holds for the following prediction model:*

$$\widehat{Y}_{i,t} = \begin{cases} \frac{1}{n_1} \sum_{i:G_i=1} Y_{i,t-1} & \text{if } G_i = 1 \\ \\ \frac{1}{n_0} \sum_{i:G_i=0} Y_{i,t-1} & \text{if } G_i = 0 \end{cases} \tag{6}$$

for $t = 2, \ldots, T$ *where* $n_g = \sum_i \mathbb{1}\{G_i = g\}$

The proof is in Appendix A.1. This shows that we can use a prediction model that is simply each group's mean in the immediately preceding period. In the comparison group, we observe the difference from this prediction and actual outcomes (which is the first "difference" of difference-in-differences). We then use these prediction errors to correct the treated group's predictions (which is the second "difference" of difference-in-differences).

2.1.2. *Two-way fixed effects models.* Two-way fixed effects (TWFE) regression models, where "two-way" refers to unit and time fixed effects, are very popular in controlled pre/post studies (de Chaisemartin and D'Haultfœuille, 2023). Several papers have described problems with using TWFE regression estimators in the setting of staggered treatment timing and treatment effect heterogeneity (Goodman-Bacon, 2021; Borusyak, Jaravel and Spiess, 2022;

de Chaisemartin and D'Haultfœuille, 2023; Sun and Abraham, 2021). Kropko and Kubinec (2020) showed that while one-way fixed effects cleanly capture either over-time or cross-sectional dimensions, the TWFE model unhelpfully combines within-unit and cross-sectional variation . Using similar logic, Imai and Kim (2021) showed that the TWFE model's promise of simultaneous adjustment for unobserved unit and time confounders depends crucially on linearity and additivity.

Setting aside these complexities for the moment, we assume a basic TWFE model in which untreated potential outcomes are generated by a linear, additive structural model,

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

where $\mathrm{E}[\epsilon_{i,t}] = 0$ for all $i = 1, \ldots, n$ and $t = 1, \ldots, T$. This is also a special case of equal expected prediction errors.

COROLLARY 2. *If Assumption 1 and Eq. (7) hold, then Assumption 2 also holds for the following prediction model:*

$$\widehat{Y}_{i,t} = \hat{\alpha}_i, \text{ where } \hat{\alpha}_i = \arg\min_{\alpha_i} \sum_{s=1}^{t-1} \left(Y_{i,s} - \alpha_i\right)^2. \tag{8}$$

The proof is in Appendix A.2. This shows that we can use ordinary least squares to obtain predictions for post-period outcomes. It is easy to imagine elaborations on this structural outcome model that include covariates, more complicated time functions, etc.

2.2. *Existing designs that are not special cases.* The designs we have considered so far combine a pre-versus-post contrast with a treated-versus-comparison contrast. This is because they are motivated by a model for outcomes that combines stable group-level differences and unpredictable shocks that are common across groups. The pre-versus-post difference adjusts for time-invariant group differences, while the treated-versus-comparison difference adjusts for common shocks.

These differ from methods like interrupted time series and synthetic controls. Interrupted time series uses only a pre-versus-post contrast, while synthetic controls uses only a treated-versus-comparison contrast. A key motivation for synthetic control methods is an interactive fixed effects model that generates potential outcomes (Abadie, Diamond and Hainmueller, 2010). Because this produces differential evolution across groups, the prediction error in the comparison group will not inform us about prediction error in the treated group.

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. Combinations of DID and synthetic control are a topic of recent research (e.g., Arkhangelsky et al., 2021), and we leave further investigation of this idea for future research.

3. Selecting models for robustness. Having established that some familiar designs are special cases of our identifying assumption (given particular prediction models), we turn to choosing among prediction models. We first define robustness (the complement of sensitivity) and discuss the difference between robustness and plausibility.

3.1. *Design sensitivity.* Design sensitivity was developed in the context of matched observational studies, which assume equivalence to a block randomized experiment. If the assumption is exactly met, the point estimator converges in probability to a constant as the sample goes to infinity. However, for any magnitude of violation of the assumption, we can generate a *range* of point estimates consistent with the data and therefore a limiting *interval*. Design sensitivity is the length of this limiting interval (Rosenbaum, 2005, 2012). For a given magnitude of violation, a sensitive design has a wider limiting interval than a more robust one.

By contrast, to interrogate the *plausibility* of identifying assumptions, researchers often study whether a version of it holds in the pre-period. For instance, in DID designs, it is common to test for non-parallel trends in the pre-period, which resembles a Granger causality test (Granger, 1969) and other forms of "placebo" tests (see, e.g., Angrist and Pischke, 2008, p. 237). This implicitly assumes that patterns observed in the pre-period would have continued into the post-period in the absence of treatment. However, this approach replaces one unverifiable assumption about counterfactual outcomes with another (Egami and Yamauchi, 2022). Sensitivity analysis offer a practical and rigorous way out of this bind.

