

13 Reasons Why Probably Increased Emergency Room Visits for Self-Harm among Teenage Girls

Chris Felton

Harvard University

Abstract: I present evidence that the release of Netflix's *13 Reasons Why*—a fictional series about the aftermath of a teenage girl's suicide—caused a temporary spike in emergency room (ER) visits for self-harm among teenage girls in the United States. I conduct an interrupted time series analysis using monthly counts of ER visits obtained from a large, nationally representative survey. I estimate that the show caused an increase of 1,297 self-harm visits (95 percent CI: 634 to 1,965) the month it was released, a 14 percent (6.5 percent, 23 percent) spike relative to the predicted counterfactual. The effect persisted for two months, and ER visits for intentional cutting—the method of suicide portrayed in the series—were unusually high following the show's release. The findings indicate that fictional portrayals of suicide can influence real-life self-harm behavior, providing support for contagion-based explanations of suicide. Methodologically, the study showcases how to make credible causal claims when effect estimates are likely biased.

Keywords: suicide; self-harm; contagion; imitation; interrupted time series; causal inference

ON March 31, 2017, Netflix released the first season of its series *13 Reasons Why*, a fictional drama about the aftermath of a teenage girl's suicide. The show's graphic depiction of this suicide incited controversy, with some mental health professionals warning about the potential for "copycat" suicides (Libbey 2018). Netflix ultimately removed this scene, but not until years later (Brito 2019).

I present evidence that *13 Reasons Why*'s release caused an increase in ER visits for intentional self-harm among teenage girls in the United States. The findings speak to long-standing sociological debates over *contagion*—the idea that observing a behavior in others makes it more likely that a person will exhibit similar behavior (Abrutyn and Mueller 2014b). In particular, the results support the idea that contagion is partly facilitated by perceived similarity. That is, when I view someone else as similar to me, I am more inclined to adopt their behaviors as my own.

Suicide contagion—also called suicide *suggestion* or *imitation*—has a complicated history in sociology. In his foundational study of suicide, Durkheim (1897) dismissed Tarde's (1903) work on imitation as "purely psychological," and Tarde has been largely forgotten by American sociology. Phillips (1974) revived sociological interest in imitative explanations of suicide, and a growing body of empirical work supports the contagion thesis (Domaradzki 2021). These findings have inspired new theoretical work as well. Abrutyn and Mueller (2014b), for instance, revisit Tarde's work and show that it was more *social-psychological* than Durkheim suggested. They argue—compellingly—that we ought to incorporate Tarde's insights into a revised Durkheimian theory of suicide.

The empirical literature on suicide contagion, although impressive, leaves some questions unanswered. Studies largely focus on the effects of celebrity suicides—

Citation: Felton, Chris. 2023. "13 Reasons Why Probably Increased Emergency Room Visits for Self-Harm among Teenage Girls." *Sociological Science* 10: 930-963.


Received: March 15, 2023

Accepted: March 31, 2023

Published: December 11, 2023

Editor(s): Arnout van de Rijt, Peter Bearman

DOI: 10.15195/v10.a33

Copyright: © 2023 The Author(s). This open-access article has been published under a Creative Commons Attribution License, which allows unrestricted use, distribution and reproduction, in any form, as long as the original author and source have been credited. 

such as Robin Williams's or Marilyn Monroe's—which limits our understanding of contagion in two ways.¹ First, the *behavioral models*—that is, the celebrities who die by suicide—are real people. This raises the question of whether fictional suicides might produce similar imitative effects. Second, celebrities are high-status, well-known, and often widely beloved. This raises the question of whether a lower-status person's suicide might produce similar imitative effects.

Although some studies have examined the effect of fictional suicides, the findings are mixed (Domaradzki 2021), and analyses suffer from different methodological shortcomings. Gould and Shaffer (1986), for instance, rely on a before-and-after design to estimate effects of television movies on suicidal behavior, which fails to account for seasonal trends in suicide. Stack et al. (2014) rely on survey data to assess the effect of cumulative exposure to movies featuring suicide on suicide attempts, but unmeasured confounding likely biases treatment effect estimates. Some studies present evidence of contagion with non-celebrity behavioral models, but these too face methodological issues. Using social network data, Abrutyn and Mueller (2014a) argue that suicidal behavior spreads between friends, but the analysis is likely affected by unmeasured confounding as well as network-specific problems like homophily bias and partial censoring of the peer group (Shalizi and Thomas 2011; Griffith 2022). The present study seeks to provide more robust evidence that imitation persists when behavioral models are fictional or low-status.

Prior Work on 13 Reasons Why

Bridge et al. (2020) present evidence that *13 Reasons Why's* release caused a spike in suicide mortality among teenage boys—but not girls—in April 2017. Niederkrotenthaler et al. (2019) find evidence of contagion for both boys and girls, but effect estimates for both groups are noisy, making it difficult to reliably compare the magnitudes. The results are puzzling: the contagion literature suggests the effect should be larger for girls because the show revolves around a girl's suicide. Celebrity suicides, for instance, tend to produce larger spikes in suicide mortality for people of the same gender as the celebrity (Fink et al. 2018; Domaradzki 2021). The absence of effects for girls led some to cast doubt on Bridge et al.'s (2020) findings (e.g., Romer 2020a; Grady 2019).

The two studies on *13 Reasons Why* raise methodological concerns as well. Both studies employ the interrupted time series (ITS) method. This procedure consists of two steps. First, we build a forecasting model for suicide mortality using data from before the show's release but not after—that is, the pre-treatment period. Second, we use this model to predict what suicide mortality would have been in the post-treatment period absent the show's release. Consequently, the accuracy of our treatment effect estimates hinges on the accuracy of our forecaster. Romer (2020a) argues that Bridge et al.'s (2020) forecasting model failed to account for strong temporal trends in suicide. Suicide mortality for teenage boys was rising prior to the show's release, he contends, so it is plausible that mortality would have continued to rise in April 2017 absent the show's release. Media coverage of Bridge et al.'s (2020) study urged caution in drawing causal conclusions from the results (Grady 2019; D'Zurilla 2019).

More broadly, methodologists have put forward several critiques of ITS methods and practice. Zoorob (2020) highlights the model-dependent nature of ITS, and Berk (2021) argues that standard ITS model-selection strategies invalidate confidence intervals. Baicker and Svoronos (2019) find that ITS produces estimates that are large in magnitude, statistically significant, and of the wrong sign when applied to data from the Oregon Health Insurance Experiment. Romer (2020b) also raises concerns about the accuracy of forecasting models used in ITS studies.

Contributions

Three issues motivate this article. First, we lack strong evidence about suicide contagion caused by fictional and low-status suicides. *13 Reasons Why's* release offers an opportunity to study these issues because the character who dies by suicide is not popular or high-status in her school—she is bullied and assaulted by her peers. Furthermore, the actress who plays this character—Katherine Langford—was fairly unknown prior to the show's release. This study provides evidence that the suicides of low-status fictional characters can affect self-harm behavior. It also indicates that previous studies on suicide contagion—which typically restrict attention to suicide mortality—may understate the full extent of imitation effects.

Second, prior research on *13 Reasons Why* introduces a puzzle: why would the show increase suicide mortality for boys but not girls? This study resolves the puzzle by changing the outcome of interest from suicide mortality to self-harm behavior, which is typically non-fatal among teen girls. More specifically, I find that the show's release caused a temporary spike in emergency room (ER) visits for self-harm among teen girls but not teen boys. The result—coupled with null effects for suicide mortality among girls—fits with a well-known gender “paradox” in suicide: although men *die* by suicide more often than women, women *attempt* suicide more often than men (Canetto and Sakinofsky 1998).²

Finally, the methodological critiques of ITS in general—and of *13 Reasons Why* studies in particular—motivate this project. I show how to use ITS to make credible causal claims in the absence of perfect causal identification. ITS, like all observational causal inference methods, requires strong identification assumptions. These assumptions almost certainly fail in most practical settings. Accordingly, I make no claim that my ITS analysis produces exactly unbiased treatment effect estimates or perfectly valid confidence intervals. Instead, I use various supplementary analyses to show that the bias is likely small and that the large, post-release spike in self-harm visits primarily reflects a causal relationship. I also import several tools from the statistics and forecasting literatures that improve upon standard ITS practice.

