Evaluating federal policies using Bayesian time series models: estimating the causal impact of the hospital readmissions reduction program

Georgia Papadogeorgou1 · Fiammetta Menchetti2 · Christine Choirat3 · Jason H. Wasfy4 · Corwin M. Zigler5 · Fabrizia Mealli2,6

Abstract
Researchers are often faced with evaluating the effect of a policy or program that was simultaneously initiated across an entire population of units at a single point in time, and its effects over the targeted population can manifest at any time period afterwards. In the presence of data measured over time, Bayesian time series models have been used to impute what would have happened after the policy was initiated, had the policy not taken place, in order to estimate causal effects. However, the considerations regarding the definition of the target estimands, the underlying assumptions, the plausibility of such assumptions, and the choice of an appropriate model have not been thoroughly investigated. In this paper, we establish useful estimands for the evaluation of large-scale policies. We discuss that imputation of missing potential outcomes relies on an assumption which, even though untestable, can be partially evaluated using observed data. We illustrate an approach to evaluate this key causal assumption and facilitate model elicitation based on data from the time interval before policy initiation and using classic statistical techniques. As an illustration, we study the Hospital Readmissions Reduction Program (HRRP), a US federal intervention aiming to improve health outcomes for patients with pneumonia, acute myocardial infarction, or congestive failure admitted to a hospital. We evaluate the effect of the HRRP on population mortality among the elderly across the US and in four geographic subregions, and at different time windows. We find that the HRRP increased mortality from pneumonia and acute myocardial infarction across at least one geographical region and time horizon, and is likely to have had a detrimental effect on public health.

Keywords Bayesian methods · Causal inference · Hospital readmissions reduction program · Mortality · Policy evaluation · Time-series

Georgia Papadogeorgou and Fiammetta Menchetti are equally contributing authors.

* Georgia Papadogeorgou
gpapadogeorgou@ufl.edu

Extended author information available on the last page of the article
1 Introduction

Researchers and policy makers are often interested in understanding the effect of new policies or programs. The target policy is often rolled out at once over the whole region of interest, it alters the operation of a group of organizations, which might affect the region’s population in a number of ways. To evaluate the policy, available data include a time period before policy rollout (the “pre-intervention” period), and a time period after policy initiation (the “post-intervention” period). Considering that the policy rollout is experienced by everyone in the region, policy evaluation should also target the policy’s impact on the whole population. Therefore, we are faced with estimating the causal effect of a policy over a single unit measured over time, including a pre- and a post-intervention time period.

In the presence of pre- and post-intervention data, popular methods include the difference-in-difference (Athey and Imbens 2006) and synthetic control (Doudchenko and Imbens 2017) approaches. However, these approaches require the presence of control units during both time periods, whereas in our case, everyone is exposed to the policy once it is initiated. Other measured time series for the same population have been previously used as assumed control units in synthetic control methodology (Gaughan et al. 2019). Interrupted (or quasi-experimental) time-series analysis has been alternatively used for estimating causal effects from a single time series (McDowall et al. 2019), but it requires that we correctly specify the outcome model in both the pre- and in the post-intervention time period, an assumption that is rather restrictive. Bojinov and Shephard (2019) introduced causal estimands and estimation based on a single time series, but their methodology requires that the treatment can be applied and taken away at any time point, which is not true for policy settings where once the policy is initiated it remains active throughout.

In this manuscript, we suggest using the whole region where the intervention took place as a single system (unit) for policy evaluation. Under the potential outcome framework for causal inference, we define causal estimands that describe the effect of the policy on the whole system, and introduce the crucial causal assumptions for estimating these quantities using observed data. For estimation, we employ Bayesian time series modeling to impute what would have happened in the post-intervention time period, had the intervention not taken place, and we combine these imputations with observed outcomes to unbiasedly estimate causal effects, similarly to other approaches in the literature (Brodersen et al. 2015; Miratrix et al. 2019; Antonelli and Beck 2020). We illustrate how model choice and the plausibility of the causal assumptions can be partially evaluated using data from the pre-intervention time period and classic statistical techniques, and using time series data that are expected to be unaffected by the intervention. The framework presented here provides guidance for the evaluation of programs and policies of various sorts, such as educational policies, social programs, or environmental policies that are adopted by schools, neighborhoods, and polluting sources, respectively, and affect the outcomes of the region’s population (Zigler and Papadogeorgou 2021).

This work is motivated by the evaluation of the Hospital Readmissions Reduction Program (HRRP), part of the Affordable Care Act which passed in 2010. The HRRP created the prospect of financial penalties that would start being applied in 2012 based on risk-standardized readmission rates for Medicare fee-for-service patients within 30 days of discharge after index hospitalization for acute myocardial infarction (AMI), congestive heart
failure (CHF), or pneumonia. The HRRP reflected intent to incentivize providers and hospitals to improve the coordination and quality of care for patients. Even though the law was associated with substantial declines in risk-standardized readmissions (Zuckerman et al. 2016; Desai et al. 2016; Wasfy et al. 2017; MedPAC 2018), whether the law was associated with increased mortality is very controversial (Joynt et al. 2011; Joynt and Jha 2012; Fonarow et al. 2017) with studies returning drastically different results (Gupta et al. 2017; Dharmarajan et al. 2017; Wadhera et al. 2018; Khera et al. 2018; Huckfeldt et al. 2019; Sandhu and Heidenreich 2019). Improving our understanding of the causal effects of the HRRP is critical, because this type of information will inform whether similar policies should be implemented or not in the future. Although most of the aforementioned papers adopt methodologies that are often employed in the causal inference literature (e.g., interrupted time series), the underlying assumptions are not explicit and the estimands are not clearly defined. This is necessary to give a reader an understanding of the effect one is aiming to estimate and under which assumptions it can be interpreted as “causal.” Within our framework, we define and estimate the causal effect of the HRRP on mortality from AMI, CHF or pneumonia among individuals that were 65 years old or older across the United States and in four subregions, and at three time windows after the passage of the law. We find that the HRRP caused an increase in mortality from AMI across the US and from pneumonia in some geographical subregions.

2 Methods

2.1 Evaluating a policy at the same level of its scope

For ease of illustration, we focus on the HRRP throughout as a policy which affects hospital operations in an effort to improve public health. To evaluate such a policy, the prevailing analysis approach compares the outcomes of hospitals that were affected by the policy at varying degrees. However, such an analysis is faced with many challenges. Firstly, hospitals that adopt the policy at different degrees can have systematic differences due to both measured and unmeasured factors which might confound the estimation of causal effects on hospital-level outcomes. Second, the extent to which a hospital adopts the policy might affect the outcome of other hospitals, giving rise to interference, which complicates how causal effects are both defined and estimated. That can happen if, for example, hospitals with high financial penalties under the HRRP avoid admitting high-risk patients, who are afterwards admitted at a different hospital (Casalino et al. 2007). Lastly, even if effects of the intervention on hospital-level outcomes were properly defined and estimated, they would represent the effects of policy adoption of different degrees rather than the total effect of the policy, and conclusions on population-level effects based on the analysis of hospital-level outcomes might suffer from ecological bias.