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

Similarly, we suppose that violations of our identifying assumption lie in a set defined by the observed pre-period differential prediction errors. Denote the differential prediction errors by

$$\delta_t = \left(\mathrm{E}\left[Y_{i,t}(0) \mid G_i = 1\right] - \mathrm{E}\left[\widehat{Y}_{i,t} \mid G_i = 1\right]\right) - \left(\mathrm{E}\left[Y_{i,t}(0) \mid G_i = 0\right] - \mathrm{E}\left[\widehat{Y}_{i,t} \mid G_i = 0\right]\right). \tag{9}$$

The point identification of Assumption 2 can be written as $\delta_T = 0$. For set identification, we suppose $\delta_T \in \Delta_T$, where Δ_T is a compact set. The lower and upper sensitivity bounds on the ATT are therefore

$$\underline{\mathrm{ATT}} = \mathrm{ATT} + \min_{\delta \in \Delta_T} \delta \tag{10}$$

$$\overline{\mathrm{ATT}} = \mathrm{ATT} + \max_{\delta \in \Delta_T} \delta\,, \tag{11}$$

where δ is an element from the set Δ_T.

Equations (10) and (11) lead to our definition of sensitivity, which is the difference between $\overline{\mathrm{ATT}}$ and $\underline{\mathrm{ATT}}$. A smaller difference implies less sensitivity (greater robustness).

To define a relevant Δ_T, we follow Rambachan and Roth (2023) and let Δ_T encompass violations up to M times the largest absolute differential prediction error observed in the pre-period,

$$\Delta_T = \left\{\delta : |\delta| \leq M \max_{v \in \mathcal{V}} |\delta_v|\right\}, \tag{12}$$

where $M \geq 0$ and $v \in \mathcal{V} \subseteq \{2, \ldots, T-1\}$ is a researcher-specified set of pre-treatment *validation periods* in which we assess deviations from equal expected prediction errors. For each validation period, $v \in \mathcal{V}$, δ_v is differential prediction error from a prediction model fit to data from $t < v$.

We can imagine alternatives to this set restriction that reflect different beliefs about the relationship between pre- and post-periods. For instance, we could remove the absolute value in Eq. (12) to create an asymmetric set restriction. If we think more recent validation periods

are more informative, we might use the set restriction as $\Delta_T = \{\delta : |\delta| \leq M|\delta_V|\}$ for $V = \max \mathcal{V}$, i.e., bound the violation by M times the *most recent* difference in expected prediction errors. Alternatively, if we think the average pre-treatment deviation matters, we could define $\Delta_T = \left\{\delta : |\delta| \leq M/|\mathcal{V}| \sum_{v \in \mathcal{V}} |\delta_v|\right\}$. We proceed with the set restriction in Eq. (12), but these alternatives are straightforward to implement.

The sensitivity parameter M controls how tightly we constrain the assumptions. Point identification holds for $M = 0$, but as long as there is some nonzero pre-treatment difference in expected prediction errors, $\Delta_T \to \Re$ as $M \to \infty$. Thus, we fix M and choose the prediction model that minimizes sensitivity for that fixed value (which is arbitrary). It follows immediately from the bounds in Equations (10) and (11) that if we define the set restriction on δ_T using Eq. (12), we should choose the prediction model that minimizes the maximum pre-treatment differential prediction error, $\max_{v \in \mathcal{V}} |\delta_v^{(f)}|$, where we now write $\delta_v^{(f)}$ with the superscript (f) to underscore the differential prediction error's dependence on the prediction model, f. Similarly, for any other set restriction (such as the most recent or mean over all validation periods, discussed above), we choose a prediction model that minimizes the width of the implied sensitivity bounds.

PROPOSITION 1. *Let f and f' be two prediction models in a set of candidate models, $\mathcal{F}$. Let $\delta_v^{(f)}$ and $\delta_v^{(f')}$ be their respective differences in expected prediction errors over validation periods, $v \in \mathcal{V}$. Under the sensitivity model in Eq. (12), model f is more robust than f' if and only if $\max_{v \in \mathcal{V}} |\delta_v^{(f)}| \leq \max_{v \in \mathcal{V}} |\delta_v^{(f')}|$.*

Our proposal does present a potential trade-off, however. Suppose that for some prediction model, equal expected prediction errors in the post-period holds exactly, but despite this, its maximum differential prediction error in the pre-period is larger than in some other candidate models. This prediction model, though unbiased, may appear less robust and therefore will not be chosen by our procedure. We formalize this potential bias-robustness trade-off in Proposition 2 below.

PROPOSITION 2. *Let f and f' be prediction models that yield predictions $\widehat{Y}_{i,f,T}$ and $\widehat{Y}_{i,f',T}$, respectively. Suppose that equal expected prediction errors in Assumption 2 holds exactly for f', but not f. However, suppose $\max_{v \in \mathcal{V}} |\delta_v^{(f')}| \geq \max_{v \in \mathcal{V}} |\delta_v^{(f)}|$. The bias for the point-identified ATT that we incur by choosing the more robust f over the unbiased f' is*

$$\left(E\left[\widehat{Y}_{i,f,T} \mid G_i = 0\right] - E\left[\widehat{Y}_{i,f',T} \mid G_i = 0\right]\right) - \left(E\left[\widehat{Y}_{i,f,T} \mid G_i = 1\right] - E\left[\widehat{Y}_{i,f',T} \mid G_i = 1\right]\right).$$

This proposition connects to the conception of robustness as stability of point estimates across competing models (Brown and Atal, 2019) or designs (O'Neill et al., 2016). In our framework, Proposition 2 implies that if competing models yield similar predictions, the robustness versus bias trade-off is mitigated. If two models produce equal predictions (and hence equal point estimates), there is no robustness-bias trade-off at all.