The article proceeds as follows. In “*13 Reasons Why* Can Enhance Our Understanding of Suicide Contagion”, I motivate the core causal question, and “What Exactly is the Treatment? Why Focus on This Subgroup?” clarifies the nature of the treatment. “Data” describes the data, and “Methods” details the methodological approach. I present the main results in “Results” along with supplementary findings that both further our understanding of the *13 Reasons Why* effect and provide additional evidence that the post-release spike in self-harm visits reflects a causal relationship. “Addressing Methodological Concerns” reports numerous

additional analyses to address potential methodological problems underlying the results. “Conclusion” concludes.

A note on terminology: Throughout the article, I use “self-harm visits” as shorthand for “ER visits for intentional self-harm.” When no demographic group is mentioned explicitly, I am referring to self-harm visits for teen girls.

Studying *13 Reasons Why* Can Enhance Our Understanding of Suicide Contagion

I aim to answer a straightforward question: did the release of Netflix’s *13 Reasons Why* cause a temporary spike in ER visits for self-harm among teenage girls? The answer is important for at least three reasons. First, the answer can inform how studios portray suicide in TV and film. Second, the answer can shed light on how fictional stories influence real-life behavior more broadly. Third, the answer can clarify the role that social status and admiration play in suicide contagion. I expand on each reason in turn.

Understanding whether *13 Reasons Why* led to a spike in self-harm behavior can inform how writers and studios approach fictional portrayals of suicide, particularly for young viewers. The initial release of *13 Reasons Why* was controversial precisely because of concerns about “copycat” suicides. The show featured a graphic depiction of suicide that was ultimately removed years later, and Netflix later added warning cards to the start of each season informing viewers about the show’s themes (Brito 2019). Concerns about suicide contagion, then, have already influenced the way Netflix presents *13 Reasons Why* and have likely affected how writers and studios approach suicide. But providing more rigorous evidence of the effect in question can draw more attention to the need for such changes. Furthermore, demonstrating that the show’s release affected non-fatal self-harm behavior would suggest that the show caused more harm than previously thought.

The effect of *13 Reasons Why* on self-harm can also inform the broader study of how fictional stories influence real-life behavior and attitudes. The effect of fictional stories on drug use and gun violence are hotly debated in the press, and researchers have also examined how fictional stories shape political and social attitudes (Paluck 2009). Evidence of self-harm contagion might shift our expectations about the influence of fiction in these other domains, contributing to a broad evidence base for the effects of fiction on real-life behavior.

Finally, *13 Reasons Why* offers a way to study how social status and widespread admiration affect suicide contagion. A large literature demonstrates that widely-reported celebrity suicides cause short-term increases in suicide mortality (Fink et al. 2018; Phillips 1974; Domaradzki 2021). One explanation for the effect is that the deaths make the idea of suicide more acceptable or cognitively accessible to at-risk populations, such as people suffering from depression (Abrutyn and Mueller 2014b; Patterson 2014). Relatedly, Tankard and Paluck (2016) argues that stories influence norm perception—that is, what people think the current social norms actually are.

How celebrity suicides affect perceived acceptability remains an open question. A death might influence perceptions through the celebrity's social status or popularity. Virtually all celebrities are high-status, and many are widely adored. Alternatively, the effect could stem from personal identification with a celebrity.³ If *J* views *K* as similar to her, and *K* dies by suicide, this might make the idea of suicide more acceptable or cognitively accessible to *J*.

Netflix's *13 Reasons Why* provides an opportunity to study suicide contagion while separating the effects of social status from the effects of personal identification with the deceased. The character who dies by suicide is bullied and assaulted by her peers, and the actress playing her was not well-known prior to the show's release. Studying *13 Reasons Why*, then, can shed light on how it is that widely-reported events or widely-consumed stories can influence behavior.

What Exactly is the Treatment? Why Focus on This Subgroup?

Two issues require clarification: the nature of the treatment and the decision to focus on effects among teen girls. In short, the treatment is the show's *release*—not the actual watching of the show—and I focus on teen girls both to resolve the substantive puzzle raised by earlier studies and because I expect effects to be largest among this subgroup. I expand on each point below.

This study examines the effect of the show's release but not the actual watching of the show. The distinction is important: the show's release plausibly affected self-harm behavior among those who had never even seen it. [Figure 1](#) illustrates how. Suppose that *J* watches the show, causing her to self-harm. *J*'s self-harm behavior might cause her friend, *K*, to self-harm as well. Thus, even if *K* had not watched the series, the show's release could have indirectly caused her to self-harm. Furthermore, the graph does not exhaust all possible causal pathways through which the show's release might cause *J* or *K* to self-harm. For instance, seeing social media posts about the show might cause *J* to talk to *K* about self-harm, which could in turn cause *K* to self-harm.⁴

Clarifying the treatment raises another question: why focus specifically on effects for teen girls? As I have defined the treatment, everybody was exposed to it. Anyone could have watched *13 Reasons Why*, and contagion effects could be present for any age group or gender. I focus on teen girls for two reasons, one substantive and one practical. Substantively, I hope to resolve the puzzle raised by Bridge et al. (2020): why did we observe no mortality effects for teen girls, who should theoretically be the most susceptible to contagion effects? Practically, we are most likely to detect treatment effects among this group. Teen girls were among the most likely demographic groups to watch the show, and the literature on suicide contagion suggests that teen girls should be the most affected by the depicted death. Because ITS makes it difficult to detect small effects, I focus the analysis on teen girls.

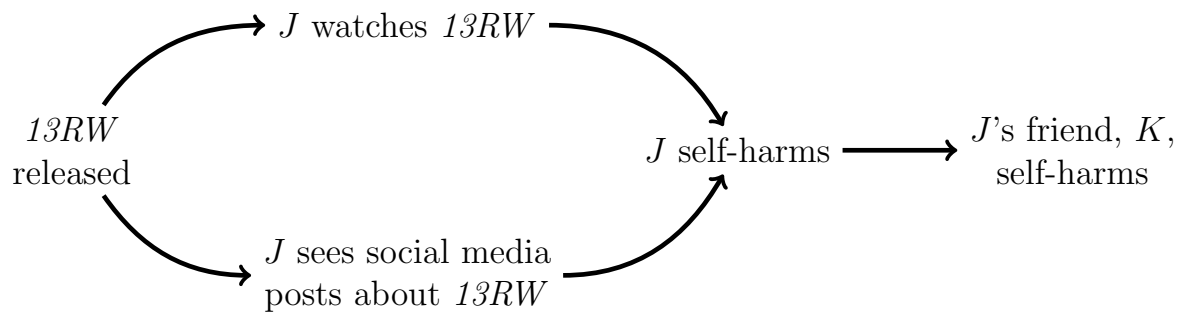


Figure 1: Releasing 13 Reasons Why can (indirectly) cause K to self-harm even if she never watches it. *Notes:* This graph does not exhaust all possible causal pathways through which the show's release might cause J or K to self-harm. The point is that there are causal pathways other than watching the show through which the treatment plausibly affects the outcome.

Data

I use data from the National Emergency Department Sample (NEDS) collected by the Healthcare Cost and Utilization Project (HCUP) to estimate national monthly counts of emergency-room (ER) visits for intentional self-harm by girls age 10–19.⁵ These counts include fatal and non-fatal visits, but the proportion of fatal visits typically falls below 1 percent.⁶ The NEDS is the largest survey of ER visits in the US. It contains data on approximately 30 million ED discharges per year. The data is collected through a stratified, single-stage cluster sample: HCUP constructs strata based on hospital characteristics, and, every year, takes a stratified random sample of hospitals, obtaining 100 percent of discharge records for each hospital. The yearly change in the sample of hospitals raises methodological concerns that I address in Results and Methodological concerns.

Intentional self-harm was categorized using the 9th and 10th editions of the International Classification of Diseases and Related Health Problems (ICD-9 and ICD-10) by the World Health Organization. Specifically, I use the ICD-9 and ICD-10 diagnosis and injury codes listed under “self-inflicted harm” by the Healthcare Cost and Utilization Project (2021). I included ER visits that had intentional self-harm listed for any injury or diagnosis (i.e., whether primary or secondary). Section A in the online supplement details the coding schemes for both the main analysis and supplementary analyses, and Section B in the supplement shows that the results are robust to alternative coding schemes. National estimates were obtained using discharge-level weights in the NEDS. HCUP calculates discharge weights by dividing the total number of ER visits within a stratum by the total number of sampled ER visits within that stratum, obtaining the total number of ER visits from the American Hospital Association. Results using unweighted counts are similar (see Section B in the online supplement).

