Research Article

Pantelis Samartsidis*, Natasha N. Martin, Victor De Gruttola, Frank De Vocht, Sharon Hutchinson, Judith J. Lok, Amy Puenpatom, Rui Wang, Matthew Hickman and Daniela De Angelis

Evaluating the power of the causal impact method in observational studies of HCV treatment as prevention

https://doi.org/10.1515/scid-2020-0005
Received June 8, 2020; accepted February 15, 2021; published online October 11, 2021

Abstract

Objectives: The causal impact method (CIM) was recently introduced for evaluation of binary interventions using observational time-series data. The CIM is appealing for practical use as it can adjust for temporal trends and account for the potential of unobserved confounding. However, the method was initially developed for applications involving large datasets and hence its potential in small epidemiological studies is still unclear. Further, the effects that measurement error can have on the performance of the CIM have not been studied yet. The objective of this work is to investigate both of these open problems.
Methods: Motivated by an existing dataset of HCV surveillance in the UK, we perform simulation experiments to investigate the effect of several characteristics of the data on the performance of the CIM. Further, we quantify the effects of measurement error on the performance of the CIM and extend the method to deal with this problem.
Results: We identify multiple characteristics of the data that affect the ability of the CIM to detect an intervention effect including the length of time-series, the variability of the outcome and the degree of correlation between the outcome of the treated unit and the outcomes of controls. We show that measurement error can introduce biases in the estimated intervention effects and heavily reduce the power of the CIM. Using an extended CIM, some of these adverse effects can be mitigated.
Conclusions: The CIM can provide satisfactory power in public health interventions. The method may provide misleading results in the presence of measurement error.

Keywords: causal impact; causal inference; HCV; measurement error.

***Corresponding author**: **Pantelis Samartsidis**, MRC Biostatistics Unit, University of Cambridge, Cambridge, UK, E-mail: pantelis.samartsidis@mrc-bsu.cam.ac.uk
Natasha N. Martin, University of California San Diego, San Diego, USA
Victor De Gruttola, Harvard University, Cambridge, USA
Frank De Vocht and Matthew Hickman, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
Sharon Hutchinson, Glasgow Caledonian University, Glasgow, UK; and Public Health Scotland, Glasgow, Scotland
Judith J. Lok, Department of Mathematics and Statistics, Boston University, Boston, USA
Amy Puenpatom, Merck & Co., Inc., Kenilworth, NJ, USA
Rui Wang, Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, USA; and Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, USA
Daniela De Angelis, MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

Introduction

The problem of assessing the causal effect of an intervention is very frequently encountered in the fields of public health and epidemiology, see for example Rothman and Greenland (2005) and Glass et al. (2013). Randomised controlled trials have long been considered the gold standard for causal effect evaluations, but such trials may be impossible to conduct, due to either cost restrictions or ethical concerns. Therefore, researchers often rely on observational studies in order to conduct their investigations. The data from observational studies are often in the form of aggregate time-series, where the outcome of interest is measured at multiple time points before and after the intervention (e.g. incidence rate of a disease within a geographical region) and there is single treated unit (or a few treated units).

Causal inference in the setup outlined above is not straightforward. First, it is important to account for the potential of unobserved confounding using the data on the control units. For instance, assume that the outcome of all units (both treated and controls) decreases during the post-intervention period due to an unobserved environmental factor. If one ignores the data on the control units, the conclusion will be that the intervention led to the decrease in outcomes. Therefore, it is essential to adjust for the fact that control units, in particular ones whose outcomes are strongly related to the outcome of the treated unit, also showed a decrease in the post-intervention outcomes. Second, it is important to account for temporal trends in the data. For example, assume that the outcome of interest is increasing over time. A pre/post-intervention comparison of the outcome in the treated unit without accounting for this trend, will suggest erroneously that the intervention had a positive effect, even if there is no treatment effect.

To overcome these challenges, several methods have been proposed, see Samartsidis et al. (2019) for a recent review. In most of these methods, following ideas first presented in Abadie and Gardeazabal (2003), the intervention effect in the treated unit is estimated as the difference between the observed outcome in the post-intervention period and the estimate of its untreated counterfactual: the outcome that would have been observed had no intervention taken place in the treated unit. Untreated counterfactuals are estimated as follows. First, a model that expresses the relationships between the observations in the treated and control units is chosen and fit to the data in the pre-intervention period. Then, by assuming that the same model would hold in the post-intervention period in the absence of the intervention, the untreated counterfactual is estimated using the parameters estimated from the pre-intervention period and the post-intervention data in control units.

In the *causal impact method* (Brodersen et al. 2015, CIM), the model that is fit to the data in the pre-intervention period is a Bayesian structural time-series model. More specifically, the outcome of the treated unit is represented as the sum of three components: a regression component that relates the outcome on the treated unit to the outcomes on controls; a time-series component that represents temporal patterns in the data; and an error component that accounts for any unexplained variability. The regression component of the CIM can provide a safeguard against some forms of unobserved confounding.[1] The time-series component is essential to reduce biases that are purely due to temporal trends. Because of its several components, the CIM allows for extremely flexible models to be fit.

The CIM generalises several existing approaches that are used for causal inference based on time-series data. More specifically, if the data on the control units are not included as covariates in CIM's regression component, then it reduces to an *interrupted time-series* (Bernal, Cummins, and Gasparrini 2016, among others) model. If the time-series component of CIM is set to zero, then CIM is akin to synthetic-control type approaches, see e.g. Abadie, Diamond, and Hainmueller (2010), Hsiao, Ching, and Wan (2012) and Amjad, Shah, and Shen (2018).

1 This is due the fact that the CIM can be viewed as a generalisation of the synthetic control Abadie, Diamond, and Hainmueller (2010) method which, as shown by (Abadie, Diamond, and Hainmueller 2010), allows for the presence of multiple, time-constant unobserved confounders whose effect on the outcome of interest can vary over time.

Despite being only recently introduced, the CIM has been employed in several applications. Brodersen et al. (2015) use the CIM to assess how much an advertising campaign contributed to the number of visits of a website. Bruhn et al. (2017) assess the impact of pneumonococcal conjugate vaccines on pneumonia-related hospitalisations in South American countries. de Vocht et al. (2017) estimate the impact of imposing stricter alcohol licensing policies on the total number of alcohol-related hospitalisations in England. Finally, de Vocht (2016) evaluates the impact of mobile phone use on selected types of brain cancer.

Despite its strengths, there are limitations to the use of the CIM. In particular, the underlying time-series model typically includes several unknown parameters and therefore a large amount of data is required to estimate these parameters (and hence the untreated counterfactual). Further, the performance of the CIM can be affected when the outcome of interest is measured with error. These limitations were not of great concern in the aforementioned applications. However, they can possibly undermine the utility of the CIM in epidemiological applications where the amount of data is limited and/or the outcomes of interest is the prevalence of disease which cannot be measured directly, but is instead estimated based on a small sample of individuals.

In this work, we perform a series of simulation experiments to evaluate the potential of the CIM for evaluating the effectiveness of a new strategy against the hepatitis C virus (HCV) namely *treatment as prevention* (TasP). Our experiments are designed to identify the characteristics of the data that mostly affect the performance of the CIM. Further, by conducting these experiments we are able to assess the implications that inability to measure the HCV prevalence without error has on the properties of the CIM. Since our simulated data are generated following existing HCV surveillance data in the UK, we expect that our findings are indicative of the performance of the CIM in more settings where one wants to evaluate HCV treatment as prevention, as well as potentially other public health applications.

The remainder of this manuscript is structured as follows. Section 2 introduces the motivating problem. Section 3 presents a series of simulations to assess the quality of the causal estimates provided by the CIM, when the prevalence is known. Section 4 includes a simulation study to investigate the effect that estimating the prevalence based on a finite sample of individuals has on the performance of the CIM and proposes an extension to the CIM that can be used to deal with this issue. Finally, Section 5 summarises the main findings of the paper and discusses some of the strengths and limitations of our work.

Motivating dataset: HCV treatment as prevention (TasP)

HCV is a blood borne virus, a leading cause of liver disease, and one of the few causes that is curable (Williams et al. 2014) in over 90% of cases through highly effective, tolerable, short-course direct acting antiviral therapies (DAAs) (Dore and Feld 2015; Gogela et al. 2015; Walker et al. 2015). In the UK and many developing countries the majority of people infected with HCV are people who inject or have injected drugs (PWID), and more than 90% of new infections occur among PWID (De Angelis et al. 2009; Harris et al. 2019; Hutchinson et al. 2006; Prevost et al. 2015). Prevention of HCV transmission among PWID is critical to strategies to 'eliminate' HCV as a public health problem.

There is good theoretical modelling evidence that introducing and scaling up HCV treatment among those at risk of HCV transmission could reduce HCV chronic prevalence among PWID at a population level (Cousien et al. 2014; Durier, Nguyen, and White 2012; De Vos and Kretzschmar 2014; Hellard et al. 2014; Martin et al. 2011; Martin, Miners, and Vickerman et al. 2012a, Martin et al. 2012b, Martin et al. 2013a, 2013b; Martin et al. 2016a, 2016b, 2016c; Rolls et al. 2013; Vickerman, Martin, and Hickman 2011; Zeiler et al. 2010). However, there are no ongoing randomised trials of HCV TasP in the community that we know of, and direct empirical evidence is yet to emerge (Hickman et al. 2015; Martin et al. 2015). In part this has been because in most settings HCV treatment rates in PWID have been too low and surveillance data are too imprecise to detect changes in HCV transmission or chronic HCV prevalence. The current scale-up of HCV treatment in some settings compared to others provides an opportunity to establish empirical evidence, if there are sufficient data available prior and after the intervention scale-up. An additional complexity with evaluating HCV TasP

is that the outcome of interest is chronic HCV prevalence among PWID in the community, which requires ongoing surveillance of PWID. As PWID are a hidden population, there will be uncertainty in prevalence of chronic HCV, prevalence of PWID, and exposure (HCV treatment per chronically infected PWID) that will need to be addressed.

The UK has ongoing surveillance of HCV in PWID in place. For example, in Scotland the *needle exchange surveillance initiative* (NESI) has been conducted on 5 occasions (years 2008–2009, 2010, 2011–2012, 2013–2014 and 2015–2016). The estimated HCV prevalence among PWID in Tayside, Glasgow and Rest of Scotland (which averages data from 5 other sides where NESI was carried out) is shown in Table 1. We consider a setting where HCV treatment is scaled-up in Tayside, which we expect could affect subsequent HCV prevalence in that region. Our objective is to evaluate under what conditions the CIM could be used to infer the magnitude of HCV TasP in Tayside.