Understanding the effects of policy adoption at different degrees can be essential for uncovering key features of the policy, but the challenges in the analysis of hospital-level data presented above have not been definitively resolved and interpretation of estimated quantities remains complicated. Here, we advocate a complementary approach that bypasses those specific issues in focusing on different, but related, causal estimands. We suggest that evaluation of large-scale policies is performed at the level of the whole
population, by viewing the whole region as one system measured over time, and considering causal estimands that describe the change in the outcome over the whole population in the presence and in the absence of the intervention. As we discuss below, this entails notions of confounding that do not necessarily require adjustment for measured differences. It also bypasses the issue of interference, since the whole population composes a single unit that is treated by the policy, and we observe the outcome at an aggregate level irrespective of which hospital provided care to whom. Finally, and importantly for policy evaluation, the framework below allows us to quantify the effect of the policy directly on the population under study, and therefore does not suffer from ecological biases. We find that providing a complimentary way to evaluate large-scale policies by shifting the goalpost from hospital-level analyses to focusing on estimands over the whole population is a key contribution of our work.

2.2 Causal estimands in single time series with one interventional time point

Since the policy could have been passed or not, we formulate our causal estimands using two potential time series: one representing the outcome had the policy taken place, which is observed, and one representing the outcome had the policy not taken place, which is not observed and has to be estimated.

Let \( t \in T = \{1, 2, \ldots, T\} \) represent \( T \) time periods of observed data, and \( t^* \in T \) denote the time period during which the policy was actually initiated. Let \( w_t = 1 \) denote that the policy is active at time period \( t \), and \( w_t = 0 \) that it is not. Then,

\[
\mathbf{w}_{1T} = (w_1 = 0, w_2 = 0, \ldots, w_{t^*-1} = 0, w_{t^*} = 1, w_{t^*+1} = 1, \ldots, w_T = 1)
\]

(1)

represents the observed scenario where the policy is initiated at time \( t^* \) and remains active thereafter, and

\[
\mathbf{w}_{0T} = (w_1 = 0, w_2 = 0, \ldots, w_{t^*-1} = 0, w_{t^*} = 0, w_{t^*+1} = 0, \ldots, w_T = 0)
\]

(2)

represents the counterfactual scenario where the policy is never initiated. We postulate the existence of potential outcomes \( Y_{jt} = Y_t(\mathbf{w}_{jt}) \) for \( j \in \{0, 1\} \), which represent the population outcome that would have been observed at time period \( t \) had the policy never been initiated, or had it been initiated at \( t^* \), respectively. Then, \( Y^1_t - Y^0_t \) represents the effect of initiating the policy at time \( t^* \) against not initiating it at all on the outcome at time period \( t \), similarly to the causal estimands in the synthetic control literature (e.g., Doudchenko and Imbens 2017). We focus at time periods after policy initiation, \( t \geq t^* \). Then, the causal effect of the policy \( k \) time periods after initiation is defined as

\[
\Delta_k = Y^{1(t^*+k)} - Y^{0(t^*+k)} , \ k \geq 0,
\]

(3)

and the cumulative effect during the first \( K \) time periods is defined as

\[
C \Delta_K = \sum_{k=0}^{K-1} \Delta_k = \sum_{k=0}^{K-1} \left[ Y^{1(t^*+k)} - Y^{0(t^*+k)} \right] .
\]

(4)

In our study, the potential outcomes \( Y_{jt} \) represent the total number of deaths across the US had the HRRP passed or not, and the cumulative effect in (4) represents the number of deaths caused or prevented by the policy within \( K \) time periods from its initiation.
Defining potential outcomes at the level of the whole region obviates the need to consider interference explicitly. In contrast, hospital-level potential outcomes should be indexed by the treatments of all hospitals which interfere with it, which would complicate the definition of subsequent hospital-level estimands. As also highlighted in Imbens and Rubin (2015), a common strategy to deal with interference between statistical units is changing the level of analysis. Therefore, considering the whole region under study as one unit for which the policy would either take place or not allows us to bypass this issue.

2.3 The causal assumptions

Since the policy was indeed initiated at time period \( t^* \), the observed treatment is equal to the treatment path \( \bar{w}_{1T} \), and the observed outcome \( Y_t \) is equal to the corresponding potential outcome, \( Y_t = Y_{1t} \). This implies that all potential outcomes corresponding to the treatment path \( \bar{w}_{1T} \) are observed, as illustrated in the top row of Table 1. Therefore, all potential outcomes \( Y_{1t} \) in the estimands (3), (4) are observed, and we are faced with estimating the potential outcomes \( Y_{0t} \) in the post-intervention period, \( t \geq t^* \). To do so, we make the following assumptions. The interpretation of these assumptions as well as how we can assess their plausibility is discussed in more detail in Sects. 2.4 and 2.5, and within the context of our study in Sect. 3.2.

**Assumption 1** The policy does not alter potential outcomes before its initiation, \( Y_{0t} = Y_{1t} \) for \( t < t^* \).

Assumption 1 states that the intervention could not have affected the outcome before it was actually initiated. Therefore, potential outcomes in the pre-intervention time period are the same, irrespective of whether the policy is initiated or not in the future. This assumption establishes that the observed outcomes in the pre-intervention period are equal to the potential outcomes in the absence of the intervention, \( Y_{0t} = Y_t \) for \( t < t^* \). The next assumption allows us to use the pre-intervention outcomes to learn the outcome process in the absence of the treatment, potentially conditional on covariates, and is discussed in more detail in Sect. 2.4.

**Assumption 2** There exist measured covariates \( X_t \) which are unaltered by the intervention, and conditional on which the outcome process is stationary with some lag \( M \geq 0 \), i.e. if \( X^t = X^t(\bar{w}_{jT}) \) are the potential covariate values under treatment path \( \bar{w}_{jT} \) for \( j = 0, 1 \), we have that \( X^0_t = X^1_t \), and there exists \( M \geq 0 \) such that, for all \( t \in T \),

\[
P(\bar{Y}^0|Y^{0(t-1)}, \ldots, Y^{0(t-M)}, X^0, \ldots, X^{0(t-M+1)}) = P(\bar{Y}^{0(t+1)}|\bar{Y}^0, \ldots, Y^{0(t-M+1)}, X^{0(t+1)}, \ldots, X^{0(t-M+2)}).
\]

Assumption 2 is the basis for imputing the missing potential outcomes in the absence of the policy. For illustration, assume that it holds for \( M = 1 \) and in the absence of covariates. Then, the conditional outcome distribution is the same for all time points in the

Table 1: Potential outcomes in the pre- and post-intervention time period, and the causal assumptions

|                         | Pre-intervention period \( t < t^* \) | Post-intervention period \( t \geq t^* \) |
|-------------------------|---------------------------------------|------------------------------------------|
| \( Y_{1t} \)            | Observed                              | Observed                                 |
| \( Y_{0t} \)            | = Observed (Assumption 1)             | Imputed based on Assumption 2             |
pre-intervention period, i.e. for $t_1, t_2 < t^*$ it holds that $P(Y^{0t_1} | Y^{0(t_1-1)}) = P(Y^{0t_2} | Y^{0(t_2-1)})$. Since the potential outcomes $Y^{0t}$ in the pre-intervention period are observed from Assumption 1, this conditional distribution can be estimated from the data. Then, setting $t = t^*$ in Assumption 2, we have $P(Y^{0t^*} | Y^{0(t^*-1)}) = P(Y^{0(t^*-1)} | Y^{0(t^*-2)})$ which allows us to impute $Y^{0t^*}$. Once $Y^{0t^*}$ is imputed based on this distribution, the same statement for $t = t^* + 1$ will be used to impute $Y^{0(t^*+1)}$ and so forth. Therefore, Assumption 2 enables us to learn the potential outcome process using the observed outcomes in the pre-intervention period, and use it to predict the missing potential outcomes in the post-intervention period had the policy not taken place.