4. An algorithm for model selection, estimation, and inference in finite samples. Thus far, we have considered population quantities only. Below, we consider implementation of this procedure in finite samples.

For each prediction model $f \in \mathcal{F}$ and each validation period $v \in \mathcal{V}$, we generate predictions $f(\mathbf{X}_{i,v}, \boldsymbol{\beta}) = \widehat{Y}_{i,f,v}$ and estimated differential prediction errors

$$\widehat{\delta}_v^{(f)} = \left(\bar{Y}_{1,v} - \bar{\widehat{Y}}_{1,f,v}\right) - \left(\bar{Y}_{0,v} - \bar{\widehat{Y}}_{0,f,v}\right),$$

and then choose the model that minimizes our sensitivity measure

$$f^\dagger := \arg\min_{f \in \mathcal{F}} \max_{v \in \mathcal{V}} |\hat{\delta}_v^{(f)}|. \tag{13}$$

Given this selected model, the point estimator of the ATT in Eq. (14) is

$$\widehat{\text{ATT}} = \left(\bar{Y}_{1,T} - \bar{\widehat{Y}}_{1,f^\dagger,T}\right) - \left(\bar{Y}_{0,T} - \bar{\widehat{Y}}_{0,f^\dagger,T}\right), \tag{14}$$

where $\bar{Y}_{g,T} = \frac{1}{n_g} \sum_{i:G_i=g} Y_{i,T}$ and $\bar{\widehat{Y}}_{g,f^\dagger,t} = \frac{1}{n_g} \sum_{i:G_i=g} \widehat{Y}_{i,f^\dagger,T}$ are the mean outcome and prediction for group g in period T, respectively, and n_g denotes the number of units in group g. For some $M > 0$, the estimators of the lower and upper sensitivity bounds on the ATT are

$$\underline{\widehat{\text{ATT}}} := \widehat{\text{ATT}} - M \max_{v \in \mathcal{V}} |\hat{\delta}_v^{(f^\dagger)}| \tag{15}$$

$$\overline{\widehat{\text{ATT}}} := \widehat{\text{ATT}} + M \max_{v \in \mathcal{V}} |\hat{\delta}_v^{(f^\dagger)}|. \tag{16}$$

As outlined above, each model generates predictions using observed data and a parameter, $\hat{\boldsymbol{\beta}}$. For instance, if the prediction model is a regression model, $\hat{\boldsymbol{\beta}}$ is the vector of regression coefficients that minimizes the sum of squared errors, and the predictions are the product of the estimated coefficients and a design matrix based on $\mathbf{X}_{i,t}$. The parameters $\hat{\boldsymbol{\beta}}$ are not known exactly, and this uncertainty has downstream effects on the subsequent predictions, errors, sensitivity measure, selection of an optimal prediction model, and ultimately our estimator.

We want an inference method that accounts for this uncertainty, but the usual approach for capturing model selection uncertainty by splitting data into testing and training subsets is not viable here. This is because estimators use the model selection data *by construction*: terms in Equations (15) and (16) use data from pre-treatment validation periods $\mathcal{V}$. Therefore, we cannot separate the training data from the test data without making the kind of assumptions about the relationship between pre- and post-period outcomes that our design sensitivity framework is designed to avoid.

Instead, we extend the quasi-Bayesian procedure of Gelman and Hill (2006), which has been employed in related designs (Miratrix, 2022; Brodersen et al., 2015; King, Tomz and Wittenberg, 2000; Tomz, Wittenberg and King, 2003; Zhang et al., 2009). For a model with parameter vector $\boldsymbol{\beta}$, this procedure takes repeated draws of the parameter from a multivariate Normal distribution centered at the estimated parameter $\hat{\boldsymbol{\beta}}$ with variance-covariance matrix equal to the parameter's estimated variance-covariance $\widehat{\boldsymbol{\Sigma}}$. These draws resemble samples from the posterior of the parameter if the prior were flat, thus the "quasi-Bayesian" moniker.

To apply this to our setting, in each pre-treatment validation period, we fit the prediction models to previous periods' data and get their estimated parameters and corresponding variance-covariance matrices. Then we repeatedly draw the parameters of each prediction model from multivariate Normal distributions with mean and variance equal to the estimates from the fitted models. For each parameter draw and prediction model, we generate outcome predictions, calculate differential prediction errors, and use these to select the best prediction model. We fit the selected prediction model to all pre-treatment data, generate predictions in the post period, and estimate the ATT. Repeating this for many draws of the parameter vectors yields a distribution for the ATT that accounts for uncertainty in the models' parameters and thus the model selection procedure. The algorithm below formalizes this procedure, and Appendix B explores its limiting behavior in large samples.