Evaluation of HCV TasP using the CIM

Let t index the various waves of NESI, where $t = 1$ corresponds to the 2008–2009 swap, $t = 2$ to 2010, etc, and let $i = 0, \dots,$ n index the various units, where $i = 0$ is the treated unit. In future NESI surveys ($t > 5$) it could be possible to evaluate the effect that HCV treatment scale-up had on virus prevalence by comparing $p_{0t}^{(1)}$, the prevalence at time t under the intervention in the treated site, to an estimate of the counterfactual $p_{0t}^{(0)}$, the prevalence that we would observe in the treated site if no intervention took place. That is, $\theta_t = p_{0t}^{(1)} - p_{0t}^{(0)}$, where θ_t is the causal effect of HCV TasP on prevalence at time t ($t > 5$).

The CIM makes use of the data in the pre-intervention period ($t \leq 5$) and post-intervention data in the control sites to obtain estimates $\hat{p}_{0t}^{(0)}$ of the counterfactuals for $t > 5$. It fits a Bayesian structural time-series model to the outcome in the pre-intervention period. Following standard modelling practice with prevalence data we choose to model $y_{0t}^{(0)} = \log \frac{p_{0t}^{(0)}}{1-p_{0t}^{(0)}}$ instead of $p_{0t}^{(0)}$ directly, and further assume that

$$y_{0t}^{(0)} = \mu_t + \boldsymbol{\beta}^\top \mathbf{y}_t + \epsilon_t \tag{1}$$

$$\mu_t = \mu_{t-1} + \delta_t, \tag{2}$$

Table 1: The NESI dataset.

NESI dataset summary					
Unit	**2008/9**	**2010**	**2011/12**	**2013/14**	**2015/16**
Estimated HCV prevalence					
Tayside	30.2	40.2	38.5	46.7	43.6
Greater Glasgow	66.1	63.6	60.1	65.9	60.8
Rest of Scotland	43.9	45.3	43.8	45.0	48.0
Sample size					
Tayside	189	219	117	169	195
Greater Glasgow	905	1336	858	813	812
Rest of Scotland	1335	1403	1048	1130	1320

Table presents the estimated HCV prevalence among PWID for the 3 sites and 5 occasions in which NESI was carried out. The sample size based on which these estimates were obtained is shown in the bottom panel.

where μ_t is the temporal local level component, $\boldsymbol{y}_t = (y_{1t}, y_{2t})^\top$ is the outcome (logit-prevalence) on control sites at time t,[2] $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$ are the regression coefficients, $\epsilon_t \sim \mathrm{N}(0, \sigma_\epsilon^2)$ and $\delta_t \sim \mathrm{N}(0, \sigma_\delta^2)$. In the model of Eqs. (1) and (2), the local level component μ_t induces temporal correlations in $y_{0t}^{(0)}$, and the regression component exploits correlations of $y_{0t}^{(0)}$ with outcomes on the control sites.

The parameters of model (1) and (2) are estimated using Markov chain Monte Carlo (MCMC) techniques (Brodersen et al. 2015). Posterior simulations are simplified using the following conditionally conjugate prior distributions. We let $\sigma_\epsilon^{-2}, \sigma_\delta^{-2} \sim \mathrm{Gamma}(\frac{\nu}{2}, \frac{s}{2})$. Brodersen et al. (2015) explain that ν can be thought of as prior sample size and s can be chosen such that s/ν is a guess for the variance. In practice, we can set $s = \nu(1 - R^2)\hat{\sigma}_y^2$, where $\hat{\sigma}_y^2$ is the sample variance of $y_{0t}^{(0)}$ and R^2 is the proportion of the variability in $y_{0t}^{(0)}$ we expect to be explained by the regression component.

For $\boldsymbol{\beta}$, a *spike-and-slab* prior (Chipman, George, and McCulloch 2001; George and McCulloch 1993, among others) is used. This prior assumes that for each β_i, there is a binary γ_i such that $\beta_i \neq 0$ when $\gamma_i = 1$ and $\beta_i = 0$ otherwise. We let each $\gamma_i \sim \mathrm{Bernoulli}(q_i)$, where q_i is the prior probability that the coefficient of unit i is non-zero. The expected number of units with $\gamma_i \neq 0$ using this prior is $\sum_{i=1}^n q_i$. Hence, we can set $q_i = k/n$ for all i to encourage only k control units with $\gamma_i = 1\,(\beta_i \neq 0)$. Conditionally on $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\top$, let $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ include the elements of $\boldsymbol{\beta}$ for which $\gamma_i = 1$. We assume that $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim \mathrm{N}\left(\mathbf{0}, \sigma_\epsilon^2 \Sigma_\beta\right)$, where $\boldsymbol{\Sigma}_\beta$ is some prior covariance matrix (e.g. the identity matrix). The use and careful tuning of the spike-and-slab prior is important in epidemiological applications for two reasons. Firstly, by setting some $\gamma_i = 0$ at each MCMC iteration, the method excludes controls whose data are not predictive of $y_{0t}^{(0)}$, and thus reduces the total number of parameters that need to be estimated. This is useful since the total number of pre-intervention time points is typically similar (or even smaller) than the total number of control units. Secondly, by calculating the posterior inclusion probability (the posterior mean of ρ_i) for each control unit, it allows us to identify the ones that mostly contribute to the estimation of the counterfactual.

As additional data ($t > 5$) become available, counterfactuals can be obtained by extrapolating the model of Eqs. (1) and (2). Let $\hat{\mu}_5, \hat{\boldsymbol{\beta}}, \hat{\sigma}_\epsilon^2, \hat{\sigma}_\delta^2$ be a sample from the posterior distribution of parameters $\mu_5, \boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\delta^2$, respectively. A sample from the posterior predictive distribution of $p_{06}^{(0)}$, the counterfactual prevalence in the first post-intervention survey, will be $\hat{p}_{06}^{(0)} = \frac{\exp\left(\hat{y}_{06}^{(0)}\right)}{1+\exp\left(\hat{y}_{06}^{(0)}\right)}$, where

$$\hat{y}_{06}^{(0)} = \hat{\mu}_5 + \hat{\delta}_6 + \hat{\boldsymbol{\beta}}^\top \boldsymbol{y}_6 + \hat{\epsilon}_6, \tag{3}$$

with $\hat{\delta}_6 \sim \mathrm{N}(0, \hat{\sigma}_\delta^2)$ and $\hat{\epsilon}_6 \sim \mathrm{N}(0, \hat{\sigma}_\epsilon^2)$. Assume that we draw L such samples, $\hat{p}_{06,\ell}^{(0)}$ ($\ell = 1, \ldots, L$). Then, L samples from the posterior distribution of the causal effect at $t = 6$ will follow as $\hat{\theta}_{6,\ell} = p_{06}^{(1)} - \hat{p}_{06,\ell}^{(0)}$, from which we obtain a point estimate (the mean of $\hat{\theta}_{6,\ell}$) and a credible interval (the 2.5 and 97.5% percentiles of $\hat{\theta}_{6,\ell}$).

Evaluating the CIM using the HCV TasP dataset

Setting

Our objective is to assess the potential of the CIM for estimating the effect of HCV TasP using the existing UK HCV data (Section 2) combined with post-intervention data that will be collected. More specifically, we investigate the performance of the estimator of the causal intervention effect provided by the CIM method and identify the characteristics of the data that most affect the quality of the estimates of $\hat{\theta}_t$. Our evaluation

2 Note that for the control units $i > 0$, we do not need to define both $y_{it}^{(0)}$ and $y_{it}^{(1)}$ since these units are not subject to the intervention: for the controls units we have that $y_{it} = y_{it}^{(0)}$ for every t.

will also inform the potential of CIM for similar datasets. To achieve these goals, we performed a series of simulations.

First, we note that the performance of the estimator of θ_t depends solely on the performance of the estimator of $\hat{p}_{0t}^{(0)}$, since $\theta_t = p_{0t}^{(1)} - p_{0t}^{(0)}$ and $p_{0t}^{(1)}$ is observed. More specifically, if $\hat{p}_{0t}^{(0)}$ is an unbiased estimate of $p_{0t}^{(0)}$, then $\hat{\theta}_t$ is an unbiased estimate of θ_t. Furthermore, then the 95% CI of θ_t will also include the true intervention effect if and only if the 95% CI of $p_{0t}^{(0)}$ includes the untreated counterfactual. Therefore, it suffices to evaluate $\hat{p}_{0t}^{(0)}$. We did this by considering the following performance measures at each post-intervention time point: (i) the mean (over simulated datasets) of the prediction error (MPE),[3] where in each simulated dataset the prediction error is defined as the difference $p_{0t}^{(0)} - \hat{p}_{0t}^{(0)}$; (ii) the standard deviation (over simulated datasets) of the prediction error (sd-PE)); (iii) the mean (over simulated datasets) width of credible intervals of $p_{0t}^{(0)}$ (CIW); (iv) the % false discovery rate (over simulated datasets) (FDR), where in each dataset a false detection occurred when $p_{0t}^{(1)} = p_{0t}^{(0)}$ (i.e. $\theta_t = 0$) and the 95% CI of $p_{0t}^{(0)}$ did not include $p_{0t}^{(1)}$; and (v) the % detection rate (over simulated datasets) (power), where in each dataset a detection occurred when $p_{0t}^{(1)} < p_{0t}^{(0)}$ (i.e. the intervention reduced prevalence) and the lower bound of the 95% CI of $p_{0t}^{(0)}$ was higher than $p_{0t}^{(1)}$.

We simulated 10,000 datasets, each consisting of HCV logit-prevalence measurements y_{it} for $n+1$ units and T times points. At each time point t, we drew $\boldsymbol{y}_t = \left(y_{0t}^{(0)}, y_{1t}, \ldots, y_{nt}\right)^\top \sim \text{MVN}(\boldsymbol{m}, \boldsymbol{S})$ that is the mean, variance and correlation of the outcomes remained constant over time. We drew the elements of $\boldsymbol{m}$ from a Uniform $(m_{\min}, m_{\max})$; $m_{\min}/m_{\max}$ is the minimum/maximum logit-prevalence found in the NESI dataset presented in Section 2. We set $\boldsymbol{S}_{11} = \sigma_y^2$. The remaining diagonal elements of $\boldsymbol{S}$ are drawn from a Uniform $\left(s_{\min}^2, s_{\max}^2\right)$; $s_{\min}^2/s_{\max}^2$ is the minimum/maximum over all i of s_i^2, where s_i^2 is the sample variance of the time-series of unit i in the NESI dataset. To obtain the off-diagonal elements of $\boldsymbol{S}$, it suffices to pick the values of ρ_{ij}, the degree of correlation between the data on units i and j, where $i, j \in \{0, \ldots, n\}$ and $i \neq j$. We set $\rho_{0j} = \rho$ for all $1 \leq j \leq k_2$ and $\rho_{0j} = 0$ when $j > k_2$. That is, the treated unit is only correlated to the first k_2 controls units. Further, for all i, j such that $1 \leq i, j \leq k_2$ and $i \neq j$, we set $\rho_{ij} = 0.8\rho$. Finally, for $i > k_2$ (i.e. the $k_1 = n - k_2$ control units that are not correlated to the treated unit), we have that $\rho_{ij} = 0$ for all $j \neq i$ i.e. these units are not correlated to any other unit in the dataset. For each simulated dataset, we introduced intervention effects (i.e. obtained $p_{0t}^{(1)}$) by reducing $p_{0t}^{(0)}$ in the post-intervention period by a certain %. More specifically, we introduced 21 different effects from 0 to 50% with increments of 2.5%.