I drop observations that are missing gender, age, or month of visit to obtain these counts. This typically results in a loss of approximately 15 percent of observations per year, with exact figures presented in Section C of the online supplement. For

supplementary analyses, I also estimated national monthly counts for intentional cutting, accidental cutting, intentional poisoning, and accidental poisoning by girls age 10–19, as well as counts for intentional self-harm by boys age 10–19, men age 40–65, and women age 40–65.

These counts are imperfect. A non-trivial proportion of observations are removed due to missing information, and as Owens et al. (2020) note, the switch from the ICD-9 coding scheme to the ICD-10 coding schemes in late 2015 likely required a transition period for hospitals to become more familiar with the latter. But as I show later, these measurement issues cannot plausibly explain the post-*13 Reasons Why* spike in self-harm visits among teen girls. The elevated counts in the two months following the show's release are uniquely large outliers in the time series, and we observe no such elevation for plausible control groups. Furthermore, simulations in Section C of the online supplement indicate that measurement error produces very little—if any—bias for ITS effect estimates.

Methods

This section outlines the methodological approach I take. “When to Use ITS and How It Works” motivates the ITS design in this particular study and reviews how ITS is implemented. “Improving the ITS Design” highlights three problems with conventional ITS practice and describes the steps I take to avoid these issues.

When to Use ITS and How It Works

The ideal setting for ITS has three distinguishing features. First, we have a time series of an aggregate outcome for some group—such as total self-harm visits for teen girls. Second, this group becomes treated all at once at some point in time. In this case, the group becomes treated on March 31, 2017, when *13 Reasons Why* is released. Recall that “treatment” refers to the release of the show rather than the actual watching of the show (see “What Exactly is the Treatment? Why Focus on This Subgroup?”). Finally, ITS is most appealing when we lack data on a plausible control group—or when plausible control groups exhibit different trends and seasonality than the treated group.⁷

ITS is especially useful in such settings because it requires no data on a control group. The core insight behind the identification strategy is that we can use the treated group's pre-treatment outcomes to predict its counterfactual outcomes in the post-treatment period, as shown in Figure 2. To begin an ITS analysis, we select a forecasting model and fit it to the *pre-treatment* series—in this case, the series of self-harm visits prior to *13 Reason Why's* release. We then use this model to forecast outcomes for the *post-treatment* period—the period immediately following the show's release. Finally, we compare these *forecasted* post-treatment outcomes with the *observed* post-treatment outcomes. The intuition is that the model can predict what would have happened in the absence of treatment because it was built using observations that had not yet been exposed to the treatment.

We can describe this procedure more formally with potential outcomes notation. Our goal is to estimate the causal effect of releasing *13 Reasons Why* on the number

Used to train forecasting model

Month	t	13RW Released	$Y_t(1)$	$Y_t(0)$	$\tau_t = Y_t(1) - Y_t(0)$
Jan 2006	1	0	—	5,101	—
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
Feb 2017	$T - 2$	0	—	8,099	—
Mar 2017	$T - 1$	0	—	9,450	—
Apr 2017	T	1	10,460	?	10,460 - ?

Imputed using forecasting model

Figure 2: The logic of interrupted time series (ITS). *Notes:* Annotated table illustrating the logic of ITS. A forecasting model is trained on the pre-treatment time series, which represents a series of potential outcomes under control ($Y_t(0)$). This model is then used to forecast, or impute, the potential outcome under control in the post-treatment period. The treatment effect is estimated by subtracting the imputed potential outcome under control ($\hat{Y}_T(0)$) from the observed potential outcome under treatment ($Y_T(1)$).

of self-harm visits among teenage girls. $Y_t(1)$ represents the number of self-harm visits we would observe in period t had *13 Reasons Why* been released that period, and $Y_t(0)$ represents the number of ER visits we would observe in period t absent *13 Reasons Why*'s release. We define the causal effect in period t as $\tau_t = Y_t(1) - Y_t(0)$ —the difference between the potential outcome under treatment in period t and the potential outcome under control in period t .

The fundamental problem of causal inference is that we observe only one of these two potential outcomes in any given period. From January 2006 to March 2017—the pre-treatment period—we observe only $Y_t(0)$. In April 2017, we observe only $Y_t(1)$. The ITS solution is to use the pre-treatment series to build a forecasting model for the potential outcomes under control, $Y_t(0)$. We can then use this model to impute $Y_t(0)$ for the post-treatment period. We then estimate our treatment effect as $\hat{\tau}_t = Y_t(1) - \hat{Y}_t(0)$, where $Y_t(1)$ is observed and $\hat{Y}_t(0)$ is predicted from our forecasting model.

ITS relies on two central assumptions (Morgan and Winship 2015). First, we assume that we can recover the correct model specification for the potential outcomes under control. This is a strong assumption that will virtually always be violated in practice. Second, we assume that this model specification remains unaltered in the post-treatment period. This latter assumption might be violated in the presence of concurrent changes that affect the outcome—that is, concurrent treatments.⁸ For instance, if a beloved celebrity died by suicide the same month as *13 Reasons Why*

Table 1: Methodological improvements to standard ITS practice.

Improvement	Purpose
1. First- and seasonally-difference the time series	→ Accounts for trends, seasonality, and other non-stationarities
2. Select the model class using a distinct but similar time series (CDC suicide mortality data)	→ Guards against overfitting
3. Construct conformal predictive intervals	→ Provides valid inference even when the model is misspecified

Notes: Methodological practices I follow for the main ITS analysis. I describe my approach in more detail in Section E of the online supplement. A supplementary analysis uses sample-splitting to further guard against overfitting, and the findings are similar (Section B in the online supplement).

was released, we would be unable to separate the effects of show from the effects of the celebrity death.

Improving the ITS Design

This subsection describes three problems with conventional ITS practice and offers three potential solutions. In this particular application, the methodological improvements change the results very little, so uninterested readers may skip to “Results” without much loss of understanding. Section E in the online supplement more thoroughly outlines the approach.

Critics have identified (at least) three problems with standard ITS practice. First, Romer (2020a) points out that practitioners may fail to account for trends and seasonality in the data. Failing to account for an upward secular trend in self-harm visits may lead us to underestimate the counterfactual outcome in the post-treatment period and overestimate the treatment effect of *13 Reasons Why*.

Second, Berk (2021) recognizes a more subtle problem with current practice. Researchers typically use the pre-treatment time series to both select *and* estimate a forecasting model. For instance, we might fit many models to the pre-treatment series of self-harm visits and select one according to the Akaike information criterion (AIC) or time-series cross-validation. We then use this same data to estimate the model parameters and forecast the counterfactual outcomes. The problem is that the selected model may overfit the data, generating artificially small residuals and producing prediction intervals that are too narrow. Re-using data in this manner can thus invalidate prediction intervals.⁹

The third and most fundamental problem with ITS is that the selected model will almost always be misspecified, generating bias and invalidating conventional confidence intervals. With potentially infinite reasonable models to choose from, it is extremely unlikely that we select the true specification—which may be too complex to reliably estimate regardless.¹⁰ The first two problems can be seen as a special case of this one, but they are worth discussing separately: even when we account for trends and avoid re-using data, our model will likely be misspecified.

I use several tools to circumvent these problems, summarized in [Table 1](#). To account for trends and seasonality, I perform all analyses on data that has been first- and seasonally-differenced.¹¹ Visual inspection and statistical tests suggest the data are stationary after this double-differencing.

I employ two strategies to avoid overfitting. First, I use national monthly suicide mortality data from the CDC as an exploratory data set to test a wide variety of forecasting models. The logic is that suicide mortality data exhibits similar seasonality and trends to the self-harm data without sharing any idiosyncratic randomness. I assess model fit using time-series cross-validation and find that AutoRegressive Integrated Moving Average (ARIMA) models with no additional covariates (e.g., monthly economic indicators) outperform others. I use the `auto.arima()` function in R, which selects the order of the ARIMA model using an AIC-guided stepwise selection (Hyndman and Khandakar 2008). I only run ARIMA models on the self-harm data, and I only use specifications chosen by this stepwise selection algorithm.

Second, I employ the main ITS analysis with and without sample-splitting. For the split-sample analysis, I use the first half of the differenced pre-treatment series to select a model specification using stepwise selection and then use the second half to estimate the model parameters with the chosen specification. Results of the analysis are extremely similar regardless of whether sample-splitting is used, and simulations suggest that, when conducting stepwise selection with no added covariates, using the entire pre-treatment series improves predictive accuracy without undercovering (Section F in online supplement). All analyses use the entire pre-treatment series to both select and fit the model unless otherwise specified.