Algorithm 1: Model selection, estimation, and inference

input : Data $\{Y_{i,t}, \boldsymbol{X}_{i,t}\}$, $i = 1, \ldots, n$ and $t = 1, \ldots, T$;
Class of prediction models, $\mathcal{F}$;
Sensitivity parameter, M

output: Simulation-based distribution of point, lower- and upper-bound estimates of ATT

for prediction models $f \in \mathcal{F} := \left\{f_1, \ldots, f_{|\mathcal{F}|}\right\}$ **do**
  for pre-treatment validation periods $v \in \mathcal{V} \subseteq \{2, \ldots, T-1\}$ **do**
    Fit f to data from periods $\{1, \ldots, v-1\}$ generating $\hat{\boldsymbol{\beta}}_{f,v}$ and $\widehat{\boldsymbol{\Sigma}}_{f,v}$
for simulations $r \in \{1, \ldots, R\}$, **do**
  for prediction models $f \in \mathcal{F} := \left\{f_1, \ldots, f_{|\mathcal{F}|}\right\}$ **do**
    for pre-treatment validation periods $v \in \mathcal{V} \subseteq \{2, \ldots, T-1\}$ **do**
      Draw $\boldsymbol{\beta}^r_{f,v} \sim N\left(\hat{\boldsymbol{\beta}}_{f,v}, \widehat{\boldsymbol{\Sigma}}_{f,v}\right)$
      Predict outcomes $\widehat{Y}^r_{i,f,v} = f\left(\boldsymbol{X}_{i,v}, \boldsymbol{\beta}^r_{f,v}\right)$
      Compute differential average prediction errors
      $\widehat{\delta}^r_{f,v} = \left(\bar{Y}_{1,v} - \bar{\widehat{Y}}^r_{1,f,v}\right) - \left(\bar{Y}_{0,v} - \bar{\widehat{Y}}^r_{0,f,v}\right)$
  Choose most robust prediction model, $f^{\dagger,r} := \underset{f \in \mathcal{F}}{\arg\min} \max_{v \in \mathcal{V}} |\widehat{\delta}^r_{f,v}|$
  Predict outcomes in post-period T from $f^{\dagger,r}$ fit to data from all pre-periods
  Calculate point estimate $\widehat{\text{ATT}}^r$
  if $M > 0$ **then**
    Calculate $\underline{\widehat{\text{ATT}}}^r$ and $\overline{\widehat{\text{ATT}}}^r$

From this algorithm, 95% confidence statements can be generated by taking by taking the middle 95% of the R estimates of the ATT. Likewise, researchers can also calculate p-values, technically posterior predictive p-values, which are known to be conservative (Meng, 1994).

5. The effect of gun laws changes on violent crime. We now return to our analysis of Missouri's repeal of its permit-to-purchase (PTP) law. To estimate the repeal's impact, we form a set of candidate prediction models drawn from the gun policy literature. Many researchers agree on a basic model with two-way fixed effects for units and time (as in Webster, Crifasi and Vernick (2014)), but disagree on other model components. Based on our survey of the literature, we divide the relevant model components into three categories:

1. Unit-specific time trends. Researchers often include unit-specific time trends, usually linear but sometimes more complicated forms (Black and Nagin, 1998; French and Heagerty, 2008). Others explicitly advocate against their inclusion (Aneja, Donohue III and Zhang, 2014; Wolfers, 2006). We consider models that include unit-specific linear, quadratic, or cubic time trends. In these models, we omit the time fixed effects to avoid over-parameterizing the models.

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

3. Outcome transformations. Linear regression is popular for outcome regressions, but can be problematic because many outcomes of interest (including the homicide rate that we consider) are naturally bounded (Moody, 2001; Plassmann and Tideman, 2001). Therefore, generalized linear models such as Poisson and negative binomial regression are often used instead. However, these non-linear models have their own challenges with interpreting the coefficients on interaction terms (Karaca-Mandic, Norton and Dowd, 2012; Ai and Norton, 2003; Puhani, 2012). We use only linear models, but we do consider transformations of the outcome variable, specifically logs and first differences (Black and Nagin, 1998). However, because we want to compare across models, we back-transform our prediction to the original outcome scale to compute prediction errors.

Obviously, this framework leaves out some modeling variations, for example, random effects (e.g., Crifasi et al., 2018), empirical Bayes estimators (Strnad, 2007) and two-stage models (Rubin and Dezhbakhsh, 2003). However, given the prominence of these three model components and the prominence of unit fixed effects and linear models, we believe the resulting set of candidate models is sufficiently broad and relevant to the gun policy literature.

From the model components above (summarized in Table 1), we take all possible combinations to derive a set of 24 candidate models. We fit each prediction model to each individual unit; thus all models effectively include unit fixed effects via the data subsetting approach (Kropko and Kubinec, 2020).

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

TABLE 1

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

To select among the 24 prediction models, we estimate the differences in expected prediction errors between treated and comparison groups. For each year prior to the law's passage in 2007, we train our prediction models on the previous years. For example, in 2006, we train a model on data from 1999 to 2005, predict in 2006, and compute the difference in average prediction errors between treated and comparison groups. To ensure adequate years of training data, we follow Hasegawa, Webster and Small (2019) in beginning the validation period in 1999. Thus, we have 5 or more years of training data, even in first validation year (1999, in which we train the model on 1994-1998 data).

Figure 2 shows the absolute differential average prediction errors for all 24 models, with the maximum for each model highlighted in black. The LDV model on the outcome's original scale without unit-specific time trends (row 3, column 1) minimizes our sensitivity criterion. The baseline mean model (row 1, column 1), which most closely corresponds to the model of choice in Webster, Crifasi and Vernick (2014) and Hasegawa, Webster and Small (2019), is the fourth-best model.