We attempted to generate data that mimic the HCV TasP application of Section 2. Therefore, we used the following simulation parameters in our *baseline* setup. The variance σ_y^2 was set equal to the variance of the logit-prevalence measurements of the treated unit in the motivating dataset. Let $T = t_1 + t_2$, where t_1 is the total number of pre-intervention data points per unit and t_2 is the total number of post-intervention observations. We set t_1 to be 6 and 12, $t_2 = 3$ and $n = 8$. In practical applications, we expect that only a small proportion of the control units to be correlated with the treated unit. Hence, in the baseline simulation we set $k_1 = 6$ and $k_2 = 2$. For the k_2 'useful' controls, we assumed that $\rho = 0.8$.

In order to identify the features of the data that mostly affect the quality of causal estimates provided by the CIM, we performed several sensitivity analyses. In each sensitivity analysis we repeated the baseline simulation altering a single characteristic of the dataset and re-evaluated the five performance measures. The characteristics that we considered are (I) the variability of the outcome, of the treated unit σ_y^2; (II) the total number of observations in the pre-intervention period, t_1; (III) the total number of control units whose outcomes are not correlated with the outcome of the treated unit, k_1; (IV) the total number of useful controls, k_2; (V) the level of correlation between y_{0t} and the outcomes of the useful controls, ρ; and (VI) the hyperparameters of the spike-and-slab prior on the regression coefficients $\boldsymbol{\beta}$. The values that we use for characteristics I–V are shown in Table 2.

3 We choose not to use the term 'bias' because the estimand is different for each dataset.

Table 2: Feature values used for the sensitivity analyses.

Characteristic	Values considered
σ_y^2	0.005, 0.04, **0.75**
t_1	**6**, 9, **12**, 24
k_1	**6**, 12, 24
k_2	**2**, 4, 6
ρ	0.6, 0.7, **0.8**
Prior	Uninformative, **Calibrated**

The values used in the baseline simulations appear in bold.

Results

The MPE, sd-PE, CIW and FDR for the baseline simulations are shown in Table 4 of Appendix A. As can be seen in Table 4, the MPE at each post-intervention time point in the baseline setup was negligible (compared to the sd(MPE)), for both $t_1 = 6$ and $t_1 = 12$. This fact implies that, over the 10,000 simulated datasets, the estimates $\hat{p}_{0t}^{(0)}$ coincided on average with the corresponding 'true' values $p_{0t}^{(0)}$. It is confirmed in Figure 4 of Appendix A, where we plot the simulated values of $p_{0t}^{(0)}$ against the estimated causal effect $\hat{\theta}_t$. However, we see that there is positive correlation between $p_{0t}^{(0)}$ and $\hat{\theta}_t$ i.e. the effect of the intervention is overestimated when the prevalence in the treated unit is high and underestimated when it is low. This correlation is expected due to the use of the logit transformation and drops with higher t_1. We also see that the FDR is very close to the nominal 5% for $t_1 = 12$, and slightly inflated for $t_1 = 6$.

The power that we obtained at each $t > t_1$ in the baseline simulations can be seen in Figure 1. As expected, the power increased with the % decrease in prevalence due to the intervention, and reached 100% when the intervention reduced prevalence by half. Lower drops in prevalence were associated with lower power. For example, a 10% decrease in HCV prevalence is only detected with probability 25 and 30% for $t_1 = 6$ and $t_1 = 12$, respectively. The power achieved was comparable at all three post-intervention time points. Nonetheless, it

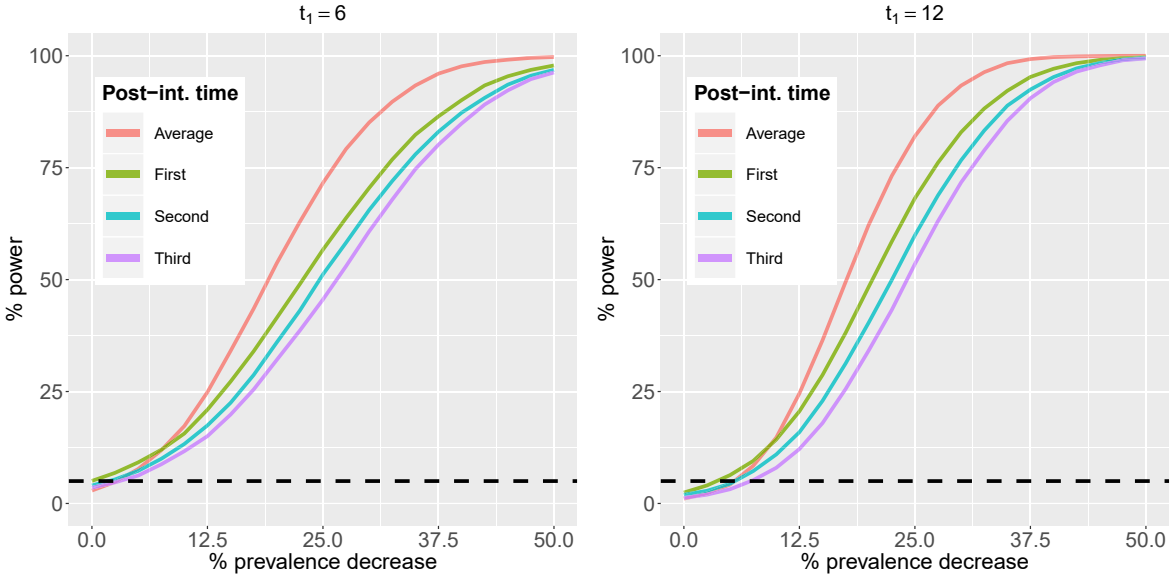

Figure 1: Baseline simulations results. The plots shows the power of detecting an intervention effect obtained by the CIM, as a function of the intervention effect magnitude. The left panel shows results for $t_1 = 6$ and the right panel for t_1=12. All results are based on 10,000 simulated datasets.

decreased with t. This decrease is due to fact that the variance of the random walk component is $(t - t_1)\sigma_\delta^2$ $(t > t_1)$ which leads to wider credible intervals as t increases. Generally, in practical applications we expect that the uncertainty in the estimates of $p_{0t}^{(0)}$ provided by the CIM will increase with t unless the time-series component has no contribution (this could be the case, for example, when all of the variability is attributed to the regression component).

One of the advantages of the Bayesian approach is that several quantities of interest can be calculated directly from the posterior distribution of the model parameters. For example, rather than testing if θ_t is zero at each $t > t_1$, one can use a summary to test for an overall effect. We examined the average causal effect in the post-intervention period defined as

$$\vartheta = \frac{1}{t_2} \sum_{t=t_1+1}^{T} \theta_t. \tag{4}$$

A credible interval for ϑ that excluded zero was considered as evidence of an overall intervention effect in the entire post-intervention period. Figure 1 presents the power that we obtained when we tested an overall intervention effect, when this effect was constant (i.e. $\theta_t = \theta$ for all $t > t_1$). As expected, there were big gains in power when we summarised the information across all $t_2 = 3$ post-intervention times. For example, for $t_1 = 6$, a prevalence decrease of 20% was detected with probability 80% when we used ϑ to test for it but only with probability 60% when we examined each post-intervention time point individually. Hence, in practical applications, it is worth monitoring the outcome of the treated units on multiple time points after the intervention is introduced, as this can increase the chances to detect an intervention effect. Moreover, it might be worth considering the average effect only in the last $s < t_2$ post-intervention time points since some interventions might not be effective immediately after introduction.

The results of our sensitivity analyses are summarised in Table 4 of Appendix A, Figure 2 and Figures 5 and 6 of Appendix A. Figures 2, 5, and 6 plot the power achieved by the CIM against the magnitude of the intervention effect at post-intervention times $t_1 + 1$, $t_1 + 2$ and $t_1 + 3$, respectively. The MPE was negligible (compared to its standard error) across all sensitivity analyses and therefore is not further discussed. For the remaining performance measures, the results that we obtained for $t = t_1 + 1$ were similar to the results obtained for $t = t_1 + 2$ and $t = t_1 + 3$. Hence, for the remainder of this section we focus attention to the first post-intervention time point.

The value of t_1 largely affected the performance of the CIM. As expected, increasing t_1 caused all sd-PE and CIW to decrease (Figure 2(a)), since the parameters were estimated with higher accuracy. Further, the FDR was inflated for low values of t_1. One possible explanation is that when t_1 was low, it was more likely to observe strong correlations between the treated and a control unit by chance, thus assigning non-zero regression coefficients to control units whose outcomes were not truly correlated to the outcome of the treated unit. As expected, the variance of the outcome σ_y^2 was crucial for the performance of the CIM. Larger outcome variance led to larger sd-PE and CIW (Table 4), and to a substantial drop in power (Figure 2(b)). Nonetheless, for fixed t_1, the values of the FDR were similar across all values of σ_y^2 considered.

The sd-MPE, CIW and power were not very sensitive the total number of 'unrelated' controls k_1 (Figure 2(c)). With increasing k_1, sd-PE and CIW increased, whereas the power dropped. The reason could be that there was a need to estimate more regression parameters as k_1 increased. This effect was more prominent when $t_1 = 6$. However, this drop in performance was negligible. We believe that this robustness to the addition of controls whose outcomes are not informative of the outcome of the treated unit is due to the spike-and-slab prior which successfully identifies these controls and, on average, set their coefficients to zero. This finding suggests that in real problems, since the expected drop in power is negligible, it is preferable to include all the available control units and allow the CIM to identify the ones that are important.