### 2.4 Confounding with a single unit measured over time

Here, we start by discussing the notion of confounding in the setting of a single unit followed over time, and then we relate it to our Assumption 2. An analogous discussion in the context of multiple units measured over time is given in Antonelli and Beck (2020).

In most settings in causal inference, measured covariates are assumed to satisfy a no-unmeasured confounding assumption in that they represent all the ways in which treated and control units differ, and that, within cells defined by these covariates, treatment assignment is independent of any potential outcome. In our setting, the whole system is viewed as a single unit subject to the intervention for which we have access to a time period before the intervention when the unit is untreated, and a time period after the intervention when the unit is treated. Construing the different time periods as the most primitive observational units, confounding would represent systematic differences in the pre- and post-intervention time periods. Then, the primary threat to validity with estimating the missing potential outcome $Y^{0t}$ in the post-intervention time period is parsing changes due to the policy from other temporal changes that would have occurred regardless. In the HRRP, this could correspond to secular trends in population lifestyle factors or improvements in medical care technology that are not related to the policy, but may have changed mortality outcomes over time. To estimate the causal effect by comparing the outcomes in the pre- and post-intervention time period, we would have to account for all such variables through matching, sub-classification, regression, or other methods.

Rather than finding pre- and post intervention time points with similar values of covariates (which would not generally be available) to account for factors that vary coincidentally with the policy and outcomes, Assumption 2 focuses on the potential outcomes in the absence of the intervention and states that secular and periodic trends in the pre-intervention outcome time-series would have persisted in the post-intervention period, had the intervention not taken place. Therefore, this assumption could be satisfied if temporal trends in the pre-intervention time period are properly accommodated using flexible functions of time or covariates. Flexible functions of time can be used to account for the overall decreasing trends in mortality during the pre-intervention period reported in Fig. S.2, without requiring to have access to all variables that might drive these trends. In fact, flexible temporal trends can adequately capture covariate time series that predict the outcome and have long term temporal trends, such as population comorbidities which have overall persistent trends during our study period. Therefore, the assumption might hold without any covariates, in which case covariates are not needed for unbiased estimation of causal effects. However, researchers have to be mindful to include covariates that predict the outcome time series and have a temporal pattern that changes post-intervention, since these covariates will not be captured by a flexible temporal trend. Importantly, the covariates that
are used in the outcome model should *not* be influenced by the treatment themselves. For example, in a hypothetical study which evaluates the effect of a transportation policy to reduce air pollution, the number of circulating cars should not be used as a covariate, as it is likely to be affected by the intervention.

As a result, the potential drawback of the proposed approach lies in that it cannot capture changes in the temporal trends of the counterfactual time series after the intervention which occur due to unpredictable changes in missing predictors, or other co-occurring causes. Causal inference methodologies for panel data with control time series partially alleviate this issue. For example, synthetic control methodology could accommodate unpredictable changes in the counterfactual time series as long as these changes in the temporal and secular trends manifest similarly in the treated and control units, and regression approaches with treated and control units can improve effect estimation in the presence of co-occurring policies, as long as these policies are not enacted in near succession (Griffin et al. 2022). Similarly to our approach, synthetic control and difference-in-differences methodologies are also based on alternatives to the classic no-unmeasured confounding assumption, though they rely on control time series for effect estimation. Synthetic control methodology requires that the weights learnt from data in the pre-intervention period are stable into the post-intervention period, and difference-in-difference methodology is based on the famous parallel trends assumption which, even though can hold conditional on covariates, does not require that all confounders are measured (Angrist and Pischke 2008).

### 2.5 Assessing the plausibility of the causal assumptions

Section 2.3 formalizes the assumptions necessary for unbiased estimation of causal effects. Even though Assumptions 1 and 2 cannot be formally tested, they can be qualitatively and quantitatively evaluated based on the observed data. We discuss how to assess the plausibility of the causal assumptions.

Assumption 1 describes that outcomes before the interventional time period could not have been altered in *anticipation* of the intervention. Therefore, to satisfy this assumption, the interventional time point \( t^* \) should be set as the first time period during which the policy might have altered the outcome under study. That might be at the time of policy implementation, or even earlier when the policy is passed or first announced. Specifying \( t^* \) earlier than necessary will not bias effect estimation; instead, estimation will not be necessarily unbiased if \( t^* \) is not set early enough (for an approach that identifies the time period when the treatment effect materializes see Cruz et al. 2017). An approach to quantitatively assess the plausibility of Assumption 1 is to use the proposed method with an earlier intervention time period, and ensure that the estimated effect during the pre-intervention period is null.

Secondly, Assumption 2 allows for the use of covariates that are unaltered by the policy. To check the plausibility that our covariates are not affected by the policy we could, in principle, test whether the intervention impacted the considered covariates by replacing the outcome under study with each covariate and repeating the policy evaluation. In practice, this step can be avoided in many instances. For example, exogenous covariates can often be used straightforwardly. In the study of the HRRP, such covariates could represent weather information which is clearly not affected by the policy. After a set of covariates that are not affected by the intervention has been chosen, the pre-intervention outcome data can be used to evaluate Assumption 2 using classic statistical tools. We suggest using posterior predictive checks (Gelman et al. 1996, 2013) to do so. Posterior predictive checks involve generating replicated data by drawing from the posterior predictive distribution and
comparing these draws to the observed data using both numerical and graphical checks. A simple graphical check plots the mean of the posterior predictive distribution against the observed data. After obtaining the posterior predictive distribution of a given test quantity (e.g., mean, median, maximum), a numerical check could be based on the so-called Bayesian p-value defined as the proportion of times that the observed test quantity exceeds the one of the replicated data. In general, an extreme p-value (very close to 0 or 1) indicates that the feature of the data captured by the test quantity is inconsistent with the assumed model. In addition to posterior predictive checks, residuals diagnostics can be used to investigate the presence of residual autocorrelation and the viability of the residual normality assumption. If any of these tests fails, the assumed model is likely to be unfit to describe potential outcomes under control in the post-intervention period. In contrast, if the assumed model is deemed plausible, it does not necessarily mean that Assumption 2 holds for the chosen model. One crucial aspect of diagnostics tests for Assumption 2 is that they only entail the data from the pre-intervention period, so model evaluation can be performed completely separately from effect estimation. In Section 3.2 we show how we used posterior predictive checks to select and validate the model in the context of our study. In a related context, Antonelli and Beck (2020) proposed an alternative approach to model evaluation which pertains to using the observed pre-intervention data to simulate hypothetical policies and evaluating the model based on the estimation of their effects.