From Figure 2, we can also see which prediction models would be optimal under different sensitivity criteria. For example, the prediction model with the smallest absolute difference in average prediction errors in the last pre-period (2007), in expectation, is the linear time trend model (row 1, column 4). By contrast, the prediction model with the smallest expected absolute difference in average prediction errors, averaged over all validation periods, is the baseline mean model on the outcome's log scale (row 1, column 2). These different loss functions for choosing the optimal model can be justified by an appropriate sensitivity analysis

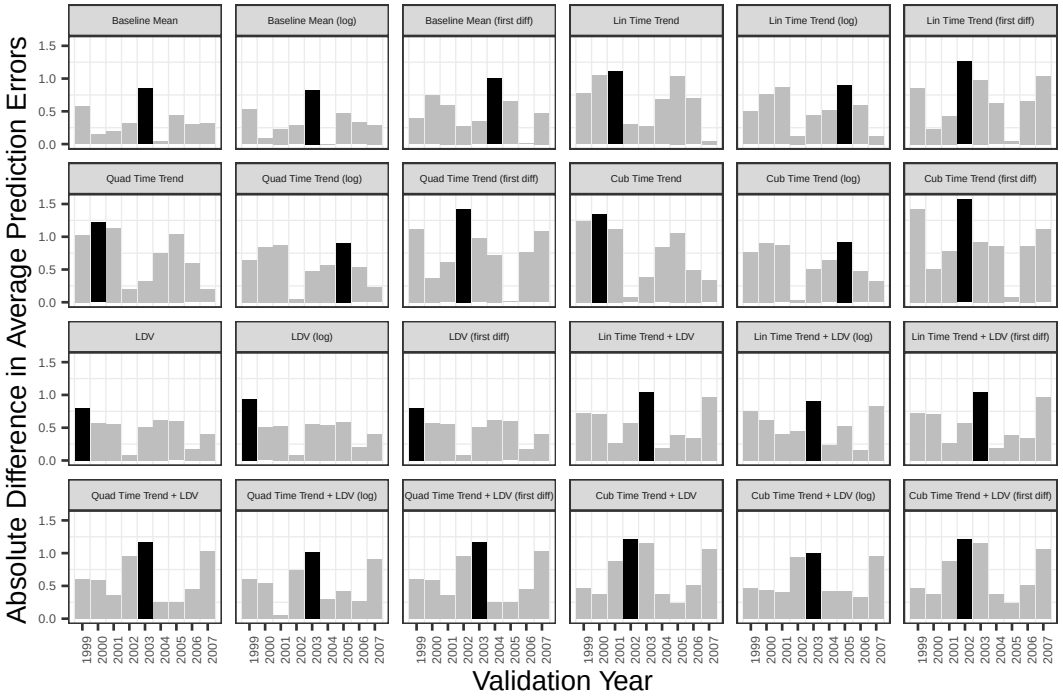

FIG 2. *Absolute difference in average prediction errors for all candidate models. The maximum for each model is highlighted in black.*

model. Given the sensitivity analysis in Eq. (12), which aligns with the sensitivity analysis proposed in recent research (Rambachan and Roth, 2023), LDV is the optimal model, in expectation. Using the algorithm above, which accounts for uncertainty in which model is optimal, yields a point estimate of 1.16 and 95% uncertainty interval of $[0.95, 1.37]$.

Figure 3 shows the predictions and their errors for all validation periods with the optimal prediction model. In it, we see that the maximum absolute differential prediction error (equivalent to the absolute difference in prediction errors) occurs in 1999 and is approximately 0.81. This is driven largely by the prediction errors in three control states, Arkansas, Oklahoma and Tennessee. (The prediction errors in the 5 other control states are relatively small.)

The LDV model fit to the training years in these three states estimates negative coefficients (-0.055, -1.486 and -1.351, respectively) on the lagged outcome in the regression. Hence, each state's drop in homicide rate in 1998 leads the model to predict even larger outcomes in 1999. However, the homicide rate drops even further in 1999. The training period in these three states has never shown two consecutive decreases, so the LDV model does not anticipate this pattern.

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

However, once we relax the assumption using the set identification strategy, the stakes of the choice of models become more important. Figure 4 shows the relationship between point estimates of the ATT and the maximum differential pre-treatment prediction errors of all 24 models. In it, we see that models yielding similar point estimates can differ dramatically in