Increasing k_2 slightly improved power but the gains were small, since each additional control could only explain a small proportion of the variability in $y_{0t}^{(0)}$ that was not already explained by the existing 'useful' controls. Another factor that affected the quality of the causal estimates was the level of correlation between the outcome of the treated unit and the outcomes of 'useful' controls. This is expected since the method uses the regression component to exploit linear relationships in the data. Therefore, the stronger

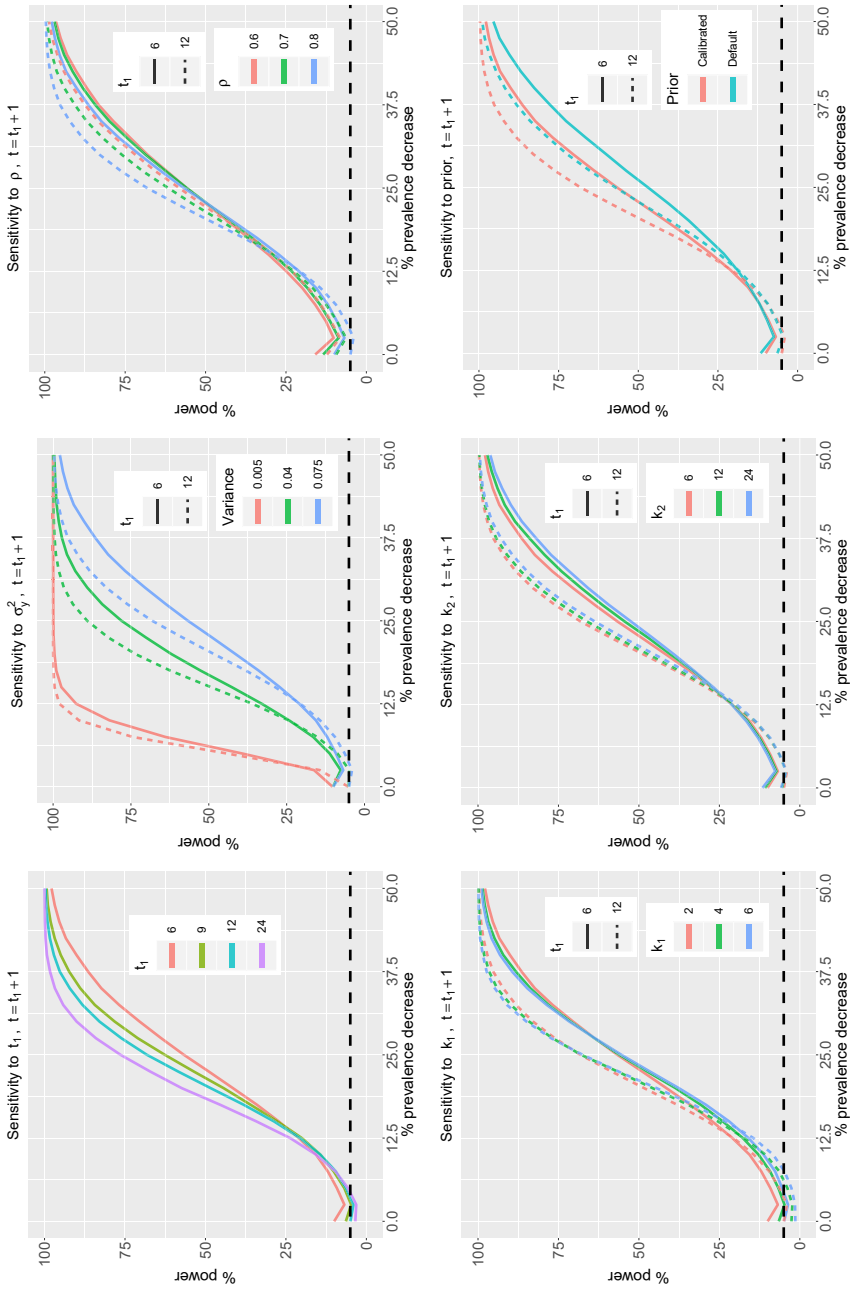

Figure 2: Results of the simulation study for the second post-intervention time point $t = t_1 + 1$. The figure presents the power for detecting an intervention effect achieved by the CIM as a function of the intervention effect magnitude, in all baseline settings and sensitivity analyses. All results are based on 10,000 simulated datasets.

these relationships were, the higher was the proportion of the variability of $y_{0t}^{(0)}$ explained by the regression component. As a result, sd-PE and CIW decreased, leading to an increased power. For small intervention effects, satisfactory power was only achieved for large values of ρ (Figure 2(e)).

Our final sensitivity analysis aimed to demonstrate the effects of the specification of the prior. Therefore, we repeated the baseline simulations, using a different spike-and-slab prior for the regression coefficients $\boldsymbol{\beta}$. To this end, the informative spike-and-slab prior presented in Section 2 was replaced by the software default. Figure 2(f) presents the power under the two prior distributions. The power to detect an effect substantially dropped under the software default prior. This was the case because this prior was less informative compared to the prior that we initially used, and thus led to greater posterior uncertainty and therefore wider credible intervals for the untreated counterfactuals.

Measuring the outcome with error

Effect on the performance of the CIM

The CIM assumes that the outcome of interest is observed without error in both the treated and the control units. This assumption is plausible in many real life problems, e.g. the one considered by Brodersen et al. (2015) where the outcome of interest is the total number of daily visits in various web-pages (the units) which can be precisely enumerated. Other examples of outcomes that can be measured without error (or estimated very precisely) include the daily sales of a product in a geographical region, the total number of deaths due to a disease in a hospital and the annual GDP of a country. However, in many epidemiological studies, it might not possible to observe the outcome without some error. For example, in the motivating application of Section 2, the true HCV prevalence in each unit is unknown and it is estimated through surveillance data as $\tilde{p}_{it} = \frac{k_{it}}{N_{it}}$, where k_{it} and N_{it} represent the total number of infected individuals and the total sample size from the surveillance study in unit i at time t, respectively. Note that we refer to $\tilde{p}_{it}$ as imprecise prevalence in order to distinguish it from the estimated prevalence $\hat{p}_{it}^{(0)}$ obtained from the CIM.

To assess the impact that the use of imprecise outcomes (instead of the true, unknown outcomes) had on the performance of the CIM, we re-analysed the same 10,000 simulated datasets that we analysed for the baseline simulation of Section 3, when $t_1 = 12$ and $t_1 = 24$. Instead of implementing the CIM to $y_{it} = \log \frac{p_{it}}{1-p_{it}}$, we implemented it to $\tilde{y}_{it} = \log \frac{\tilde{p}_{it}}{1-\tilde{p}_{it}}$, where $\tilde{p}_{it} = \frac{k_{it}}{N_{it}}$ and k_{it} were simulated from a Bin (N_{it}, p_{it}) distribution. We evaluated the performance considering the same performance measures as in Section 3. We only present the power at the first post-intervention time point because we found that the results were very similar in the remaining post-intervention time points. We artificially introduced the intervention effects that were non-zero, by drawing the k_{0t} from a Bin $\left(N_{it}, p_{0t}^{*}\right)$, where p_{0t}^{*} were obtained by reducing the original prevalence p_{0t}. The sample size across units and time points was constant, i.e. $N_{it} = \mathrm{n}$ for all units i and times t. We simulated $\mathrm{n} = 50$, $\mathrm{n} = 50$ and $\mathrm{n} = 150$.

Table 3 (sd-PE, CIW and FDR) and Figure 3 (power) summarise the results for $t_1 = 12$; the results for $t_1 = 24$ are similar and therefore not shown. For comparison, we also show the results that we obtained in the baseline simulation when we assumed outcomes were measured without error. As expected, the use of the imprecise outcomes $\tilde{p}_{it}$ instead of the perfectly measured outcomes p_{it} degrades the performance of the CIM substantially. Table 3 shows that both the sd-PE and CIW increase when the CIM is implemented using $\tilde{p}_{it}$. For example, the CIW obtained when the sample size $\mathrm{n} = 50$, was approximately twice the width that we obtained using the original data. As a result, there was also reduced power to detect an intervention effect (Figure 3). For instance, the power to detect a 25% decrease using the approximated outcomes and $\mathrm{n} = 50$ was roughly 37%, as opposed to 75% for the exact prevalence outcomes.

The increase in uncertainty (and therefore loss of power) occurred because $\tilde{y}_{it}$ are noisy observations of y_{it} and therefore the correlations between $\tilde{y}_{0t}$ and $\tilde{y}_{it}$ $(i > 0)$ were weaker than the correlations between y_{0t} and y_{it} in the original simulation study. As a result, the estimates of regression coefficients of the k_1 predictive control

Table 3: Effect of measurement error on the performance of the CIM.

Method	n		sd-PE			CIW			FDR	
CIM	∞	4.30	4.26	4.26	0.17	0.18	0.19	0.059	0.047	0.043
CIM	50	7.82	7.78	7.70	0.29	0.30	0.32	0.173	0.153	0.136
CIM	100	7.00	7.04	7.03	0.25	0.26	0.27	0.161	0.142	0.130
CIM	150	6.63	6.57	6.67	0.23	0.24	0.25	0.154	0.139	0.122
EIV	50	6.73	6.72	6.67	0.37	0.37	0.37	0.057	0.055	0.049
EIV	100	6.25	6.26	6.26	0.29	0.29	0.29	0.075	0.069	0.074
EIV	150	6.00	5.97	5.98	0.25	0.25	0.26	0.088	0.084	0.083

The table presents standard deviation of the prediction error (sd-MPE), mean credible interval width (CIW), and false discovery rate (FDR) in the baseline setting with t_1=12, when the CIM and CIM-EIV methods are applied to the imprecise outcomes $\tilde{p}_{0t}$. For reference, we show results for the CIM implemented to the true outcomes p_{0t} (CIM, n=∞). For each performance measure, the three columns correspond to the three post-intervention time points. The values of sd-PE are multiplied by 10^2. Results are based on 10,000 simulated datasets.

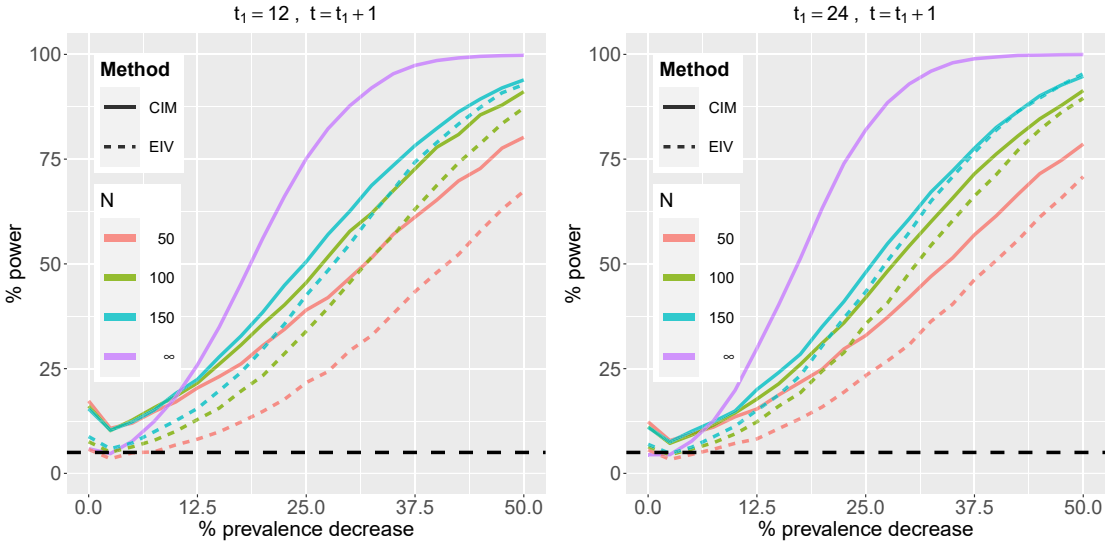

Figure 3: Effect of measurement error on the performance of the CIM. The figure presents the power achieved for $t = t_1 + 1$ when the CIM and CIM-EIV are applied to the imprecise outcomes $\tilde{p}_{0t}$. For reference, we show results for the CIM when implemented to the true outcomes (CIM, n = ∞). The left and right panels correspond to $t_1 = 12$ and $t_1 = 24$, respectively. Results are based on 10,000 simulated datasets.

units were biased downwards and the estimates of σ_ϵ^2 were biased upwards. In the classic linear regression setting this phenomenon is known as regression dilution, see e.g. Frost and Thompson (2000).