2.6 Estimating the causal effect

Once a model that passes the model assessment step is chosen, causal effect estimation and inference is straightforward by combining the observed outcomes with the samples from the posterior distribution of the imputed potential outcomes \( Y_{0t} \). Let \( Y_{0t}^{(b)} \) denote the \( b^{th} \) posterior sample for \( b = 1, 2, \ldots, B \) and \( t \geq t^* \). Then, \( \{ \Delta_{k(b)} = Y_{t+k}^{(b)} - Y_{(t+k)}^{0(b)} \} \) represent \( B \) samples from the posterior distribution of the point-wise effect \( \Delta_{k} \) in (3). These estimates are aggregated over the \( K \) time points in the post-intervention period to acquire posterior samples of \( C \Delta_{K} \) in (4). Then, we can use the posterior mean as the point-estimator of the causal effects \( \Delta_{k} \) and \( C \Delta_{K} \), and the quantiles of the posterior distribution for inference.

3 Evaluation of the Hospital Readmissions Reduction Program

3.1 Data sources and construction of the data set

We collected monthly level death counts from pneumonia, acute myocardial infarction (AMI) and congestive heart failure (CHF) among people aged more than 65 years old (all gender, all races, all origins) from January 2000 to December 2019. The data were gathered from the Centers for Disease Control and Prevention WONDER database by specifying the underlying cause of death using ICD-10 codes (J12–J18 for pneumonia, I21 for AMI, and I50 for CHF). We considered death counts at the national level, but we also performed regional-level analyses to study how treatment effects varied across regions. We used the geographical regions corresponding to the Northeast, Midwest, South and West United States as coded by the US Census Bureau (see Supplement A in the Online Resource for the definition of each region). By analyzing regions separately, we assume
that there is no interference across regions, an assumption that is realistic considering the size of each of these regions.

To improve the prediction accuracy of the outcome in the absence of intervention, we also included a set of covariates that might be linked with the outcome, they are unaffected by the intervention, and do not necessarily follow predictable temporal trends. First of all, we considered weather information:

1. Temperature (TEMP) measured as a monthly average in each region, obtained from the National Centers for Environmental Information;
2. Heat index (HEAT) defined as an indicator variable that takes value 1 when the maximum temperature registered in a given month and region is above 28°C and 0 otherwise; and
3. Cold index (COLD) defined indicator variable that takes value 1 when the minimum temperature registered in a given month and region is below 0°C and 0 otherwise.

The heat and cold indices are included in the model in addition to temperature as extremely hot or cold weather is believed to be a predictor of mortality (Medina-Ramón et al. 2006), and its occurrence, though temporally correlated, does not always exhibit a stable pattern over time. We also included

4. An air pollution metric reflecting yearly averages of PM$_{2.5}$ concentrations at the national level obtained from the US Environmental Protection Agency; and
5. A poverty measure (POV) defined as the percentage of population below the poverty level, measured yearly both at the regional and national level, obtained from the Census Bureau.

Since we are evaluating a possible impact of the HRRP on mortality, we excluded information on admitted patients such as comorbidities, since whether a patient is admitted or not, and as a result their characteristics, might be affected by the intervention (Casalino et al. 2007).

### 3.2 Causal assumptions and model evaluation for estimating the effects of the HRRP

Even though the financial penalties introduced by the HRRP started to be applied in 2012, we set the intervention date $t^*$ as March 2010, corresponding to the time period that the policy reform was passed. We use the passage of the law instead of its implementation as the time of intervention to account for the possibility that hospitals might have changed their behaviour right after the program’s announcement, and in expectation of the program’s implementation (Assumption 1). Therefore, our data include a time period of approximately 10 years pre-intervention, and 9 years post-intervention. We return to evaluating Assumption 1 in the end of this section, and after model choice according to Assumption 2 has been completed.

Since the covariates considered are exogenous to the health system, we expect that they are unaltered by the intervention (Assumption 2). While temperature data may follow cyclical patterns, the poverty measure reverts its increasing trend shortly after the HRRP (see Fig. S.3). Thus, by including it in our models we allow for changes in our outcome series in the post intervention period due to the trend change in poverty levels, which would otherwise not be possible to predict, as explained in Sect. 2.4.
To predict the counterfactual outcome in the post-intervention period, we estimate the outcome model process based on data from the control, pre-intervention period, and using Bayesian Structural Time Series models (West and Harrison 2006). These models require a preliminary analysis to characterize the features of the generating process (e.g., linear trend, seasonality, cycle). In this specific case, the data exhibit a yearly seasonal pattern, which is visible from the autocorrelation functions in Fig. S.1 of the Online Resource. Similar seasonal patterns are observed across the three conditions, and both at a national and regional level (see Fig. 1 and S.2 in the Online Resource). The outcome evolution in the pre-intervention period exhibits decreasing trends for pneumonia and AMI (Fig. 1). We formally investigated whether a local level or a local linear trend best describes these trends in the outcome process based on the posterior predictive checks described in Sect. 2.5. We also investigated whether the exogenous covariates and seasonality should be included in the model or not in a similar manner. The results of the model evaluation process for all models considered are included in the Online Resource. Based on these results, we find that, for each of the three outcomes, a local linear trend model with covariates and a seasonal component fits the observed, pre-intervention outcome data the best. This model can be written as

$$
Y_t = \mu_t + \gamma_t + X_t \beta + \varepsilon_t \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)
$$

$$
\mu_{t+1} = \mu_t + \nu_t + \eta_{t,\mu} \quad \eta_{t,\mu} \sim N(0, \sigma_\eta^2)
$$

$$
\nu_{t+1} = \nu_t + \eta_{t,\nu} \quad \eta_{t,\nu} \sim N(0, \sigma_\eta^2)
$$

$$
\gamma_{t+1} = -\sum_{s=1}^{S-2} \gamma_{t-s} + \eta_{t,\gamma} \quad \eta_{t,\gamma} \sim N(0, \sigma_\eta^2)
$$

where $Y_t$ is the outcome time series; $X_t$ is a vector of covariates with coefficients $\beta$; $\mu_t$, $\nu_t$, and $\gamma_t$ are, respectively, the trend component, the random slope and the seasonal component; and $\varepsilon_t, \eta_{t,\mu}, \eta_{t,\nu}, \eta_{t,\gamma}$ are serially independent random variables with zero mean and constant variances. To understand the contribution of each component, consider a simple case with no covariates and no seasonal term: if $\nu_t = 0$, the trend component evolves as a random walk and model (5) reduces to a local level; instead, if $\eta_{t,\mu} = \eta_{t,\nu} = 0$, then $\nu_{t+1} = \nu_t = \nu$ and $\mu_{t+1} = \mu_t + \nu$, the trend would be exactly linear making (5) a deterministic linear trend.
plus noise model. Therefore, a local linear trend model allows for flexible, non-linear temporal trends in the outcome time series, that could capture predictors of mortality whose long term trends persist during our study period (see Sect. 2.4). Finally, the seasonal component $\gamma_t$ allows each season to have a different contribution to the overall mean, while ensuring that over $S$ seasons the aggregate contribution of $\gamma_t$ is centered at zero. This model is also used for the regional analyses. Even though models with Gaussian errors for ITS analyses do not perform well for count data in general (Ye et al. 2022), it is expected that it is a sufficiently good approximation here since the observed counts are very large (Agresti 2013; Coxe et al. 2009), and our posterior predictive checks illustrate good performance in predicting the observed counts (first column in Fig. S.4). In Sect. 3.4 we consider alternative model specifications and discuss how the results from these alternative analyses relate to the one presented here.