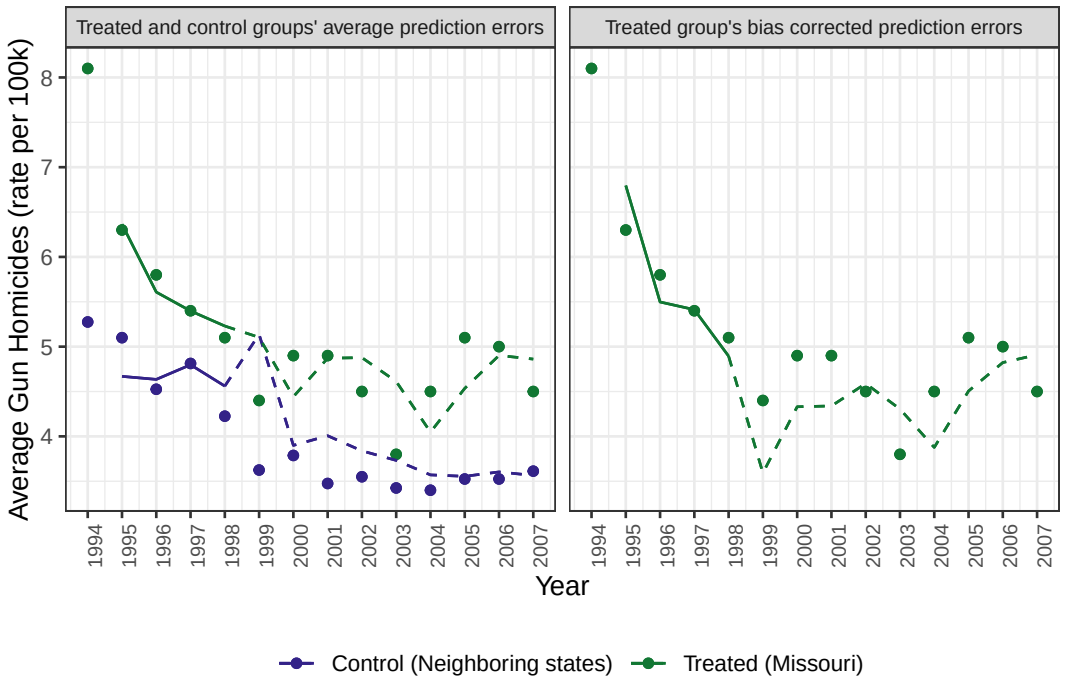

FIG 3. *Average outcomes and the optimal model's average predictions for all pre-treatment validation periods in treated and control states. The points are observed outcomes, the solid lines are model-fitted values in the training periods and the dashed lines are modeled predictions in the validation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

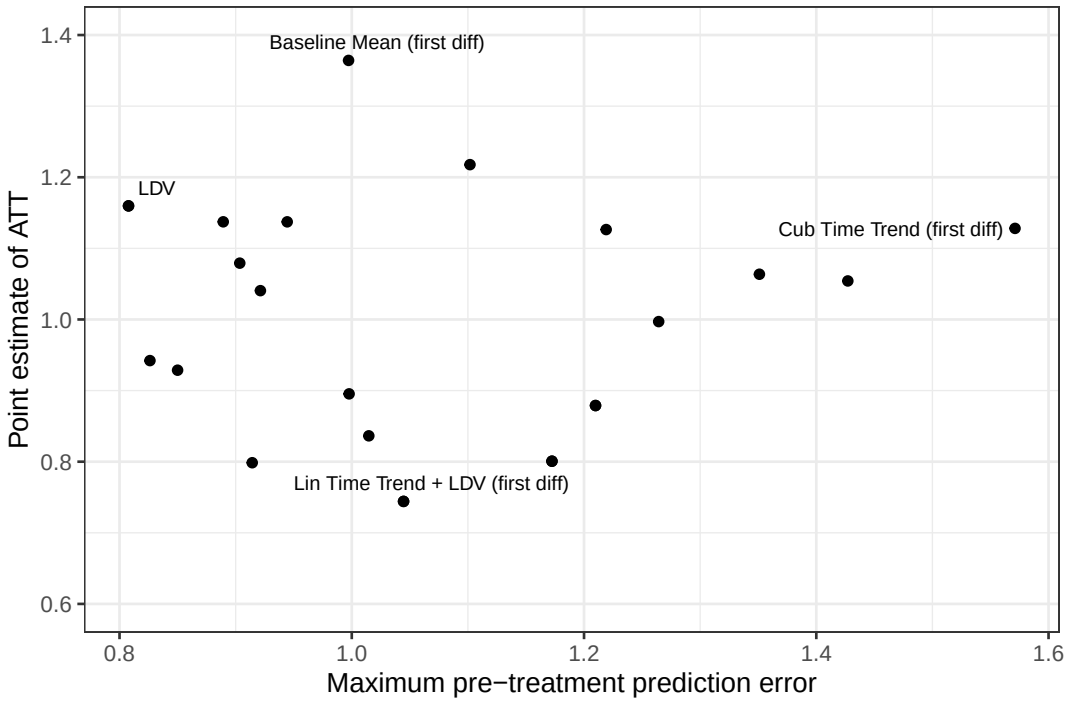

FIG 4. *Point estimates of ATT from each model (y axis) and corresponding maximum absolute differential prediction errors in the pre-period (x axis).*

terms of robustness. For instance, the most and least robust models yield similar point estimates of 1.16 and 1.13, respectively. Because of their different levels of robustness, however, the sensitivity bounds of the most robust model would include 0 with a value of M greater than or equal to 1.4. That is, the violation of equal-expected-prediction errors would have to be over 1.4 times greater than the worst violation in 9 prior years. For the least robust model, the value of M that causes the sensitivity bounds to include 0 is much smaller, only 0.7.

6. Conclusion and open questions. In this paper, we introduce a new method for identifying the ATT based on a "predict, correct, select" procedure. We *predict* untreated potential outcomes, *correct* them for unexpected shocks using the observed prediction error in the comparison group, and *select* the optimal prediction model using a robustness criterion. Our causal identification of the ATT based on these predictions assumes equal expected prediction errors in the treated and comparison groups.

We developed these ideas in an attempt to reconcile the wide variety of models used to evaluated gun policy effects. Specifically, we studied the repeal of Missouri's permit-to-purchase law in 2007 using models drawn from the literature. Rather than making claims about the plausibility of any underlying causal models, we selected the optimal model based on robustness. We found that a lagged dependent variable model had low average prediction errors throughout the pre-period and minimized our robustness criterion. The resulting point estimate was similar to point estimates under alternative models, but the differences became stark under mild violations of the identifying assumption.