In addition to the increased uncertainty in the estimates of the causal effect, the use of imprecise measurements also led to an increased FDR (Table 3). More specifically, for n = 50, the FDR at $t = t_1 + 1$ was roughly 16%, more than triple the desired nominal level of 5%. A potential explanation is that some control units appeared to be highly correlated with the treated unit in the pre-intervention period by chance, because of the error in $\tilde{p}_{it}$. As a result, the coefficients of these units were over-estimated, leading to inaccurate prediction of the untreated counterfactual in the post-intervention period.

Both of these problems, i.e. increased uncertainty in the causal estimate and increased false positive rate, became more profound when the sample size n was reduced. The reason is that as n decreased, the $\tilde{y}_{it}$ become more variable (i.e. the measurement error increased).

An errors-in-variables causal impact method

The simulation study of Section 4.1 shows that measurement error has an adverse impact on the performance of the CIM, reducing power and increasing the false positive rate. The former is expected and inevitable when it is not possible to measure the outcome precisely. However, an increased false positive rate is an undesirable property which can reduce the reliability of a significant finding obtained using the CIM, especially when the estimated intervention effect is small. In this section, we extend the CIM in an attempt to deal with this problem.

We propose a two-level Bayesian hierarchical. At the first level, we have the data. Let $k_{0t}^{(0)}$ be the total number of infected individuals in the treated units when there is no intervention. We have that $k_{0t}^{(0)} = k_{0t}$ when $t \leq t_1$ and missing for $t > t_1$. Further, let $k_{it}^{(1)} = k_{it}$ $(t > t_1)$ be the total number of infected individuals in the treated unit when the intervention is in effect. We assume that

$$\begin{aligned} k_{0t}^{(0)} &\sim \text{Bin}\left(N_{0t}, \frac{\exp\left(y_{0t}^{(0)}\right)}{1+\exp\left(y_{0t}^{(0)}\right)}\right), \\ k_{0t}^{(1)} &\sim \text{Bin}\left(N_{0t}, p_{0t}^{(1)}\right) \quad (t > t_1), \\ k_{it} &\sim \text{Bin}\left(N_{it}, \frac{\exp\left(y_{it}\right)}{1+\exp\left(y_{it}\right)}\right) \quad (i > 0). \end{aligned} \tag{5}$$

Equation (5) relates the logit-prevalence to the observed data thus acknowledging that there is uncertainty regarding its true value. The smaller N_{it} is, the larger the uncertainty regarding the true value of y_{it}. Depending on the application, one might need to adopt observation Eq. (5) in order to account for more complex relationships between the observed data and the logit-prevalence (e.g. when there are data from multiple sub-populations of individuals).

At the second level, we have the unknown prevalence parameters. Similar to the CIM, we assume that the untreated logit-prevalence $y_{0t}^{(0)}$ in the treated unit can be written as

$$y_{0t}^{(0)} = \alpha_t + \boldsymbol{\beta}^\top \boldsymbol{y}_t + \varepsilon_t, \tag{6}$$

where $\varepsilon_t \sim \text{Normal}(0, \sigma_\varepsilon^2)$ for all t. For the treated prevalence $p_{0t}^{(1)}$ in the treated unit we *a priori* assume that $p_{0t}^{(1)} \sim \text{Beta}(1,1)$ for all $t > t_1$. For σ_ϵ^2 and $\boldsymbol{\beta}$ we use the same prior specifications as in Section 2.1. The intercept α_t arises from an AR(1) process i.e.

$$\alpha_t = \mu + \phi(\alpha_{t-1} - \mu) + \eta_t, \tag{7}$$

where $\phi \in (-1,1)$ is the persistent parameter and $\eta_t \sim \text{N}(0, \sigma_\eta^2)$. For the AR hyperparameters μ, σ_η^2 and ϕ we use similar priors as Kastner and Frühwirth-Schnatter (2014). More specifically, we let $\mu \sim \text{n}(0, 10^3)$, $\sigma_\eta^2 \sim \text{Gamma}(0.5, \frac{0.5}{(1-R^2)\hat{\sigma}_{y_0}^2})$ and $\frac{\phi+1}{2} \sim \text{Beta}(1,1)$, where R^2 and $\hat{\sigma}_y^2$ are defined as in Section 2.1.

Samples $\theta_{t,\ell}$ $(\ell = 1, \ldots, L$ and $t > t_1)$ from the posterior distribution of the causal effects are obtained as $p_{0t,\ell}^{(1)} - p_{0t,\ell}^{(0)}$. The $p_{0t,\ell}^{(1)}$ are drawn from their $\text{Beta}(1 + k_{0t}, 1 + \text{n}_{0t} - k_{0t})$ posterior distributions. The $p_{0t,\ell}^{(0)}$ are drawn from their posterior predictive distributions via MCMC. The proposed algorithm is a block Gibbs sampler that is on each iteration, one parameter (or block of parameters) is drawn from its full conditional distribution given the remaining parameters and data. The indicator variables γ_i are drawn one at a time, see e.g. Sutton (2020). The AR hyperparameters μ, σ_η^2 and ϕ are jointly updated using a Metropolis-Hastings step (Kastner and Frühwirth-Schnatter 2014). The unknown logit-prevalence $\boldsymbol{y}_i = \left(y_{i1}, \ldots, y_{it_1}\right)^\top$ are drawn

one at a time from their normal full conditionals; for this to be possible we make use of the Pólya–Gamma representation of the Binomial likelihood as proposed by Polson, Scott, and Windle (2013). The remaining model parameters $\boldsymbol{\alpha} = \left(\alpha_1, \ldots, \alpha_{t_1}\right)^\top$, $\boldsymbol{\beta}$ and σ_ε^2 have conjugate prior distributions and are therefore easy to update. The code that we used has been made publicly available.[4]

In the model of Eqs. (5)–(7), the covariates $\boldsymbol{y}_t$ are random variables. Hence, the model is similar in spirit to errors-in-variables (EIV) models often used to deal with the problem of regression dilution in practice, see for example Dellaportas and Stephens (1995). We therefore refer to this model as the CIM-EIV approach. However, it is more general than an EIV model as it allows for the response variable y_{0t} to be measured with error as well.

Application to the simulated data

We applied the proposed CIM-EIV method to the data that we simulated for the experiment of Section 4.1. We used the same R^2 and $\hat{\sigma}_y^2$ (for the priors on the variance parameters) as we did for the CIM. The prior distributions for the spike-and-slab parameters were the same as for the CIM with the exception that $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ was set to $10^3\boldsymbol{I}$.

The sd-PE, CIW and FDR are presented in Table 3. We see that for fixed n, the sd-PE obtained by the CIM-EIV was lower compared to the one obtained by the CIM. However, the proposed method successfully adjusted for the uncertainty regarding the true values of the prevalence thus leading to wider credible intervals. As a result, we see that the proposed EIV approach reduced the FDR compared to the CIM, and that the benefits were more apparent when n was small. When we increased n, the magnitude of the difference $\tilde{y}_{it} - y_{it}$ decreased and therefore the CIM-EIV did not improve much compared to the CIM (whose performance in terms of the FDR was already satisfactory). Therefore, we recommend that the CIM-EIV method is used especially in cases where the problem of dilution is expected to be high.

Note that the power of the CIM-EIV method was lower compared to the power of the CIM (see Figure 3). This is expected, since the CIM-EIV relaxes the assumption of the CIM that the outcomes are known precisely. In order to increase the power, one can combine post-intervention time-points as explained in Section 3.

Discussion

Main findings

Using an HCV treatment as prevention intervention as a case study, our paper presents a series of simulations studies to investigate the potential of the CIM for use in observational epidemiological/public health studies aiming to estimate the causal effect of an intervention on an outcome of interest using aggregate time-series observational data. Overall, our experiments show that if the untreated outcome of the treated unit is linearly related to the (untreated) outcomes of some of the controls units and the effect of the intervention is effective, then the method will provide satisfactory power. We have found that the main characteristics of the data that affect the ability of the CIM to detect a non-zero intervention effect are the length of the time-series in the pre-intervention period, the variability of the outcome and the strength of the linear relationships between the pre-intervention data on the treated unit and the control units.

This work has demonstrated some of the potential merits of adopting a Bayesian approach for this problem. In particular, we have shown that it is possible to improve power by summarising information from all post-intervention time points rather than considering each one separately. Moreover, our simulation experiments suggest that if the prior distributions for the CIM model parameters are not chosen carefully

4 https://osf.io/4cwps/?view_only=0f6071a38d5e472dbcabc20d99dcb2e6.

then the method may provide misleading results. Finally, we have studied the implications of prevalence being measured with error on the performance of the CIM. Specifically, our simulations show that when the prevalence is estimated based on a small sample of individuals, the power of the method drops substantially and the false positives rates are inflated. In such cases, it might be preferable to use the proposed CIM-EIV approach.

Our work has important implications for HCV elimination initiatives and HCV TasP researchers. Theoretical modelling studies have shown the substantial potential benefits of scaled-up HCV treatment for PWID on reducing HCV chronic prevalence and incidence (Cousien et al. 2014; Durier, Nguyen, and White 2012; De Vos and Kretzschmar 2014; Hellard et al. 2014; Martin et al. 2011; Martin, Miners, and Vickerman 2012a; Martin et al. 2013a, 2013b; Martin et al. 2016a, 2016b, 2016c; Rolls et al. 2013; Vickerman, Martin, and Hickman 2011; Zeiler et al. 2010). However, empirical studies are needed to confirm that HCV treatment as prevention expansion can yield population declines in prevalence and incidence. As randomized controlled trials testing HCV TasP may be logistically difficult, prohibitively expensive, or ethically questionable, observational studies may provide alternative evidence for a TasP effect. Our findings, that the CIM method is a robust method for detecting a TasP intervention effect using surveillance data from the UK, provide an important methodological tool for use in empirical evaluations of HCV TasP using observational data. Indeed, ongoing observational studies of HCV treatment expansion among PWID such as occurring in Dundee and across the UK as part of the EPiTOPE (Hickman et al. 2019) study will, when combined with CIM methods, shed important new information on the effectiveness of HCV TasP in the real-world.

Since our simulated data have been generated based on an existing UK HCV dataset regarding the effectiveness of TasP against the HCV, we expect that our conclusions will be relevant to other public health applications in clinical practice.