Finally, to address model misspecification, I use conformal predictive inference, which can provide exact prediction (and confidence) intervals *even when the model is misspecified* (Shafer and Vovk 2008; Lei et al. 2018; Chernozhukov et al. 2021). In particular, I employ Chernozhukov et al.'s (2021) procedure, which extends the standard conformal approach to the time-series setting.¹² As a result, even if ITS treatment effects are biased, we can still construct valid confidence intervals for the treatment effect.¹³ Conformal inference ensures that, even when our model is wrong, it can still be useful. This shifts the goal of the analysis from unbiased treatment effect estimation to valid inference (i.e., confidence intervals) for the treatment effect. In this setting, the conformal inference prediction intervals and the standard ARIMA intervals are virtually identical. I briefly walk through the mechanics and assumptions behind conformal inference in Section E of the online supplement.

Results

This section presents the core substantive results of the study. “Main Results” shows the results of ITS analyses for both teen girls and other demographic groups. “Understanding the 13 Reasons Why Effect” displays two supplementary analyses that serve to both expand our understanding of the 13 Reasons Why effect and provide additional evidence that the post-release spike in self-harm visits reflects a causal relationship.

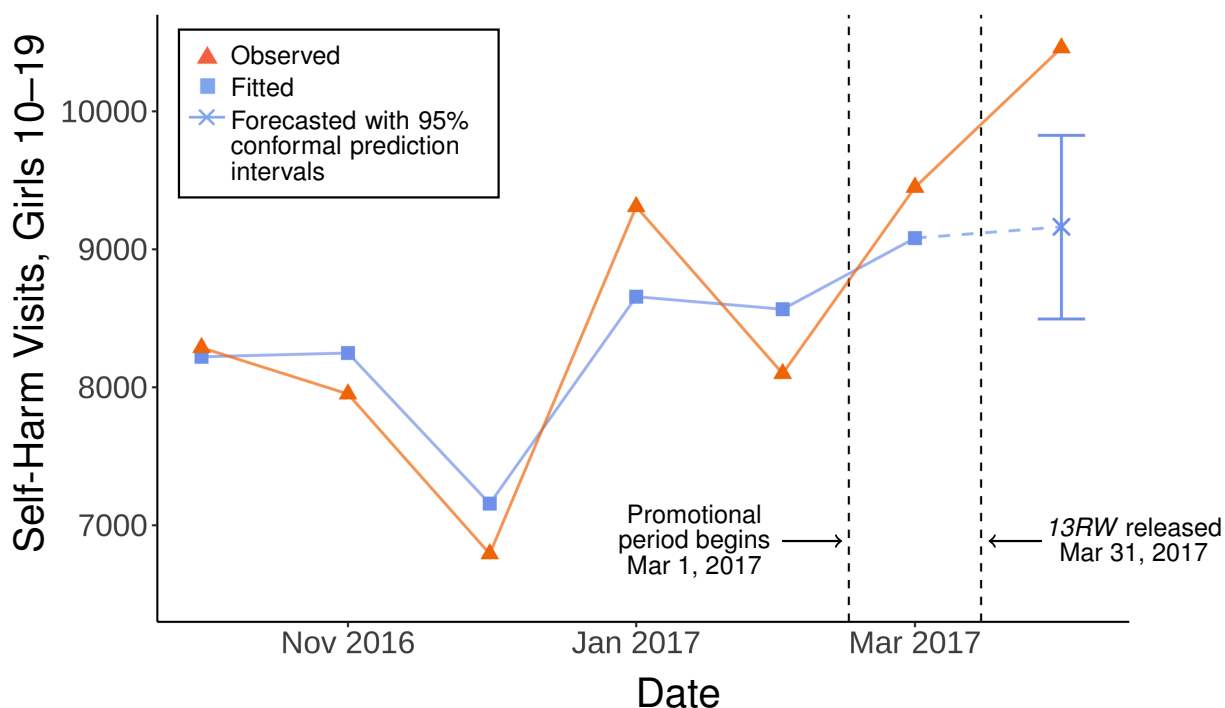


Figure 3: Main ITS results. *Notes:* Plot showing the final 6 pre-treatment periods and first post-treatment period. The model is fit to 122 differenced pre-treatment periods, but all values are de-differenced for visualization. The estimated treatment effect is an increase of 1,297 self-harm visits, with 95 percent conformal confidence intervals ranging from 634 to 1,965. This is a 14 percent (6.5 percent, 23 percent) spike relative to the predicted counterfactual.

Main Results

Figure 3 shows the results of the main ITS analysis using 95 percent conformal prediction intervals. The model is fit to 122 differenced pre-treatment periods. The estimated treatment effect is an increase of 1,297 self-harm visits, with 95 percent conformal confidence intervals ranging from 634 to 1,965. This is a 14 percent (6.5 percent, 23 percent) spike relative to the predicted counterfactual.¹⁴ It is worth reiterating that very few of these visits—typically less than 1 percent—result in death (see “Data”).

How does the magnitude of this effect compare with the effect of celebrity suicides? Fink et al. (2018) found that, in the month of Robin Williams’s death, suicide mortality was 18.3 percent higher than the forecasted counterfactual among people age 30–44. The effects are close in magnitude, but it is worth noting that Robin Williams’s death occurred in the middle of the month rather than at the beginning, so the Fink et al. (2018) estimate understates the one-month treatment effect of his death. Furthermore, the Fink et al. (2018) estimate is for all people age 30–44, not just men. Among men age 30–44, the effect may have been more pronounced.

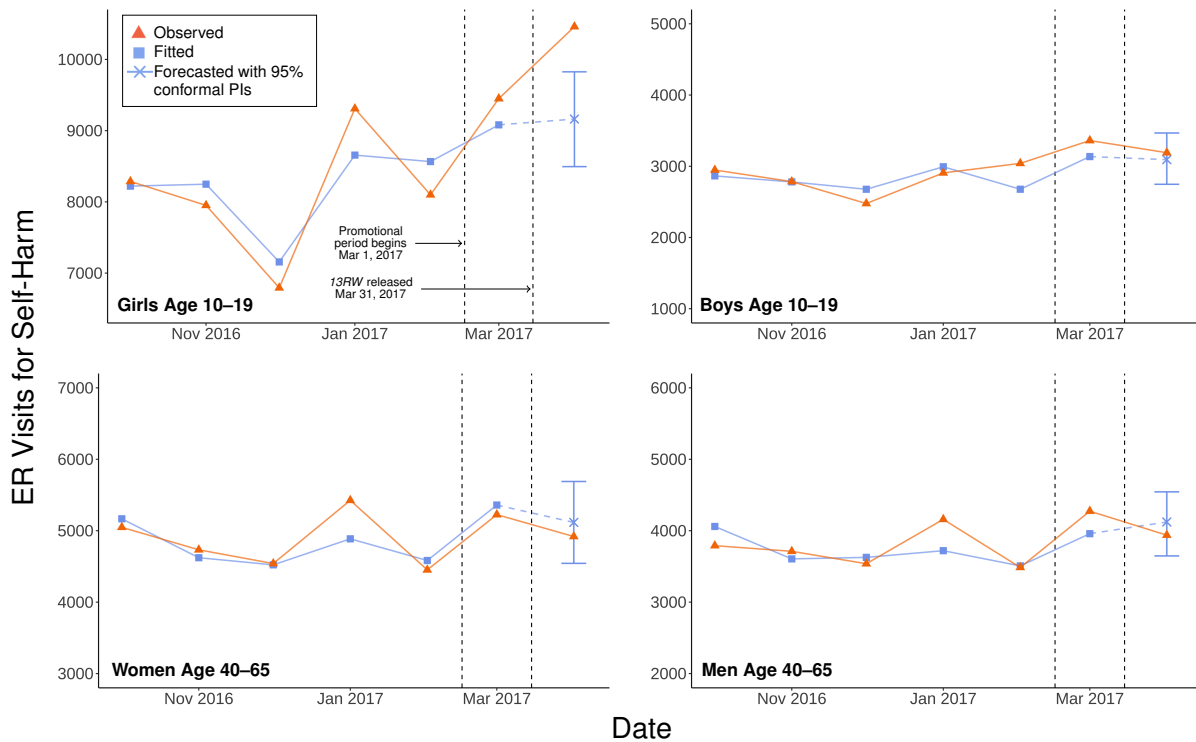


Figure 4: Estimated treatment effects are negligible for teen boys and older adults. *Notes:* For each group, I perform AIC-guided stepwise selection to choose an ARIMA specification using only the pre-treatment series. All models include first- and seasonal-differencing to account for secular trends and seasonality. The result for boys fits with the gender “paradox” in suicide, and the analyses for adults serve as placebo tests for the ITS method. Predicted values for each group except teen girls lie within the prediction intervals, boosting confidence in the out-of-sample predictive accuracy of the model-selection procedure. Errorbars show 95 percent conformal prediction intervals.