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

Second, our method is scale-dependent because we measure prediction error as a linear difference on the scale of the outcome variable. This limits our approach, but we believe this limitation is compatible with applied models in common use.

Third, prediction models can only use variables that are measured prior to treatment or are structurally unaffected by treatment. For some data-generating models, such as interactive fixed effects, the correction step will not de-bias the estimator because shocks to the system will not affect treated and comparison groups equally.

Fourth, by switching to a robustness criterion for model selection, we introduce a potential bias-robustness trade-off. Our conception of robustness is different from conceptions based on the stability of point estimates across competing models (Brown and Atal, 2019) or designs (O'Neill et al., 2016). In our framework, two identical point estimates can differ in robustness, and we have no guarantee that choosing based on robustness will choose the true model (if it is among the candidate models). However, Proposition 2 implies that if competing models yield the same predictions, there is no robustness versus bias trade-off.

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

Third, many researchers already interpret robustness in terms of bias. In difference-in-differences, for instance, researchers interpret parallel trends in the pre-period as evidence for the plausibility of the true identifying assumption of parallel trends from the pre- to post-periods. Yet pre-period parallel trends provide evidence of counterfactual parallel trends only

under additional assumptions, and violations of pre-period parallel trends can still be consistent with the identifying assumption (Kahn-Lang and Lang, 2020; Roth and Sant'Anna, 2023). Therefore, our proposal offers a more transparent version of this practice, recasting the evaluation of pre-period violations as a sensitivity analysis rather than as a test of (untestable) assumptions.

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

However, we need not be limited to models already in use. A fifth benefit of our proposed method is its potential ability to draw upon flexible and modern prediction models, e.g., machine learning methods. Because our proposed model selection procedure is grounded in a model-free, causal identification framework, we need not believe the model. In fact, the inner workings of the prediction models can remain a black box. As long as it generates equally good predictions in the treated and comparison group, we can identify our target causal estimand.

APPENDIX A: EXISTING MODELS AS SPECIAL CASES

Our proofs each follow the steps sketched out below.

1. Use the design's identification assumptions to re-express the treated and comparison groups' expected untreated potential outcomes in the post-period, $\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 1\right]$ and $\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 0\right]$.
2. Write the expected prediction errors in treated and comparison groups:
   a) First, use Assumption 1 to substitute untreated potential outcomes for observed outcomes in the prediction model, $\widehat{Y}_{i,T} = f\left(\boldsymbol{X}_{i,t}\right)$.[2]
   b) Next, take expectation (with respect to the identification assumptions) of predictions in treatment and comparison groups, $\mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 1\right]$ and $\mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 0\right]$.
   c) Finally, compute the expected prediction errors in each group, $\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 1\right] - \mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 1\right]$ and $\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 0\right] - \mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 0\right]$.
3. Show that difference in expected prediction errors is equal to 0, thereby implying Assumption 2 and, consequently, the identified estimand in Eq. (4).

A.1. Difference-in-Differences. First, use parallel trends in Eq. (5) to write the treated and comparison groups expected untreated potential outcomes in the post-treatment period as

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

Then, using the prediction model in Eq. (6) and Assumption 1, write the expected predictions in treated and comparison groups as

$$\mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 1\right] = \mathrm{E}\left[Y_{i,T-1}(0) \mid G_i = 1\right]$$
$$\mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 0\right] = \mathrm{E}\left[Y_{i,T-1}(0) \mid G_i = 0\right].$$

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

Now we can write the expected prediction error in each group as

$$\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 1\right] - \mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 1\right] = \mathrm{E}\left[Y_{i,T}(0) \mid G_i = 0\right] - \mathrm{E}\left[Y_{i,T-1}(0) \mid G_i = 0\right]$$
$$\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 0\right] - \mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 0\right] = \mathrm{E}\left[Y_{i,T}(0) \mid G_i = 1\right] - \mathrm{E}\left[Y_{i,T-1}(0) \mid G_i = 1\right].$$

By Eq. (5), these two are equal, so Assumption 2 also holds.

A.2. Two-way Fixed Effects. First, note that Eq. (7) and the linearity of expectations imply that the treated and comparison groups expected untreated potential outcomes in the post-period are

$$\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 1\right] = \mathrm{E}\left[\alpha_i \mid G_i = 1\right] + \gamma_T$$
$$\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 0\right] = \mathrm{E}\left[\alpha_i \mid G_i = 0\right] + \gamma_T.$$

The prediction model in Eq. (8) is simply each unit's average outcome prior to t:

$$\hat{\alpha}_i = \arg\min_{\alpha_i} \sum_{s=1}^{t-1} \left(Y_{i,s} - \alpha_i\right)^2 = \frac{1}{(t-1)} \sum_{s=1}^{t-1} Y_{i,s}. \tag{17}$$

Given Assumption 1 and Eq. (7), we substitute for $Y_{i,t}$ in the prediction model in Eq. (17), which yields expected predictions for each group in period T of

$$\mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 1\right] = \mathrm{E}\left[\alpha_i \mid G_i = 1\right] + \left(\frac{1}{T-1}\right) \sum_{t=1}^{T-1} \gamma_t$$
$$\mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 0\right] = \mathrm{E}\left[\alpha_i \mid G_i = 0\right] + \left(\frac{1}{T-1}\right) \sum_{t=1}^{T-1} \gamma_t.$$

Now we can write the expected prediction error in each group as

$$\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 1\right] - \mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 1\right] = \gamma_T - \left(\frac{1}{T-1}\right) \sum_{t=1}^{T-1} \gamma_t$$
$$\mathrm{E}\left[Y_{i,T}(0) \mid G_i = 0\right] - \mathrm{E}\left[\widehat{Y}_{i,T} \mid G_i = 0\right] = \gamma_T - \left(\frac{1}{T-1}\right) \sum_{t=1}^{T-1} \gamma_t.$$

It immediately follows that the difference in expected prediction errors is equal to 0, thereby implying Assumption 2.