Limitations and future research

This work has limitations. First, in our simulation experiments we have assumed that the mean untreated logit-prevalence remains constant over time in both control and treated units. Further, we have assumed that the correlation between the logit-prevalence of the treated unit and the logit-prevalence of control units lalso remains constant. Hence, in the future, it is worth studying the performance of the CIM (in terms of both bias and power) under data generating mechanisms where these assumptions do not hold. This could be done, for example, by introducing a declining trend in a subset of the units. Second, in future research, it is worth comparing the performance of the CIM with other existing methodologies, such as difference-in-differences and generalised linear mixed models, since the results from existing comparative studies (Gobillon and Magnac 2016; Kinn 2018; O'Neill et al. 2016; O'Neill et al. 2020) may not generalise to the type of data that we consider.

There are many ways in which the proposed CIM-EIV approach can be improved. One idea is to account for the fact the unknown prevalence in control units is likely to show serial correlations. For example, one could assume that the logit-prevalence in control units is an AR(1) process. Another option is to account for correlations between controls units. Both of these extensions are likely to improve the precision of the causal estimates provided by the method.

Finally, we note that we use UK surveillance data to construct our case study, which incorporates regular, routine surveillance among PWID. In many settings, surveillance among PWID occurs more sporadically, or among fewer sites, or does not occur at all. In these settings, CIM methods may not generate sufficient power to detect an intervention effect, or the observational period may need to be lengthened. Further studies in different settings with alternative surveillance systems are warranted.

Research funding: This study was funded by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA, and the National Institute for Health Research (NIHR) Programme Grants for Applied Research programme (Grant Reference Number RP-PG-0616-20008). The study was further supported by the National Institute for Health Research Health Protection Unit on Evaluation of Interventions. NNM and VDG

were partially supported from the San Diego Center for AIDS Research (SD CFAR), an NIH-funded program (P30 AI036214). JJL acknowledges NIH funding from NIH/NIAID R01 AI10072 and NSF funding from NSF DMS 1854934. RW acknowledges support R01 AI136947 from the National Institute of Allergy and Infectious Disease (NIAID). DDA was funded by the UK Medical Research Council grant MC_UU_00002/11. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Appendix A: Supplementary simulation results

In this section we provide further results for the simulation study of Section 3. Table 4 presents mean prediction error (MPE) of the causal estimates, standard error of the PE (sd-PE), mean credible interval width (CIW) and false discovery rate (FDR) in all baseline settings and sensitivity analyses. Figure 4 shows simulated $p_{0t}^{(0)}$ against estimates $\hat{\theta}_t$ obtained in baseline simulations, for all three post-intervention times. Figures 5 and 6 show the power achieved by the CIM in all baseline settings and sensitivity analyses, at $t_1 + 2$ and $t_1 + 3$, respectively.

Table 4: Simulation results for Section 3.

t_1		MPE			sd-PE			CIW			FDR	
t_1=6*	−0.35	0.70	0.37	11.36	11.30	11.33	0.211	0.226	0.239	0.100	0.086	0.071
t_1=9	−0.31	0.61	0.40	11.61	11.52	11.59	0.196	0.213	0.226	0.064	0.052	0.038
t_1=12*	0.04	0.44	0.50	11.73	11.69	11.76	0.185	0.202	0.217	0.050	0.038	0.027
t_1=24	−0.26	0.12	0.32	11.96	11.93	11.97	0.167	0.184	0.199	0.034	0.023	0.013
σ_y^2		MPE			sd-PE			CIW			FDR	
0.005 (t_1=6)	−0.06	0.23	0.17	10.88	10.86	10.87	0.056	0.060	0.063	0.103	0.088	0.073
0.005 (t_1=12)	0.04	0.17	0.19	10.90	10.89	10.91	0.049	0.054	0.057	0.052	0.039	0.028
0.04 (t_1=6)	−0.24	0.55	0.32	11.13	11.08	11.11	0.156	0.167	0.177	0.101	0.086	0.071
0.04 (t_1=12)	0.08	0.37	0.44	11.33	11.30	11.36	0.137	0.150	0.160	0.050	0.038	0.027
0.075 (t_1=6)*	−0.35	0.70	0.37	11.36	11.30	11.33	0.211	0.226	0.239	0.100	0.086	0.071
0.075 (t_1=12)*	0.04	0.44	0.50	11.73	11.69	11.76	0.185	0.202	0.217	0.050	0.038	0.027
k_1		MPE			sd-PE			CIW			FDR	
6 (t_1=6)*	−0.35	0.70	0.37	11.36	11.30	11.33	0.211	0.226	0.239	0.100	0.086	0.071
6 (t_1=12)*	0.04	0.44	0.50	11.73	11.69	11.76	0.185	0.202	0.217	0.050	0.038	0.027
12 (t_1=6)	−0.28	0.63	0.20	11.28	11.23	11.26	0.217	0.232	0.245	0.108	0.094	0.077
12 (t_1=12)	0.02	0.57	0.39	11.68	11.64	11.69	0.190	0.206	0.220	0.055	0.040	0.030
24 (t_1=6)	−0.19	0.72	0.13	11.22	11.17	11.19	0.223	0.237	0.250	0.116	0.103	0.085
24 (t_1=12)	0.05	0.65	0.40	11.60	11.56	11.61	0.195	0.212	0.225	0.058	0.044	0.034
k_2		MPE			sd-PE			CIW			FDR	
2 (t_1=6)*	−0.35	0.70	0.37	11.36	11.30	11.33	0.211	0.226	0.239	0.100	0.086	0.071
2 (t_1=12)*	0.04	0.44	0.50	11.73	11.69	11.76	0.185	0.202	0.217	0.050	0.038	0.027
4 (t_1=6)	−0.21	0.42	0.35	11.35	11.28	11.31	0.213	0.228	0.240	0.065	0.059	0.047
4 (t_1=12)	0.50	0.28	0.20	11.71	11.65	11.72	0.188	0.204	0.217	0.026	0.019	0.013
6 (t_1=6)	−0.18	0.38	0.34	11.35	11.27	11.31	0.214	0.229	0.241	0.049	0.043	0.035
6 (t_1=12)	0.02	0.20	−0.12	11.70	11.61	11.71	0.190	0.206	0.219	0.014	0.012	0.007
ρ		MPE			sd-PE			CIW			FDR	
0.6 (t_1=6)	−0.16	0.68	0.37	11.34	11.31	11.28	0.211	0.227	0.240	0.159	0.137	0.113
0.6 (t_1=12)	0.31	0.57	0.31	11.47	11.46	11.46	0.203	0.220	0.233	0.122	0.103	0.075
0.7 (t_1=6)	−0.23	0.70	0.39	11.35	11.30	11.29	0.211	0.227	0.240	0.134	0.115	0.098
0.7 (t_1=12)	0.16	0.53	0.45	11.60	11.57	11.59	0.196	0.213	0.227	0.094	0.074	0.055
0.8 (t_1=6)*	−0.35	0.70	0.37	11.36	11.30	11.33	0.211	0.226	0.239	0.100	0.086	0.071
0.8 (t_1=12)*	0.04	0.44	0.50	11.73	11.69	11.76	0.185	0.202	0.217	0.050	0.038	0.027

Table 4: (continued)

Prior	MPE			sd-PE			CIW			FDR		
Default (t_1=6)	0.06	0.80	0.43	11.06	11.03	11.02	0.239	0.251	0.262	0.116	0.104	0.088
Default (t_1=12)	0.06	0.68	0.28	11.41	11.39	11.41	0.211	0.227	0.241	0.063	0.048	0.034
Calibrated (t_1=6)*	−0.35	0.70	0.37	11.36	11.30	11.33	0.211	0.226	0.239	0.100	0.086	0.071
Calibrated (t_1=12)*	0.04	0.44	0.50	11.73	11.69	11.76	0.185	0.202	0.217	0.050	0.038	0.027

The table presents the mean prediction error (MPE) of the point estimates, the standard error of the MPE (sd-PE), the mean credible interval width (CIW) and the false discovery rate (FDR), in all baseline settings and sensitivity analyses. Baseline settings are indicated by a (*) symbol. For each performance measure, the three columns correspond to the three post-intervention time points. The values of the MPE and sd=MPE are multiplied by 10^3 and 10^2, respectively. Results are based on 10,000 simulated datasets.

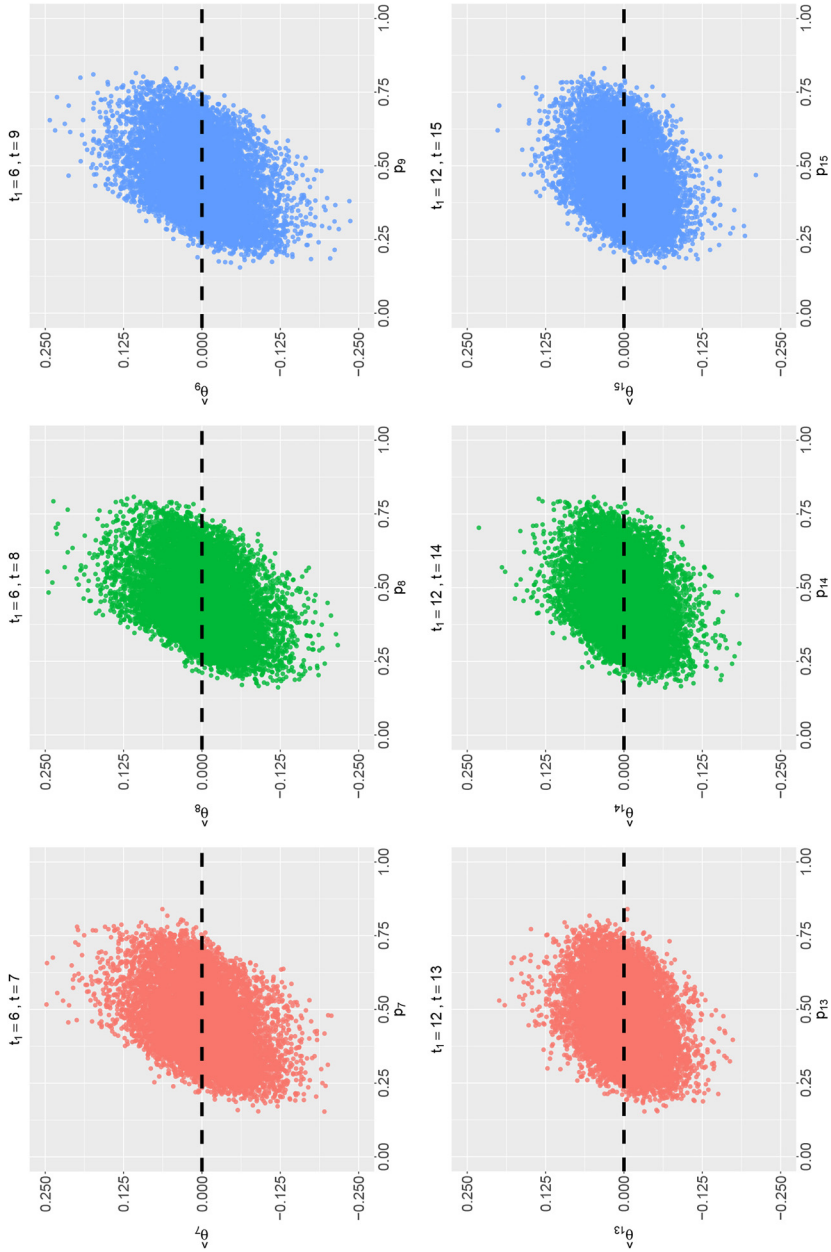

Figure 4: Baseline simulation results. The figure shows the simulated values of the prevalence $p_{0t}^{(0)}$ in the post-intervention period against the estimates $\hat{\theta}_t$ provided by the CIM. The top and bottom rows correspond to $t_1 = 6$ and 12, respectively. The left, middle and right columns correspond to the first, second and third post-intervention time point respectively. Each plot contains 10,000 points, one for each simulated dataset.