Figure 4 shows results for teen girls alongside three other demographic groups: teen boys, men age 40–65, and women age 40–65. ARIMA models are chosen for these groups using the same selection method used for girls.¹⁵ That is, I perform AIC-guided stepwise selection on each time series to choose the ARIMA specification. The analysis for teen boys—coupled with Bridge et al.’s (2020) results on boys’ mortality—shows that the gender “paradox” applies here. Mortality rose for boys but not girls; ER visits for self-harm rose for girls but not boys.

The analyses for men age 40–65 and women age 40–65 serve as placebo tests for the ITS method applied. Because these groups were both less likely to watch the series and less likely to identify with the deceased character, we expect negligible treatment effects in these groups. If the method works, it should estimate treatment effects close to zero—and it does. Although the estimated treatment effects likely suffer from some bias, the placebo test results suggest that the bias might not be too severe. For instance, if poor out-of-sample prediction were causing the discrepancy between the observed and forecasted values for teen girls, we would expect to see

similar discrepancies for the other groups, because the model-selection procedure was identical across all four analyses. It is worth noting, however, that the models in the placebo groups (as well as teen boys) show closer in-sample fit and narrower prediction intervals than the model for teen girls, suggesting the treated group's series is noisier.

Understanding the 13 Reasons Why Effect

In this section, I present results from two additional analyses. The first explores the duration of the treatment effect, and the second examines forms of self-harm used during the post-treatment period. I find that the effect persists for two months, which is consistent with Niederkrötenhaller et al.'s (2019) finding that social media posts about the show had largely dissipated by June 2017. I also find that ER visits for cutting—the method of suicide portrayed in *13 Reasons Why*—are particularly high in the post-treatment period.

These analyses serve two purposes. First, they help us better understand the *13 Reasons Why* effect. The imitation of method, for instance, suggests that the effect may operate through similar mechanisms as celebrity effects. Second, the results make us more confident that the post-release spike in self-harm visits primarily reflects a causal effect. Although the estimated treatment effects are imprecise and likely biased, the additional evidence makes it more difficult to believe that model misspecification or concurrent changes can explain the entire post-release rise in self-harm visits.

The treatment effect persists for two months. To assess the duration of the treatment effect, I use the fitted ARIMA model from “Main Results” to forecast 4 months ahead with 95 percent prediction intervals. The prediction intervals are constructed using simulated forecasting paths with bootstrapped residuals because the conformal inference approach of Chernozhukov et al. (2021) does not easily extend to the multi-step-ahead setting (Hyndman and Khandakar 2008; Miratrix 2022). The post-treatment observations are not used for these forecasts because the goal is to forecast the potential outcomes under control.

Figure 5 plots the results. The observed counts of self-harm visits are much higher than the forecasted values in April and May. However, in June and July the observed counts are close to the forecasted values and well within the prediction intervals. A duration of two months is consistent with findings from Niederkrötenhaller et al. (2019). They find that social media posts related to *13 Reasons Why* were highest in April 2017 and had largely dissipated by June 2017.

Two factors could explain why these durations align. First, social media activity serves as a proxy for public interest in the show. Second, social media coverage of the show might mediate the effect of the show's release on self-harm visits (see Figure 1). That is, viewing social media posts related to the show may trigger self-harm behavior even among people who have never watched it. The analysis provides no direct evidence of mediation, but the results suggest it cannot be ruled out.

The findings also shed doubt on competing, non-causal stories about the post-release spike in self-harm visits. Consider the idea that model misspecification

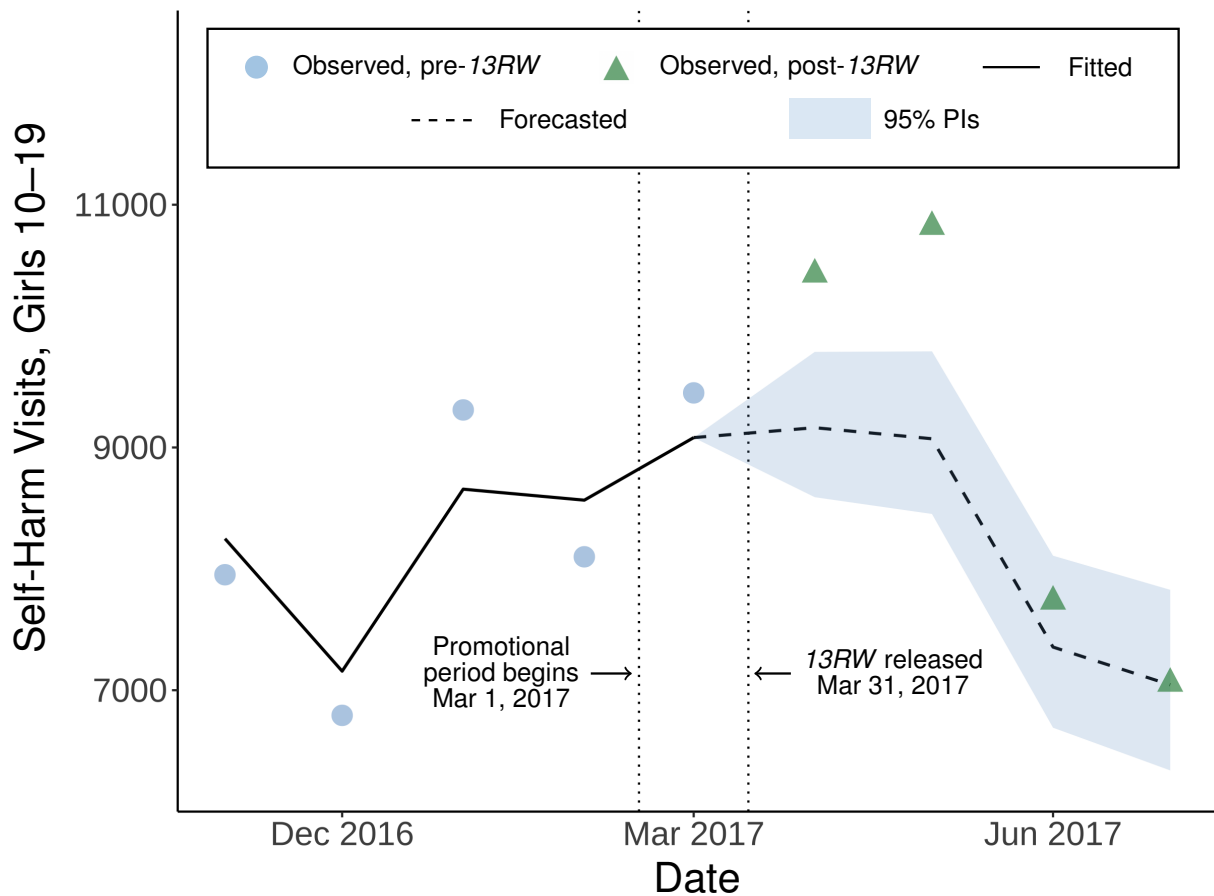


Figure 5: The treatment effect persists for two months. *Notes:* Multi-step-ahead forecasts with 95 percent prediction intervals constructed from simulated paths using bootstrapped residuals. The duration of the effect is consistent with trends in *13 Reasons Why*-related social media posts. The prediction intervals likely undercover the potential outcomes under control, but analyses in “Methodological Concerns” provide more evidence that the effect persisted for two months.

explains the gap between observed and forecasted counts in April and May. Why, then, are the forecasts for June and July so accurate? And why does the duration of the apparent treatment effect happen to align with *13 Reasons Why*-related social-media activity? The results are far from conclusive, but they make competing stories less plausible.¹⁶

ER visits for intentional cutting are particularly high in April and May 2017. The second analysis explores the rise in ER visits for cutting among girls age 10–19. The literature on celebrity suicides shows that people imitate the method of suicide. Rates of suicide by asphyxiation, for instance, were particularly high following Robin Williams’s death. Because *13 Reasons Why* portrayed a suicide death by cutting, we should expect ER visits for cutting to be particularly high in April and May 2017. This is exactly what we find.