Note that although the structural model in Eq. (7) contains time fixed effects, the prediction model in Eq. (8) does not. A prediction model that includes time fixed effects would be equivalent to fitting the same model in Eq. (8) to mean-centered outcomes (Kropko and Kubinec, 2020). However, under the TWFE model in Eq. (7), this inclusion of time fixed effects in the predictive model would be unnecessary. Time fixed effects are constant across units, so they are eliminated by the treated-minus-control difference between groups.

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

To understand the properties of our model selection, estimation, and inference procedure, we show that the probability of selecting the truly optimal model goes to 1 as $n \to \infty$. This implies that, in a sufficiently large sample, model selection uncertainty can be ignored and inference can be conducted conditional on the chosen optimal model.

First, limit the set of candidate models to *stable* prediction models. Under random sampling from a target population, our prediction models are random since the prediction in period t is based on a model fit to data from periods before t; hence, the model that generates predictions in period t will vary depending on which data happen to be realized. By stability of a random sequence of these prediction models, where the sequence's index, $n = 2, 3, \ldots$, is suppressed for notational simplicity, we follow Guo and Basse (2023) in defining stability as

$$\text{Stability} := \|\hat{f}\left(\boldsymbol{X}_{i,t}\right) - f\left(\mathrm{E}\left[\boldsymbol{X}_{i,t}\right]\right)\|^2 = o_p(1) \text{ for all } i = 1, \ldots, n \text{ and } t = 1, \ldots, T, \tag{18}$$

where the norm, $\|\cdot\|$, satisfies the properties of a norm on functions and $f\left(\mathrm{E}\left[\boldsymbol{X}_{i,t}\right]\right)$ is the prediction model, f, applied to the population-level analogue of the vector of predictors, $\boldsymbol{X}_{i,t}$. Under standard conditions on population-level outcomes and predictors, the myriad ordinary least squares functions in common use are all stable. Guo and Basse (2023) show that a much larger class of smooth, parametric models are also stable.

Stability implies only that the sample prediction approximates the population-level prediction for a sufficiently large sample size. This definition of stability differs from alternatives (e.g., Bühlmann and Yu, 2002) in that a prediction model need not be consistent for a model-based parameter. In our setting, a prediction model is an algorithm used only to generate predictions, not to correctly explain an underlying stochastic process giving rise to potential outcomes. (For this conception of prediction models, albeit in different statistical contexts, see Huang et al. 2023, Rosenbaum 2002 and Sales, Hansen and Rowan 2018.)

Proposition 3 establishes that choosing the model that minimizes the estimated sensitivity will end up choosing the truly optimal prediction model with probability that goes to 1 as $n \to \infty$.

PROPOSITION 3. *Suppose a class of stable prediction models, $\mathcal{F}$ and independent and identically distributed (i.i.d.) sampling. Without loss of generality, let f^* denote the unique optimal model in the population whereby $f^* := \arg\min_{f \in \mathcal{F}} \max_{v \in \mathcal{V}} |\delta_v^{(f)}|$. It follows that*

$$\lim_{n \to \infty} \Pr\left(\max_{v \in \mathcal{V}} |\hat{\delta}_v^{(f^*)}| < \max_{v \in \mathcal{V}} |\hat{\delta}_v^{(f)}| \right) = 1 \textit{ for all } f \in \{\mathcal{F} \setminus f^*\}.$$

Proposition 3 states that the optimal model in the population yields the smallest absolute prediction error with limiting probability 1. Hence, in concert with Proposition 1, so long as the sample size is sufficiently large, choosing the model that minimizes an estimate of the worst-case difference in prediction errors will choose the optimal model. This provides a large-sample justification for the algorithm above.

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

BRODERSEN, K. H., GALLUSSER, F., KOEHLER, J., REMY, N., SCOTT, S. L. et al. (2015). Inferring Causal Impact Using Bayesian Structural Time-Series Models. *The Annals of Applied Statistics* **9** 247–274.

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

BÜHLMANN, P. and YU, B. (2002). Analyzing Bagging. *The Annals of Statistics* **30** 927–961.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

LIU, L., WANG, Y. and XU, Y. (2023). A Practical Guide to Counterfactual Estimators for Causal Inference with TimeSeries CrossSectional Data. *American Journal of Political Science*.

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

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

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

MENG, X.-L. (1994). Posterior Predictive p-Values. *The Annals of Statistics* **22** 1142–1160.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

STRNAD, J. (2007). Should Legal Empiricists Go Bayesian? *American Law and Economics Review* **9** 195-303.

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

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

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

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

ZHANG, F., WAGNER, A. K., SOUMERAI, S. B. and ROSS-DEGNAN, D. (2009). Methods for Estimating Confidence Intervals in Interrupted Time Series Analyses of Health Interventions. *Journal of Clinical Epidemiology* **62** 143–148.