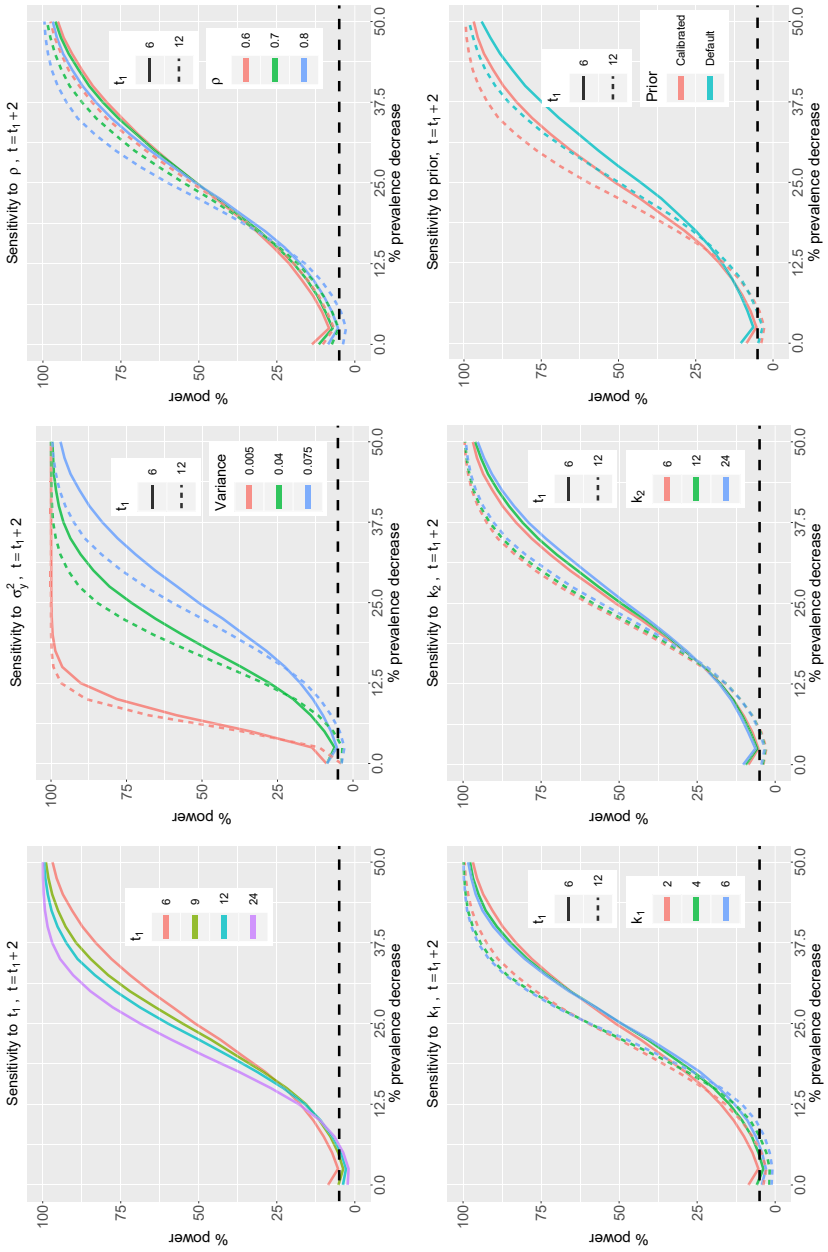

Figure 5: Results of the simulation study for the second post-intervention time point $t = t_1 + 2$. The figure presents the power for detecting an intervention effect achieved by the CIM as a function of the intervention effect magnitude, in all baseline settings and sensitivity analyses. All results are based on 10,000 simulated datasets.

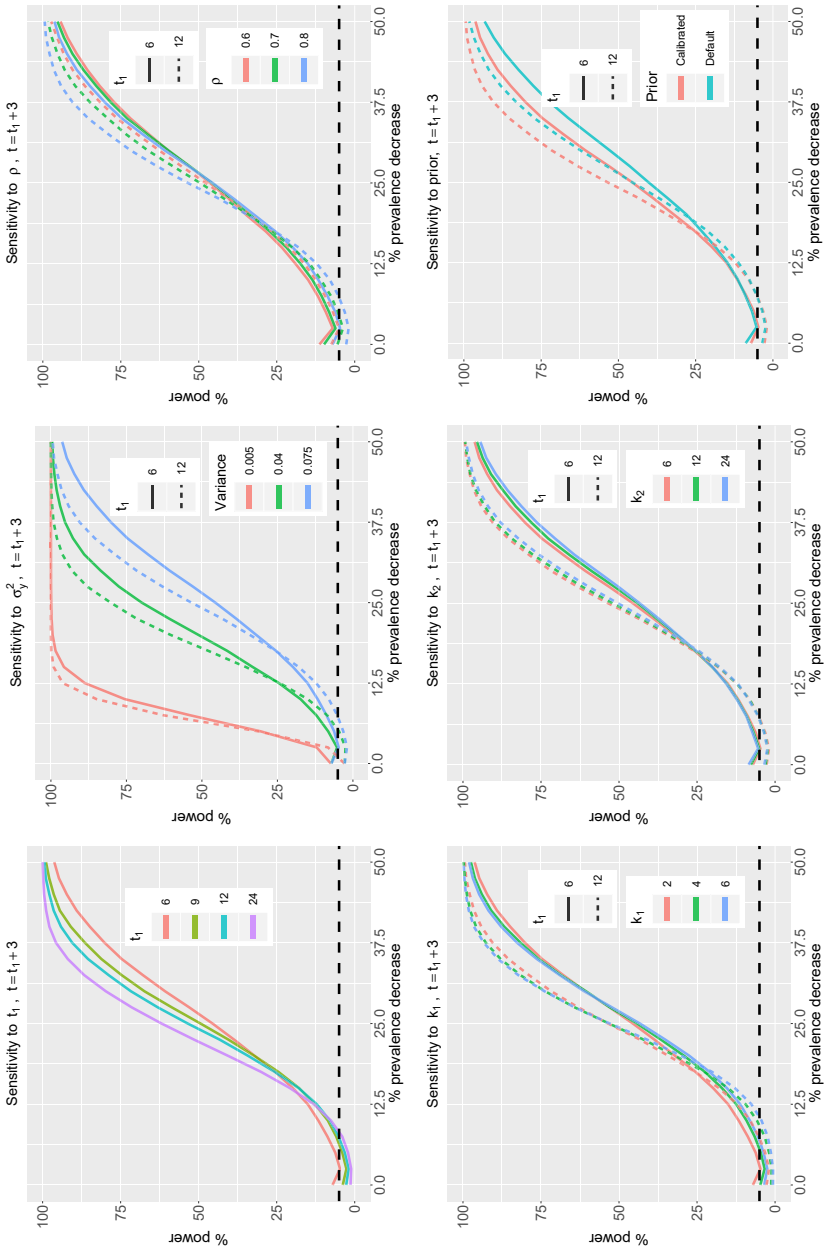

Figure 6: Results of the simulation study for the second post-intervention time point $t = t_1 + 3$. The figure presents the power for detecting an intervention effect achieved by the CIM as a function of the intervention effect magnitude, in all baseline settings and sensitivity analyses. All results are based on 10,000 simulated datasets.

References

Abadie, A., and J. Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *The American Economic Review* 93 (1): 113–32.

Abadie, A., A. Diamond, and J. Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490): 493–505.

Amjad, M., D. Shah, and D. Shen. 2018. "Robust Synthetic Control." *Journal of Machine Learning Research* 19 (1): 802–52.

Bernal, J. L., S. Cummins, and A. Gasparrini. 2016. "Interrupted Time Series Regression for the Evaluation of Public Health Interventions: A Tutorial." *International Journal of Epidemiology* 46 (1): 348–55.

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy, and S. L. Scott. 2015. "Inferring Causal Impact Using Bayesian Structural Time-Series Models." *Annals of Applied Statistics* 9 (1): 247–74.

Bruhn, C. A., S. Hetterich, C. Schuck-Paim, E. Kürüm, R. J. Taylor, R. Lustig, E. D. Shapiro, J. L. Warren, L. Simonsen, and D. M. Weinberger. 2017. "Estimating the Population-Level Impact of Vaccines Using Synthetic Controls." *Proceedings of the National Academy of Sciences* 114 (7): 1524–9.

Chipman, H., E. I. George, and R. E. McCulloch. 2001. *The Practical Implementation of Bayesian Model Selection* In *Volume 38 of Lecture Notes–Monograph Series*, 65–116. Beachwood, OH: Institute of Mathematical Statistics.

Cousien, A., V. Tran, M. Jauffret-Roustide, S. Deuffic-Burban, J.-S. Dhersin, and Y. Yazdanpanah. 2014. "Impact of New DAA-Containing Regimens on HCV Transmission Among Injecting Drug Users (Idus): A Model-Based Analysis (Anrs 12376)." *Journal of Hepatology* 60 (1): S36–7.

De Angelis, D., M. Sweeting, A. Ades, M. Hickman, V. Hope, and M. Ramsay. 2009. "An Evidence Synthesis Approach to Estimating Hepatitis C Prevalence in England and Wales." *Statistical Methods in Medical Research* 18 (4): 361–79.

de Vocht, F. 2016. "Inferring the 1985–2014 Impact of Mobile Phone Use on Selected Brain Cancer Subtypes Using Bayesian Structural Time Series and Synthetic Controls." *Environment International* 97: 100–7.

de Vocht, F., K. Tilling, T. Pliakas, C. Angus, M. Egan, A. Brennan, R. Campbell, and M. Hickman. 2017. "Estimating the Population-Level Impact of Vaccines Using Synthetic Controls." under review.

De Vos, A., and M. Kretzschmar. 2014. "Benefits of Hepatitis C Virus Treatment: A Balance of Preventing Onward Transmission and Re-infection." *Mathematical Biosciences* 258: 11–8.

Dellaportas, P., and D. A. Stephens. 1995. "Bayesian Analysis of Errors-in-Variables Regression Models." *Biometrics* 51 (3): 1085–95.

Dore, G. J., and J. J. Feld. 2015. "Hepatitis C Virus Therapeutic Development: In Pursuit of "Perfectovir"." *Clinical Infectious Diseases* 60 (12): 1829–36.