With this model choice, we return to evaluating the presence of anticipation effects in Assumption 1 following the instructions given in Sect. 2.5. More specifically, we fictionally set the intervention date earlier on December 2009 and then forecast what would have happened in the following three months. We estimate anticipatory effects for all three conditions with 95% credible intervals overlapping zero (effects represent number of deaths avoided or caused by the HRRP; pneumonia: $-1323$ with 95% CI from $-4112$ to $1329$; AMI: $-1344$ with 95% CI from $-3997$ to $1338$; and CHF: $579$ with 95% CI from $-390$ to $1451$). Therefore, there is no indication that Assumption 1 is violated.

### 3.3 Evaluation of the impact of the HRRP on mortality

We estimated the effect of the passage of the HRRP on mortality from pneumonia, AMI and CHF among the elderly at the national level and in each of the four regions, and at three different time horizons after the intervention: short-term (March 2010–December 2012), mid-term (March 2010–December 2015), and long-term (March 2010–December 2019). For the regional analyses, we estimated the model on mortality data from pneumonia, AMI and CHF for each region separately. We excluded the PM$_{2.5}$ concentration due to lack of easily-available regional data. All the computations were done using the 

\[ \text{CausalImpact} \]

R package (Brodersen et al. 2015). Fig. 1 shows the evolution of the raw death counts from the three conditions at the national level. The presence of seasonal patterns in the considered time series are evident.

Table 2 presents the results of our national analysis in terms of estimates of $C_{\Delta K}$ and corresponding 95% credible intervals for the short-, mid-, and long-term. In Table 3 we provide estimates of $C_{\Delta K}$ for the four regions and the three time horizons. To ease cross-regional comparisons, we also provide estimates and 95% credible intervals for the proportion of deaths by each condition that is attributed to the HRRP (by diving our estimates and credible interval bounds with the observed number of deaths in each region). Lastly, Fig. 2 shows the cumulative effect of the HRRP across the US and in the Northeast for each condition and until December 2015 (short to mid-term). The plots for the other regions can be found in the Online Resource.

At the national level, we find that the HRRP led to a medium and long-term increase in mortality from AMI. By comparing the national results with those emerging from the

---

1 Looking at Fig. S.16 in the Online Resource, we believe that PM$_{2.5}$ can be excluded from the regional analysis without concerns. Indeed, the inclusion probabilities of PM$_{2.5}$ are small in all models; moreover, the sensitivity analysis performed at the national level shows robustness to different choices of predictors (namely, the effects in Table 2 are in line with the estimates resulting from a local linear trend and seasonal model without PM$_{2.5}$, see panel (1) in Table 4).
In this table, $C_{\Delta}$ denotes the cumulative impact at the end of three different time horizons (short, mid, and long-term). Results with 95% CIs that do not include 0 are shown in bold.

| Table 2 | Cumulative causal effect and 95% credibility intervals of the HRRP on mortality from pneumonia, AMI and CHF at the national level |
|---------|------------------------------------------------------------------------------------------------------------------------------------|
| PNE     | AMI                                                                                                                                | CHF                                                                     |
| C$\Delta$ | 2.5%  | 97.5% | C$\Delta$ | 2.5%  | 97.5% | C$\Delta$ | 2.5%  | 97.5%       |
| Short-term | 10,795 | –2092 | 23,665 | 12,333 | –734  | 25,239 | 3780  | –3035  | 10,509 |
| Mid-term   | 49,445 | –36,549 | 134,641 | 83,970 | 3555  | 162,972 | 35,300 | –26,811 | 94,469 |
| Long-term  | 98,849 | –149,887 | 340,540 | 223,041 | 2644  | 442,617 | 94,727 | –96,895 | 27,8605 |

As we see in Fig. 2, the uncertainty bands for the cumulative effect estimates get wider as we predict further into the future. This is also true for the point-wise effect estimates (see Fig. S.19). This is due to the auto-regressive mean structure which requires uncertainty.

Table 3 Causal effect of the HRRP on mortality across US regions

| PNE     | AMI                                                                 | CHF                                                                 |
|---------|-------------------------------------------------------------------|---------------------------------------------------------------------|
| C$\Delta$ | C$\Delta^p$ (95% CI) | C$\Delta$ | C$\Delta^p$ (95% CI) | C$\Delta$ | C$\Delta^p$ (95% CI) |
| Short-term |                   |                                       |                   |                                       |
| Northeast | 3931   | 22.2 (5.6, 38.5) | 3541   | 10.7 (1.4, 20.1) | 741  | 4.4 (–4.7, 13.2) |
| Midwest  | 3517   | 19.4 (1.4, 37.5) | 3644   | 8.9 (0.2, 17.5)  | 995  | 4.2 (–4.4, 12.7) |
| South    | 3669   | 13.5 (–2.4, 29.1) | 3394   | 5.3 (–2.3, 12.7) | 582  | 1.7 (–5.8, 9.3)  |
| West     | 2131   | 13.4 (–3.2, 30.1) | 2270   | 7.8 (0.9, 14.5)  | 1319 | 8.7 (–0.1, 17.5) |
| Mid-term  |                   |                                       |                   |                                       |
| Northeast | 16,155 | 28.9 (–5.4, 62.6) | 20,294 | 20.3 (2, 38.7)   | 5958 | 10.3 (–12.8, 32.8) |
| Midwest  | 18,957 | 32.6 (–3.8, 69.4) | 23,100 | 18.4 (1.2, 35.4) | 10,157 | 26.3 (–10.6, 33.6) |
| South    | 16,954 | 19.3 (–13.9, 51.8) | 28,313 | 14.1 (–0.7, 28.9) | 11,898 | 10.1 (–9.1, 28.9) |
| West     | 14,247 | 27.7 (–8.2, 63.4) | 18,404 | 20.2 (6.1, 34.7) | 9795  | 19 (–5.4, 43.1)  |
| Long-term |                   |                                       |                   |                                       |
| Northeast | 31,269 | 34.2 (–28.5, 95.2) | 52,046 | 33 (0.2, 65.6)   | 14,865 | 14.7 (–26.9, 55.6) |
| Midwest  | 41,004 | 44.2 (–20.9, 110.7) | 57,955 | 28.5 (–1.3, 58)  | 25,984 | 18.1 (–21.6, 56.6) |
| South    | 35,671 | 24.8 (–33.8, 82.5) | 77,855 | 23.4 (–2.6, 49.6) | 32,349 | 15.5 (–18.1, 48.9) |
| West     | 33,938 | 40.2 (–22.7, 102.8) | 55,022 | 36.1 (10.8, 62.1) | 27,003 | 28.6 (–14, 71.2) |

C$\Delta$ denotes the number of deaths attributed to the HRRP. C$\Delta^p$ shows the proportion of deaths attributed to the HRRP (and 95% credible intervals) in each region. Results are shown for the three time horizons: i) 1 year (December 2011); ii) 5 years (December 2015); iii) 9 years (December 2019). Bold face is used for estimates with corresponding credible interval that does not include zero.