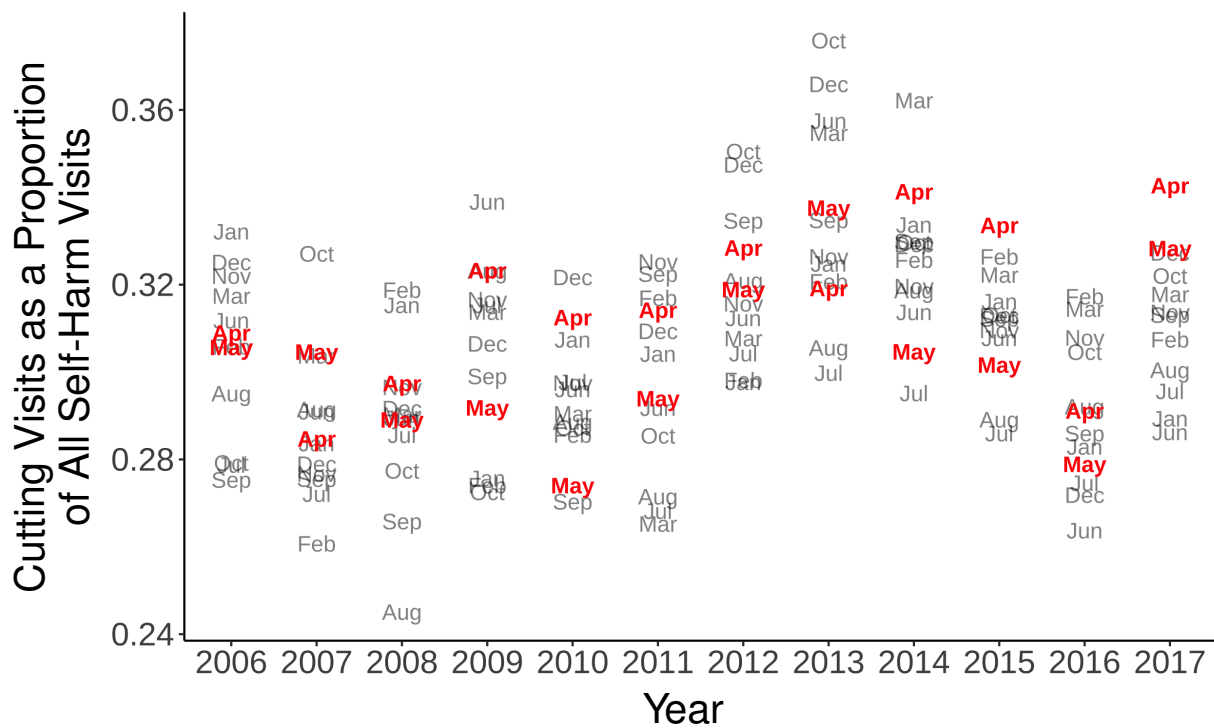


Figure 6: ER visits for intentional cutting are particularly high following 13 Reasons Why’s release. *Notes:* The plot shows the monthly proportions of self-harm visits that are for intentional cutting among teen girls. Each month is labeled with its abbreviation, and April and May are highlighted in red. The proportions show noticeable variation by year, likely due to yearly changes in the sample of hospitals, but in no year other than 2017 do April and May have the two highest counts.

Figure 6 plots the proportion of total self-harm visits that were for intentional cutting by month and year. The plot shows a curious trend: proportions are unusually high from 2012–2015 before dropping in 2016. This is likely a consequence of the NEDS sampling procedure, whereby a different sample of hospitals is selected each year. It is possible that the particular samples in these years resulted in unusually high proportions of cutting. The decline in 2016 may stem from the switch from ICD-9 to ICD-10 in late 2015, but the source of the earlier discontinuity is unclear.¹⁷

Because of these strange jumps, I plot the data by year and label each month, highlighting April and May in red. The plot helps us appreciate how jointly unusual April and May 2017 are. In some years, April has the highest or second-highest proportion of self-harm visits for cutting, but in no other year are both April and May together at the top of the pack. I plot the proportions sequentially in Section B of the online supplement.

Looking at each month separately, April 2017 has the 8th largest proportion of cutting visits out of 144 months, placing it in the 95th percentile, whereas May 2017 is the 21st highest (86th percentile).¹⁸ If we demean the proportions by year, the figures are 6th (97th percentile) and 29th (81st percentile), respectively. Furthermore,

as I show in “ER Visits for Accidental Cutting and Poisoning are Not Especially Low in April or May 2017”, these estimated proportions may understate the spike in cutting visits: ER visits for *accidental* cutting—but not accidental poisonings—rise substantially following the show’s release, suggesting that some intentional cuttings were misclassified as accidental.

The cutting results allow us to say nothing definitive about the mechanisms through which the *13 Reasons Why* effect operates, but they do offer an opportunity to speculate. First, the cutting results point to a commonality with celebrity effects: the imitation of method. The similarity suggests that the *13 Reasons Why* effect may operate through similar social-psychological mechanisms as celebrity effects. In particular, imitation of method seems to make cognitive accessibility a plausible mediator of the effect. The results also suggest that masking the method of suicide in its fictional depictions could curb imitation effects—as Pirkis et al. (2006) have recommended for media coverage of celebrity suicides.

The results also provide more evidence that the post-release spike in self-harm visits reflects a causal relationship. The ITS treatment effect estimates are likely biased, and, as I show later, the conformal prediction intervals probably provide imperfect coverage. But the apparent imitation of method—especially when coupled with the two-month duration of the post-release spike—points toward a causal effect of the show’s release.

Addressing Methodological Concerns

In section, I address potential methodological problems that might undermine the results, such as model dependence, invalid prediction intervals, anticipation effects, and differential measurement error. The discussion reinforces the study’s main thesis: although treatment effect estimates likely suffer from some degree of bias, the causal effect of *13 Reasons Why*’s release probably explains the vast majority of the post-release spike in self-harm visits. Indeed, results in “ER Visits for Accidental Cutting and Poisoning are Not Especially Low in April or May 2017” suggest that some ER visits for intentional cutting may have been misclassified as accidental, leading us to underestimate the effect of the show on self-harm visits.

April and May 2017 Forecasts are Virtually Unaffected by Alternative Post-Treatment-Period Cutoffs

The canonical ITS setup requires a clean separation between pre- and post-treatment periods. The unit is untreated in one period and becomes treated in the next. The *13 Reasons Why* promotional period complicates this separation. On March 1, 2017, a month before the show came out, Netflix began promoting the show’s first season and airing a trailer for the show (Bridge et al. 2020). Because the widespread promotion plausibly affected self-harm behavior, we might think of March 2017 as a “partially treated” period.

If March 2017 represents a partially treated period—rather than a true control period—it poses two problems for the analysis. First, recall that ITS relies on using a forecasting model for the potential outcomes under control $Y_t(0)$ —the count of