Durier, N., C. Nguyen, and L. J. White. 2012. "Treatment of Hepatitis C as Prevention: A Modeling Case Study in Vietnam." *PloS One* 7 (4): e34548.

Frost, C., and S. G. Thompson. 2000. "Correcting for Regression Dilution Bias: Comparison of Methods for a Single Predictor Variable." *Journal of the Royal Statistical Society: Series A* 163 (2): 173–89.

George, E. I., and R. E. McCulloch. 1993. "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association* 88 (423): 881–9.

Glass, T. A., S. N. Goodman, M. A. Hernán, and J. M. Samet. 2013. "Causal Inference in Public Health." *Annual Review of Public Health* 34: 61–75.

Gobillon, L., and T. Magnac. 2016. "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls." *The Review of Economics and Statistics* 98 (3): 535–51.

Gogela, N. A., M. V. Lin, J. L. Wisocky, and R. T. Chung. 2015. "Enhancing Our Understanding of Current Therapies for Hepatitis C Virus (HCV)." *Current HIV* 12 (1): 68–78.

Harris, R. J., H. E. Harris, S. Mandal, M. Ramsay, P. Vickerman, M. Hickman, and D. De Angelis. 2019. "Monitoring the Hepatitis C Epidemic in England and Evaluating Intervention Scale-Up Using Routinely Collected Data." *Journal of Viral Hepatitis* 26 (5): 541–51.

Hellard, M., D. A. Rolls, R. Sacks-Davis, G. Robins, P. Pattison, P. Higgs, C. Aitken, and E. McBryde. 2014. "The Impact of Injecting Networks on Hepatitis C Transmission and Treatment in People Who Inject Drugs." *Hepatology* 60 (6): 1861–70.

Hickman, M., D. De Angelis, P. Vickerman, S. Hutchinson, and N. Martin. 2015. "Hcv Treatment as Prevention in People Who Inject Drugs–Testing the Evidence." *Current Opinion in Infectious Diseases* 28 (6): 576.

Hickman, M., J. F. Dillon, L. Elliott, D. De Angelis, P. Vickerman, G. Foster, P. Donnan, A. Eriksen, P. Flowers, D. Goldberg, W. Hollingworth, S. Ijaz, D. Liddell, S. Mandal, N. Martin, L. J. Z. Beer, K. Drysdale, H. Fraser, R. Glass, L. Graham, R. N. Gunson, E. Hamilton, H. Harris, M. Harris, R. Harris, E. Heinsbroek, V. Hope, J. Horwood, S. K. Inglis, H. Innes, A. Lane, J. Meadows, A. McAuley, C. Metcalfe, S. Migchelsen, A. Murray, G. Myring, N. E. Palmateer, A. Presanis, A. Radley, M. Ramsay, P. Samartsidis, R. Simmons, K. Sinka, G. Vojt, Z. Ward, D. Whiteley, A. Yeung, and S. J. Hutchinson. 2019. "Evaluating the Population Impact of Hepatitis C Direct Acting Antiviral Treatment as Prevention for People Who Inject Drugs (Epitope) – A Natural Experiment (Protocol)." *BMJ Open* 9 (9): e029538.

Hsiao, C., S. H. Ching, and S. K. Wan. 2012. "A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong with Mainland China." *Journal of Applied Econometrics* 27 (5): 705–40.

Hutchinson, S., K. Roy, S. Wadd, S. Bird, A. Taylor, E. Anderson, L. Shaw, G. Codere, and D. Goldberg. 2006. "Hepatitis C Virus Infection in Scotland: Epidemiological Review and Public Health Challenges." *Scottish Medical Journal* 51 (2): 8–15.

Kastner, G., and S. Frühwirth-Schnatter. 2014. "Ancillarity-sufficiency Interweaving Strategy (Asis) for Boosting Mcmc Estimation of Stochastic Volatility Models." *Computational Statistics & Data Analysis* 76: 408–23.

Kinn, D. 2018. "Synthetic Control Methods and Big Data." arXiv preprint arXiv:1803.00096.

Martin, N., A. Miners, and P. Vickerman. 2012a. *Assessing the Cost-Effectiveness of Interventions Aimed at Promoting and Offering Hepatitis C Testing in Injecting Drug Users: An Economic Modelling Report*. National Institute for Health and Clinical Excellence (NICE).

Martin, N., P. Vickerman, G. Foster, A. Miners, S. Hutchinson, D. Goldberg, and M. Hickman. 2012b. "The Cost-Effectiveness of Hcv Antiviral Treatment for Injecting Drug User Populations." *Hepatology* 55: 49–57.

Martin, N. K., P. Vickerman, G. R. Foster, S. J. Hutchinson, D. J. Goldberg, and M. Hickman. 2011. "Can Antiviral Therapy for Hepatitis C Reduce the Prevalence of Hcv Among Injecting Drug User Populations? A Modeling Analysis of its Prevention Utility." *Journal of Hepatology* 54 (6): 1137–44.

Martin, N. K., M. Hickman, S. J. Hutchinson, D. J. Goldberg, and P. Vickerman. 2013a. "Combination Interventions to Prevent Hcv Transmission Among People Who Inject Drugs: Modeling the Impact of Antiviral Treatment, Needle and Syringe Programs, and Opiate Substitution Therapy." *Clinical Infectious Diseases* 57 (suppl_2): S39–45.

Martin, N. K., P. Vickerman, J. Grebely, M. Hellard, S. J. Hutchinson, V. D. Lima, G. R. Foster, J. F. Dillon, D. J. Goldberg, G. J. Dore, and M. Hickman. 2013b. "Hepatitis C Virus Treatment for Prevention Among People Who Inject Drugs: Modeling Treatment Scale-Up in the Age of Direct-Acting Antivirals." *Hepatology* 58 (5): 1598–609.

Martin, N. K., P. Vickerman, G. J. Dore, and M. Hickman. 2015. "The Hepatitis C Virus Epidemics in Key Populations (Including People Who Inject Drugs, Prisoners and Msm): The Use of Direct-Acting Antivirals as Treatment for Prevention." *Current Opinion in HIV and AIDS* 10 (5): 374–80.

Martin, N. K., A. Thornton, M. Hickman, C. Sabin, M. Nelson, G. S. Cooke, T. C. Martin, V. Delpech, M. Ruf, H. Price, Y. Azad, E. C. Thomson, and P. Vickerman. 2016a. "Can Hepatitis C Virus (HCV) Direct-Acting Antiviral Treatment as Prevention Reverse the Hcv Epidemic Among Men Who Have Sex with Men in the United Kingdom? Epidemiological and Modeling Insights." *Clinical Infectious Diseases* 62 (9): 1072–80.

Martin, N. K., P. Vickerman, I. F. Brew, J. Williamson, A. Miners, W. L. Irving, S. Saksena, S. J. Hutchinson, S. Mandal, E. O'moore, and M. Hickman. 2016b. "Is Increased Hepatitis C Virus Case-Finding Combined with Current or 8-week to 12-week Direct-Acting Antiviral Therapy Cost-Effective in UK Prisons? A Prevention Benefit Analysis." *Hepatology* 63 (6): 1796–808.

Martin, N. K., P. Vickerman, G. J. Dore, J. Grebely, A. Miners, J. Cairns, G. R. Foster, S. J. Hutchinson, D. J. Goldberg, T. C. Martin, M. Ramsay, STOP-HCV Consortium, and M. Hickman. 2016c. "Prioritization of Hcv Treatment in the Direct-Acting Antiviral Era: An Economic Evaluation." *Journal of Hepatology* 65 (1): 17–25.

O'Neill, S., N. Kreif, R. Grieve, M. Sutton, and J. S. Sekhon. 2016. "Estimating Causal Effects: Considering Three Alternatives to Difference-in-Differences Estimation." *Health Services & Outcomes Research Methodology* 16 (1–2): 1–21.

O'Neill, S., N. Kreif, M. Sutton, and R. Grieve. 2020. "A Comparison of Methods for Health Policy Evaluation with Controlled Pre-post Designs." *Health Services Research* 55 (2): 328–38.

Polson, N. G., J. G. Scott, and J. Windle. 2013. "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables." *Journal of the American Statistical Association* 108 (504): 1339–49.

Prevost, T. C., A. M. Presanis, A. Taylor, D. J. Goldberg, S. J. Hutchinson, and D. De Angelis. 2015. "Estimating the Number of People with Hepatitis C Virus Who Have Ever Injected Drugs and Have yet to Be Diagnosed: An Evidence Synthesis Approach for Scotland." *Addiction* 110 (8): 1287–300.

Rolls, D. A., R. Sacks-Davis, R. Jenkinson, E. McBryde, P. Pattison, G. Robins, and M. Hellard. 2013. "Hepatitis C Transmission and Treatment in Contact Networks of People Who Inject Drugs." *PloS One* 8 (11): e78286.

Rothman, K. J., and S. Greenland. 2005. "Causation and Causal Inference in Epidemiology." *American Journal of Public Health* 95 (S1): S144–S150.

Samartsidis, P., S. R. Seaman, A. M. Presanis, M. Hickman, and D. De Angelis. 2019. "Assessing the Causal Effect of Binary Interventions from Observational Panel Data with Few Treated Units." *Statistical Science* 34 (3): 486–503.

Sutton, M. 2020. "Bayesian Variable Selection." In *Case Studies in Applied Bayesian Data Science*, 121–35. Cham: Springer.

Vickerman, P., N. Martin, and M. Hickman. 2011. "Can Hepatitis C Virus Treatment Be Used as a Prevention Strategy? Additional Model Projections for Australia and Elsewhere." *Drug and Alcohol Dependence* 113 (2): 83–5.

Walker, D. R., M. C. Pedrosa, S. R. Manthena, N. Patel, and S. E. Marx. 2015. "Early View of the Effectiveness of New Direct-Acting Antiviral (DAA) Regimens in Patients with Hepatitis C Virus (HCV)." *Advances in Therapy* 32 (11): 1117–27.

Williams, R., R. Aspinall, M. Bellis, G. Camps-Walsh, M. Cramp, A. Dhawan, J. Ferguson, D. Forton, G. Foster, I. Gilmore, M. Hickman, M. Hudson, D. Kelly, A. Langford, M. Lombard, L. Longworth, N. Martin, K. Moriarty, P. Newsome, J. O'Grady, R. Pryke, H. Rutter, S. Ryder, N. Sheron, and T. Smith. 2014. "Addressing Liver Disease in the UK: A Blueprint for Attaining

Excellence in Health Care and Reducing Premature Mortality from Lifestyle Issues of Excess Consumption of Alcohol, Obesity, and Viral Hepatitis.'' *The Lancet* 384 (9958): 1953–97.

Zeiler, I., T. Langlands, J. M. Murray, and A. Ritter. 2010. ''Optimal Targeting of Hepatitis C Virus Treatment Among Injecting Drug Users to Those Not Enrolled in Methadone Maintenance Programs.'' *Drug and Alcohol Dependence* 110 (3): 228–33.