As we see in Fig. 2, the uncertainty bands for the cumulative effect estimates get wider as we predict further into the future. This is also true for the point-wise effect estimates (see Fig. S.19). This is due to the auto-regressive mean structure which requires uncertainty.

regional break down, we obtain interesting findings. Short-term mortality is concentrated in the first two regions for pneumonia. Interestingly, even though we did not find the presence of a short run effect for AMI at the national level, we can see mortality from AMI increased due to the HRRP in three regions (the null result in the South US might be behind the absence of a causal effect at the national level). For the mid-term effects, we can see that the increase in mortality from AMI is mostly driven by the Northeast, Midwest and West regions. Finally, the results from our long-term regional analysis show that the national result for AMI is mostly explained by the Northeast and West US regions. Finally, both the national and regional analyses reflected null effect estimates of the HRRP on CHF mortality.

As we see in Fig. 2, the uncertainty bands for the cumulative effect estimates get wider as we predict further into the future. This is also true for the point-wise effect estimates (see Fig. S.19). This is due to the auto-regressive mean structure which requires uncertainty. 

 Springer
quantification for outcomes further in the future to incorporate the uncertainty of parameters and imputed potential outcomes for all intermediate time periods. We find this to be an attractive property of this approach since it represents that the observed outcome time series trends are more trustworthy when extrapolating in the near future. For a fixed $k$, the width of the uncertainty band for our causal effect estimators will not go to zero, irrespective of the length of the pre-intervention period. That is because we estimate the causal effect for a single unit, we do not benefit from estimating averages over a sample, and as a result the uncertainty embedded in the posterior predictive distribution for the unit’s missing potential outcome will propagate to our estimator. Even though including predictors of the outcome could reduce the outcome model’s residual variance and as a result the variance of the imputed potential outcomes, the potential outcomes’ inherent variability will persist.

### 3.4 Additional analyses of our data

Considering the policy importance of our results, we performed several checks to evaluate the performance of our approach and the robustness of our qualitative results to the model chosen and the methodology used. These checks correspond to analyses of the effect of the HRRP under the following general categories: (1) analyses based on our model and alternative choice of predictors, (2) analyses on control time series that are expected to be unaffected by the intervention, (3) analyses based on alternative methodologies that employ a different set of causal assumptions, and (4) analyses based on transformations of our data. The analyses are discussed in detail in Sect. E in the Online Resource, and the acquired results are summarized here.

First, we evaluate the robustness of our quantitative results to alternative choices of predictor variables. We follow a three-step procedure: We establish alternative sets of predictor variables that one could consider. Then, for each set of predictors and without looking at the effect estimates based on each model, we inspect the posterior predictive checks of the model fits for each outcome and each set of covariates. If the check shows that the model is not a good fit for the observed data, the corresponding set of covariates is discarded.
Using the models with predictors that passed the posterior predictive checks, we estimate the causal effects, and we compare them to the results from our original analysis. For our study, we considered three alternative sets of predictors: (1) no predictors; (2) all predictors excluding PM$_{2.5}$; and (3) only weather covariates. The posterior predictive checks of the model without predictors (Fig. S.5 in the Online Resource) show high residual autocorrelation, and therefore this model is discarded. Conversely, the model without PM$_{2.5}$ and the model with only the weather covariates fit the data well (see Figs. S.8, S.9 and S.14, S.15 in the Online Resource). The causal effect estimates from these two models are shown in Table 4 and they are extremely similar to our original results in Table 2, illustrating that our results are robust to the choice of predictor variables.

As an additional robustness check, we performed falsification tests by repeating the analysis on death counts from conditions that, in principle, should not be affected by the HRRP: if the estimated impact on such conditions is close to zero, we can be more confident about the results obtained for pneumonia, AMI, and CHF. To do so, we acquired monthly level death counts from the Centers for Disease Control and Prevention WONDER database by specifying as underlying cause of death the following codes: A00–B99 (infectious and parasitic diseases); C00–D49 (tumors); D50–D89 (diseases of the blood); F01–F99 (mental disorders); V00–Y99 (external causes of morbidity). Figure S.20 in the Online Resource shows the evolution in mortality from each disease from January 2000 to December 2019. We can observe an increasing trend in mortality from all causes in the considered period. The results from each estimated model are reported in Table 5 and show no evidence of a causal effect for any of the control outcomes. These results show that if unpredictable changes in mortality trends exist, they are not present in control conditions so we can be more reassured that such changes are also not present in the target conditions. In addition, falsification tests are also important to ascertain the impact of possible co-occurring policies: if no evidence of an effect is found on control conditions, we can be more confident that co-occurring policies had no impact on target conditions as well. Finally, we verified the absence of anticipatory effects on both the true outcomes and on control conditions (see Table S.1 in the Online Resources).

As we discussed in Sect. 2.4, our approach can suffer if there exist structural changes in the temporal trends of the outcome time series after the intervention that are not due to the intervention. To evaluate the robustness of our qualitative analyses, we considered alternative methodologies that aim to address these issues, namely: interrupted time series,

|                | PNE          | AMI          | CHF          |
|----------------|--------------|--------------|--------------|
|                | CΔ 2.5%      | CΔ 2.5%      | CΔ 2.5%      |
| (1) Short-term | 10,756 –1814 | 12,386 –372  | 3432 –3244   | 9906          |
| Mid–term       | 47,828 –36,526| 133,777 86,054| 163,997 31,374| 91,649        |
| Long–term      | 98,180 –146,105| 342,845 216,877| 437,220 91,494| 273,341       |
| (2) Short-term | 10,555 –1511 | 12,570 –583  | 3709 –2841   | 10,357        |
| Mid–term       | 48,263 –35,138| 130,502 84,239| 3634 39,010 | 95,792        |
| Long–term      | 105,614 –127,016| 22,753 211,273| 439,858 98,079| 274,474       |

Table 4 Cumulative causal effect and 95% credibility intervals of the HRRP on mortality from pneumonia, AMI and CHF at the national level from two local linear trend models with different choices of regressors: (1) all regressors excluding PM$_{2.5}$; (2) only weather covariates. Bold face is used for estimates with corresponding credible interval that does not include zero.
We find that the main methodological contributions of this paper are three-fold. In a time-series setting with one time series and one interventional time point, we (1) defined quantities to be estimated in terms of potential outcomes, (2) formalized the assumptions on which a previously-developed (Brodersen et al. 2015) Bayesian time series model was based (which had not been done until now), and importantly (3) provided a comprehensive interrogation of the assumptions underlying the estimation of such causal estimands in the

### Discussion

We find that the main methodological contributions of this paper are three-fold. In a time-series setting with one time series and one interventional time point, we (1) defined quantities to be estimated in terms of potential outcomes, (2) formalized the assumptions on which a previously-developed (Brodersen et al. 2015) Bayesian time series model was based (which had not been done until now), and importantly (3) provided a comprehensive interrogation of the assumptions underlying the estimation of such causal estimands in the
context of our study, laying the foundation for the use of this method in a wide range of applications assessing clinical outcomes for health policy in non-randomized settings.