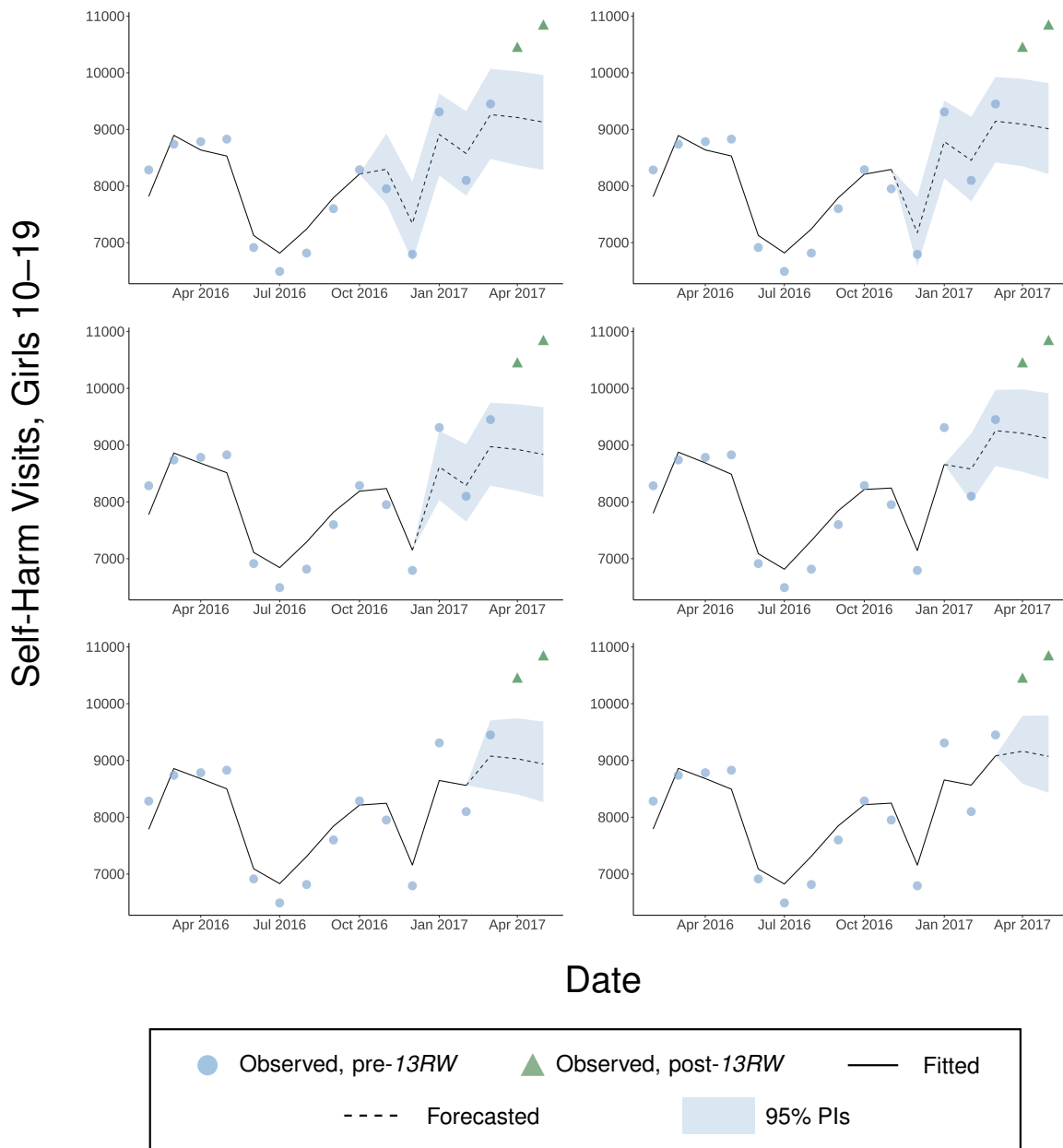


Figure 7: Multi-step-ahead prediction plots. *Notes:* Multi-step-ahead prediction plots using simulated forecasting paths with bootstrapped residuals to construct prediction intervals. Models are fit up to the period indicated by the switch from a solid line to a dashed line. These intervals lack the theoretical guarantees of conformal inference, but they also allow us to conduct multi-step-ahead predictive inference. The forecasts for April and May of 2017 are insensitive to the start of the forecasting period and always lie well outside the prediction intervals, which should mitigate concerns about anticipation effects.

self-harm visits in period t we would observe absent the show's release. Using March 2017 to select this model may result in a worse model choice for the potential outcomes under control. Second, even if the observation is not used to select a model, an autoregressive model will use March's count of self-harm visits to forecast April's count of self-harm visits. The model is built to forecast a future control outcome using a past control outcome, but using March means we are using a partially treated outcome to forecast a control outcome. Section G in the online supplement provides a more formal discussion of these problems.

Fortunately, using March 2017 to select a model—or forecast April 2017's count of self-harm visits—matters very little. Figure 7 shows that the forecasts for April 2017 and May 2017 remain virtually unchanged across six different cutoff points (each month from October 2016 to March 2017). For each cutoff point, I select an ARIMA specification using only pre-cutoff data and then use that model to forecast multiple periods ahead. The plots show prediction intervals based on simulated forecasting paths with bootstrapped residuals. For each cutoff point, the observed values for April and May 2017 lie outside the 95 percent prediction intervals.

ER Visits for Accidental Cutting and Poisoning are Not Especially Low in April or May 2017

The central causal claim I have advanced is that the release of the show *13 Reasons Why* temporarily increased self-harm behavior among teen girls. But there is a competing story we can tell about the data: the show's release could have caused physicians to misclassify accidental cuttings and poisonings as intentional.¹⁹ If a doctor's awareness of *13 Reasons Why* increased her propensity to classify a particular drug overdose as intentional, we might see a rise in ER visits for "intentional" self-harm in the absence of any change in actual self-harm behavior. In short, *13 Reasons Why* may have changed doctors' behavior rather than girls'.

To explore this possibility, I compare ER visits for intentional and accidental injuries. Specifically, I focus on ER visits for poisoning and contact with sharp objects among teen girls, which make up roughly 97 percent of intentional self-harm visits for teen girls in April 2017. The ICD-10 classifies poisonings and contact with sharp objects as either accidental, intentional, or of undetermined intent. If the data reflect a change in doctors' behavior, we would expect the rise in "intentional" poisonings to be coupled with a *decline* in poisonings coded as accidental or of undetermined intent.²⁰ Figure 8 shows trends in ER visits for poisonings and cuttings among teen girls by intentionality. To account for seasonality, I show trends from both 2016 and 2017.

The poisoning trends show no evidence that *13 Reasons Why* affected doctors' propensity to classify poisonings as intentional. The 2017 trends in accidental and undetermined-intent poisonings are virtually identical to their 2016 counterparts. In contrast, 2017 counts for intentional poisoning trend upward relative to 2016.

The trends for contact with sharp objects tell a different story, but it is one that supports the central thesis. Counts of ER visits for accidental contact with sharp objects are noticeably *higher* in 2017 than in 2016. If doctors were simply misclassifying accidental cuttings as intentional, however, we would see the 2017