In our illustrative study, we evaluated the effect of the HRRP passage on mortality among the elderly. One of the key differences of our approach compared to previous associational studies is our focus on the whole region of interest as one unit measured over time, which allows us to estimate the effect of the HRRP passage on the whole population it meant to serve. Since the presented approach bypasses some complicating aspects of policy evaluation, we view our approach as complementary to approaches that compare penalized to non-penalized hospitals.

To our knowledge, this is the first paper evaluating the causal effect of the HRRP. Previously published analyses have suggested that after enactment of the HRRP, readmissions have decreased for all of the 3 initial penalty conditions (Zuckerman et al. 2016; Desai et al. 2016; Wasfy et al. 2017), even though these reductions in readmissions are likely due to increases in post-discharge emergency department visits or observation stays (Wadhera et al. 2019). Although our results are not directly comparable to the above studies, a fraction of reduced readmissions may result in increased mortality; in addition, our analysis is also able to capture the portion of mortality due to the HRRP arising from other sources (e.g., delayed first admissions and in-hospital mortality), or due to the possibility that hospital resources were redistributed away from reducing mortality to reducing readmissions.

Even though reductions in readmissions are encouraging, our analysis raises concerns about increased mortality caused by the HRRP. From the perspective of policy evaluation, our results illustrate that the future of the HRRP might have to be reassessed and future amendments might be necessary to avoid a detrimental impact of the policy on the population it was meant to serve.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10742-022-00294-8.

Acknowledgements The authors would like to thank Francesca Dominici, Alessandra Mattei, and Joseph Antonelli for their constructive comments.

Funding Funding was provided by National Institutes of Health R01 GM111339, R01 ES024332, R01 ES026217, P50MD010428, DP2MD012722, R01 ES028033, USEPA 83615601, and Health Effects Institute 4953-RFA14-3/16-4. The contents of this work are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication. Research described in this article was conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the United States Environmental Protection Agency (EPA) (Assistance Award No.CR-83467701) and certain motor vehicle and engine manufacturers. The contents of this article do not necessarily reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers. Dr. Wasfy is supported by National Institutes of Health and Harvard Catalyst KL2 TR001100 and American Heart Association 18CDA34110215. Dr. Menchetti and Prof. Mealli thank the Department of Excellence 2018-2022 funding provided by the Italian Ministry of Education, University and Research (MIUR).

Data availability The code and datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

Declaration

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.
References

Abadie, A., Gardeazabal, J.: The economic costs of conflict: a case study of the Basque Country. Am. Econ. Rev. 93(1), 113–132 (2003)

Abadie, A., Diamond, A., Hainmueller, J.: Synthetic control methods for comparative case studies: estimating the effect of California’s tobacco control program. J. Am. Stat. Assoc. 105, 493–505 (2010)

Abadie, A., Diamond, A., Hainmueller, J.: Comparative politics and the synthetic control method. Am. J. Polit. Sci. 59, 495–510 (2015)

Agresti, A.: Categorical Data Analysis. John Wiley & Sons, Hoboken (2013)

Angrist, J.D., Pischke, J.S.: Mostly Harmless Econometrics: An Empiricist’s Companion. Princeton University Press, Princeton (2008)

Antonelli, J., Beck, B.: Heterogeneous causal effects of neighborhood policing in New York city with staggered adoption of the policy (2020). arXiv preprint arXiv:2006.07681

Athey, S., Imbens, G.W.: Identification and inference in nonlinear difference-in-differences models. Econometrica 74(2), 431–497 (2006). https://doi.org/10.1111/j.1468-0262.2006.00668.x

Ben-Michael, E., Feller, A., Rothstein, J.: The augmented synthetic control method. J. Am. Stat. Assoc. 116(536), 1789–1803 (2021)

Bernal, J.L., Cummins, S., Gasparrini, A.: Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int. J. Epidemiol. 46(1), 348–355 (2017)

Bojinov, I., Shepherd, N.: Time series experiments and causal estimands: exact randomization tests and trading. J. Am. Stat. Assoc. 114(528), 1665–1682 (2019)

Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods. Springer science & business media, Berlin (1993)

Brodersen, K.H., Gallusser, F., Koehler, J., et al.: Inferring causal impact using Bayesian structural time-series models. Annals Appl. Stat. 9(1), 247–274 (2015). https://doi.org/10.1214/14-AOAS788

Card, D., Krueger, A.B.: Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. Tech. rep., Working Paper No. 4509, National Bureau of Economic Research (1993)

Casalino, L.P., Elster, A., Eisenberg, A., et al.: Will pay-for-performance and quality reporting affect health care disparities? These rapidly proliferating programs do not appear to be devoting much attention to the possible impact on disparities in health care. Health Aff. 26(Suppl2), w405–w414 (2007)

Coxe, S., West, S.G., Aiken, L.S.: The analysis of count data: a gentle introduction to Poisson regression and its alternatives. J. Pers. Assess. 91(2), 121–136 (2009)

Cruz, M., Bender, M., Ombao, H.: A robust interrupted time series model for analyzing complex health care intervention data. Stat. Med. 36(29), 4660–4676 (2017)

Desai, N.R., Ross, J.S., Kwon, J.Y., et al.: Association between hospital penalty status under the hospital readmission reduction program and readmission rates for target and nontarget conditions. J. Am. Med. Assoc. 316(24), 2647–2656 (2016). https://doi.org/10.1001/jama.2016.18533

Dharmarajan, K., Wang, Y., Lin, Z., et al.: Association of changing hospital readmission rates with mortality rates after hospital discharge. J. Am. Med. Assoc. 318(3), 270–278 (2017). https://doi.org/10.1001/jama.2017.8444

Doudchenko, N., Imbens, G.: Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. no. 22791. National Bureau of Economic Research (2016)

Doudchenko, N., Imbens, G.W.: Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis (2017). arXiv preprint arXiv:1610.07748v2

Fonarow, G.C., Konstam, M.A., Yancy, C.W.: The hospital readmission reduction program is associated with fewer readmissions, more deaths: time to reconsider. J. Am. Coll. Cardiol. 70(15), 1931–1934 (2017). https://doi.org/10.1016/j.jacc.2017.08.046

Gaughan, J., Gutacker, N., Grašič, K., et al.: Paying for efficiency: incentivising same-day discharges in the English NHS. J. Health Econ. 68(102), 226 (2019)

Gelman, A., Carlin, J.B., Stern, H.S., et al.: Bayesian Data Analysis. CRC Press, Boca Raton (2013)

Gelman, A., Meng, X.L., Stern, H.: Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica (1996)

Griffer, B.A., Schuler, M.S., Pane, J., et al.: Methodological considerations for estimating policy effects in the context of co-occurring policies. Health Serv. Outcomes Res. Methodol. pp 1–17 (2022)

Gupta, A., Allen, L.A., Bhatt, D.L., et al.: Association of the hospital readmissions reduction program implementation with readmission and mortality outcomes in heart failure. JAMA Cardiol. 3(1), 44–53 (2017). https://doi.org/10.1001/jamacardio.2017.4265

Huckfeldt, P., Escarce, J., Sood, N., et al.: Thirty-day postdischarge mortality among black and white patients 65 years and older in the medicare hospital readmissions reduction program postdischarge
mortality among black and white patients with medicarepostdischarge mortality among black and white P. JAMA Netw. Open 2(3), e190,634-e190,634 (2019). https://doi.org/10.1001/jamanetworkopen.2019.0634

Imbens, G.W., Rubin, D.B.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge (2015)

Joynt, K.E., Jha, A.K.: Thirty-day readmissions—truth and consequences. N. Engl. J. Med. 366(15), 1366–1369 (2012). https://doi.org/10.1056/NEJMp1201598

Joynt, K.E., Orav, E.J., Jha, A.K.: Thirty-day readmission rates for medicare beneficiaries by race and site of care. J. Am. Med. Assoc. 305(7), 675 (2011). https://doi.org/10.1001/jama.2011.123

Khera, R., Dharmarajan, K., Wang, Y., et al.: Association of the hospital readmissions reduction program with mortality during and after hospitalization for acute myocardial infarction, heart failure, and pneumonia. JAMA Netw. Open 1(5), e182777–e182777 (2018). https://doi.org/10.1001/jamanetworkopen.2018.2777

McDowall, D., McCleary, R., Bartos, B.J.: Interrupted Time Series Analysis. Oxford University Press, Oxford (2019)

Medina-Ramón, M., Zanobetti, A., Cavanagh, D.P., et al.: Extreme temperatures and mortality: assessing effect modification by personal characteristics and specific cause of death in a multi-city case-only analysis. Environ. Health Perspect. 114(9), 1331–1336 (2006)

MedPAC.: The effects of the Hospital Readmissions Reduction Program. Tech. rep., (2018) http://www.medpac.gov/docs/default-source/reports/jun18_ch1_medicarereport_sec.pdf?sfvrsn=0

Menchetti, F., Cipollini, F., Mealli, F.: Combining counterfactual outcomes and ARIMA models for policy evaluation., forthcoming. The Econometrics Journal. Earlier version available at (2021) https://arxiv.org/abs/2103.06740

Meyer, B.D., Viscusi, W.K., Durbin, D.L.: Workers’ compensation and injury duration: evidence from a natural experiment. Am. Econ. Rev. 85(3), 322–340 (1995)

Miratrix, L.: Package simits: analysis via simulation of interrupted time series (ITS) data (2020)

Miratrix, L., Anderson, C., Henderson, B., et al.: Simulating for uncertainty with interrupted time series designs. Tech. rep., (2019) https://www.mdrc.org/sites/default/files/img/methods_for_ITS.pdf

O’Neill, S., Kreif, N., Grieve, R., et al.: Estimating causal effects: considering three alternatives to difference-in-differences estimation. Health Serv. Outcomes Res. Method. 16, 1–21 (2016)

Ryan, A.M., Burgess, J.F., Jr., Dimick, J.B.: Why we should not be indifferent to specification choices for difference-in-differences. Health Serv. Res. 50, 1211–1235 (2015)

Sandhu, A.T., Heidenreich, P.A.: Comparison of the change in heart failure readmission and mortality rates between hospitals subject to hospital readmission reduction program penalties and critical access hospitals. Am. Heart J. 209, 63–67 (2019). https://doi.org/10.1016/j.ahj.2018.12.002

Schaffer, A.L., Dobbins, T.A., Pearson, S.A.: Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. BMC Med. Res. Methodol. 21(1), 1–12 (2021)

Wadhera, R.K., Joynt Maddox, K.E., Wasfy, J.H., et al.: Association of the hospital readmissions reduction program with mortality among medicare beneficiaries hospitalized for heart failure, acute myocardial infarction, and pneumonia. JAMA 320(24), 2542–2552 (2018). https://doi.org/10.1001/jama.2018.19232

Wagner, A.K., Soumerai, S.B., Zhang, F., et al.: Segmented regression analysis of interrupted time series studies in medication use research. J. Clin. Pharm. Ther. 27(4), 299–309 (2002)

Wasfy, J.H., Zigler, C.M., Chiorat, C., et al.: Readmission rates after passage of the hospital readmissions reduction program: a pre-post analysis. Annals Internal Med. 166(5), 324–331 (2017). https://doi.org/10.7326/M16-0185

West, M., Harrison, J.: Bayesian Forecasting and Dynamic Models. Springer-Verlag, New York (2006)

Wheeler, A.P.: The effect of 311 calls for service on crime in DC at microplaces. Crime Delinq. 64(14), 1882–1903 (2018)

Wadhera, R.K., Joynt, Maddox, K.E., Kazi, D.S., et al.: Hospital revisits within 30 days after discharge for medical conditions targeted by the Hospital Readmissions Reduction Program in the United States: national retrospective analysis. BMJ 366(14563) (2019). https://doi.org/10.1136/bmj.j4563

Ye, S., Wang, R., Zhang, B.: Comparison of estimation methods and sample size calculation for parameter-driven interrupted time series models with count outcomes. Health Serv. Outcomes Res. Methodol. pp 1–48 (2022)

Zigler, C.M., Papadogeorgou, G.: Bipartite causal inference with interference. Stat. Sci. Rev. J. Inst. Math. Stat. 36(1), 109 (2021)
Zuckerman, R.B., Sheingold, S.H., Orav, E.J., et al.: Readmissions, observation, and the hospital readmissions reduction program. N. Engl. J. Med. 374(16), 1543–1551 (2016). https://doi.org/10.1056/NEJMsa1513024

Publisher’s Note  Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Georgia Papadogeorgou1 · Fiammetta Menchetti2 · Christine Choirat3 · Jason H. Wasfy4 · Corwin M. Zigler5 · Fabrizia Mealli2,6

1 Department of Statistics, University of Florida, Gainesville, FL, USA
2 DiSIA and Florence Center for Data Science, University of Florence, Florence, Italy
3 Swiss Data Science Center, ETH Zürich and EPFL, Lausanne & Institute of Global Health, Faculty of medicine, University of Geneva, Geneva, Switzerland
4 Cardiology Division, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
5 Department of Statistics and Data Sciences, Department of Women’s Health, University of Texas at Austin, Austin, TX, USA
6 Department of Economics, University of Florence, DiSIA and Florence Center for Data Science, and European University Institute, Florence, Italy