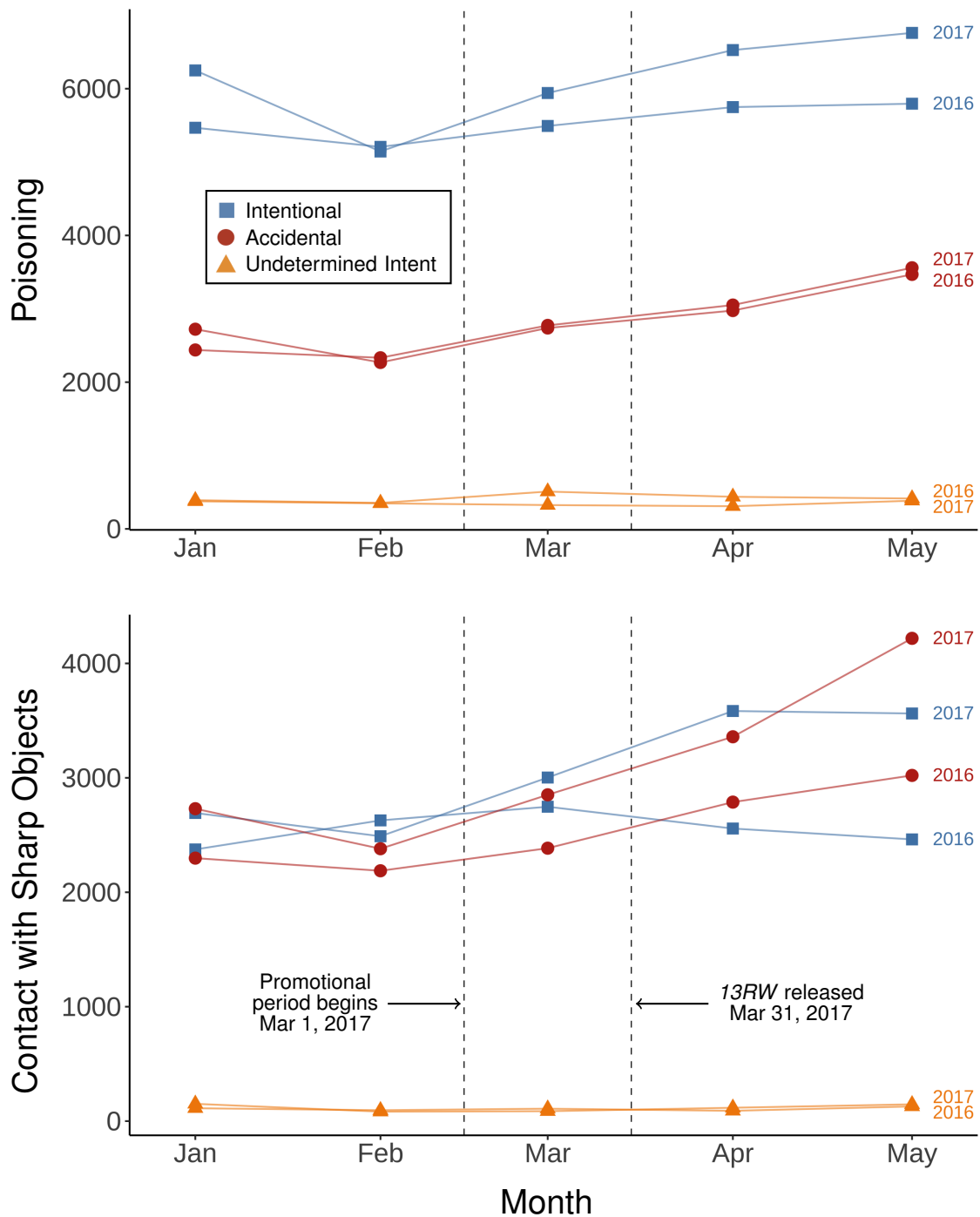


Figure 8: Injury visits by type and intent. *Notes:* Counts of visits for poisoning and contact with sharp objects. I plot 2016 data to show the trend and seasonality we might expect to see in the absence of treatment. Accidental poisonings in 2017 closely track their 2016 counterparts, but intentional poisonings after February 2017 increase relative to 2016 in the post-treatment period. This suggests that doctors were not merely reclassifying accidental poisonings as intentional. For contact with sharp objects, accidental injuries actually increase relative to 2016, indicating the main treatment effect might be underestimated. The rise in intentional self-harm visits was unaccompanied by a drop in accidental injury visits.

counts fall. The plot suggests that the estimated treatment effect may *understate* the true effect of *13 Reasons Why*, with intentional cuttings being misclassified as accidental.²¹

One potential cause for concern is the unusually high counts of ER visits for intentional poisoning, intentional cutting, and accidental cutting in January 2017. Perhaps the draw of hospitals in 2017 resulted in unusually high counts for these injuries, resulting in unusually high counts in April and May 2017. Two pieces of evidence can put us at ease. First, as I show in later sections, January counts are often high. Second, as shown in [Figure 4](#), other demographic groups show elevated counts of self-harm in January 2017 but not in April or May 2017, suggesting the elevation is specific to teen girls. This makes it unlikely that the elevated counts are the result of an unusual sample of hospitals.

Prediction Intervals Show Good but Imperfect Coverage in Repeated Placebo Tests

Conformal prediction intervals rely on weaker assumptions than conventional prediction intervals, but these assumptions might still be violated.²² One way to assess the validity of the intervals is with a repeated placebo test procedure. More precisely, this procedure will assess the validity of conformal intervals when we have used the data to select the model. We can thus see it as assessing the validity of the ITS method (as implemented in this study) as a whole.

To understand the logic of this procedure, consider one placebo test in isolation. First, we select an ARIMA model using the first half of the pre-treatment data—that is, the first 61 first- and seasonally-differenced periods. Call this the *training* period. Next, we fit the chosen model to these 61 periods. Finally, we use the fitted model to forecast the 62nd period's count of self-harm visits with a 95 percent conformal prediction interval. Call the 62nd period the *test* period.

This test period is pre-treatment, and thus represents a potential outcome under control. Consequently, the observed count should lie within the prediction interval with probability 0.95. If the prediction interval fails to cover the observed count of self-harm visits, we might doubt the validity of the intervals. If the prediction interval covers the observed count, we will say that the ITS method *passes* the placebo test.

Conducting just one placebo test, however, fails to provide an accurate picture of how well our prediction intervals cover. For instance, if nominal 95 percent prediction intervals cover the truth only 80 percent of the time, they are invalid but still provide a high probability of passing the placebo test. To better gauge the performance of these intervals, we can repeat this placebo test many times.

To repeat the test, I iteratively expand the training period, refitting the ARIMA model each time (but without changing the specification chosen for the initial test). For instance, the second placebo test fits the ARIMA model to the first 62 periods and forecasts the 63rd period with 95 percent prediction intervals. The third placebo test fits the ARIMA model to the first 63 periods and forecasts the 64th period, and so on. I repeat this procedure until I reach the post-treatment period. [Figure 9](#) and [Figure 10](#) plot the placebo test results alongside the estimated treatment effect in the

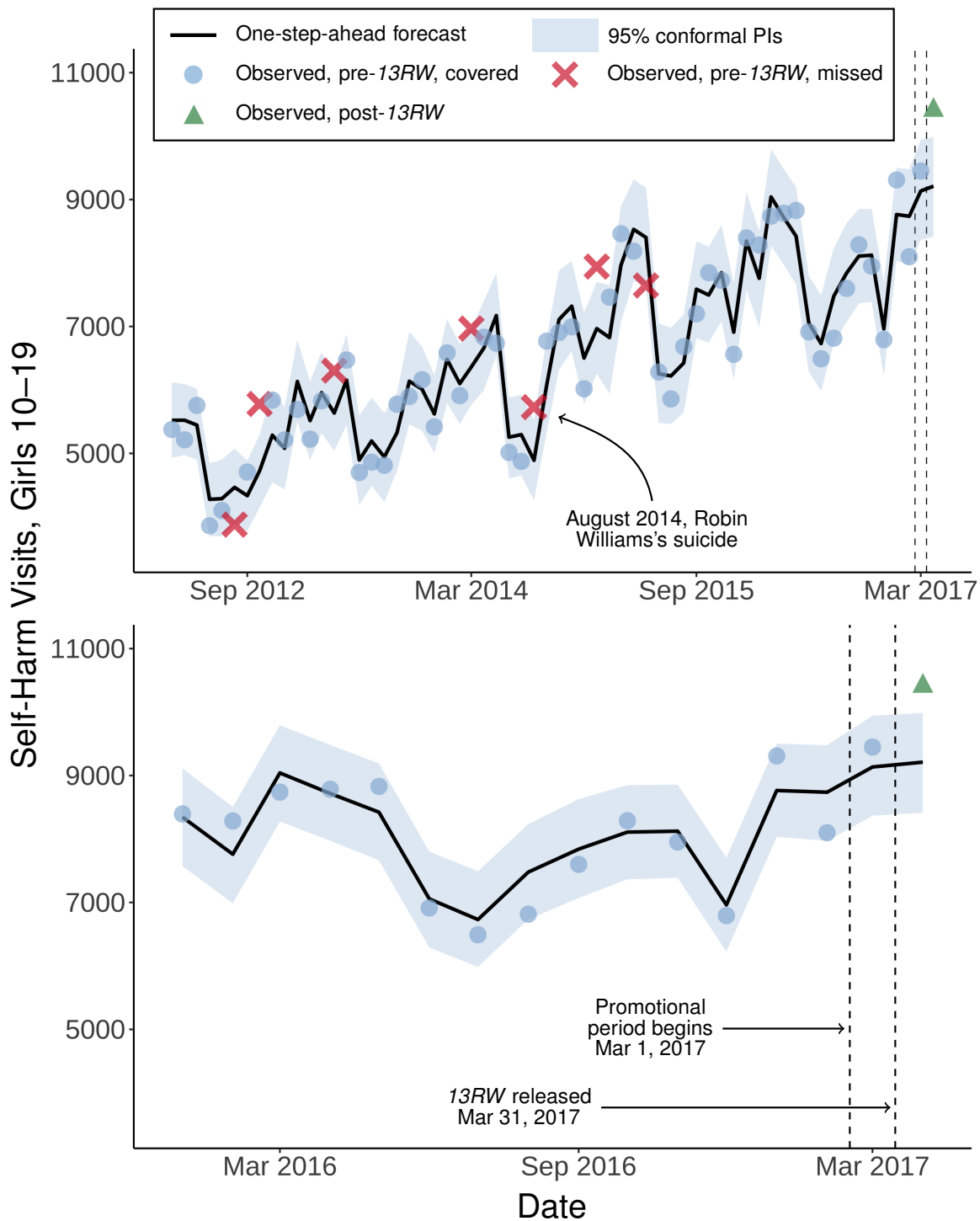


Figure 9: Repeated placebo test results. *Notes:* One-step-ahead forecast plots with 95 percent conformal prediction intervals. The black line represents out-of-sample forecasts, not fitted values. The ARIMA specification was chosen using the first half of the time series only. For the initial test period, the model is trained on the first half of the series. For each subsequent test period, an additional observation is added to the training series. Of 61 placebo tests, 7 fail, indicating good but imperfect coverage. Excluding the Robin Williams period, the numbers are 60 and 6, respectively.

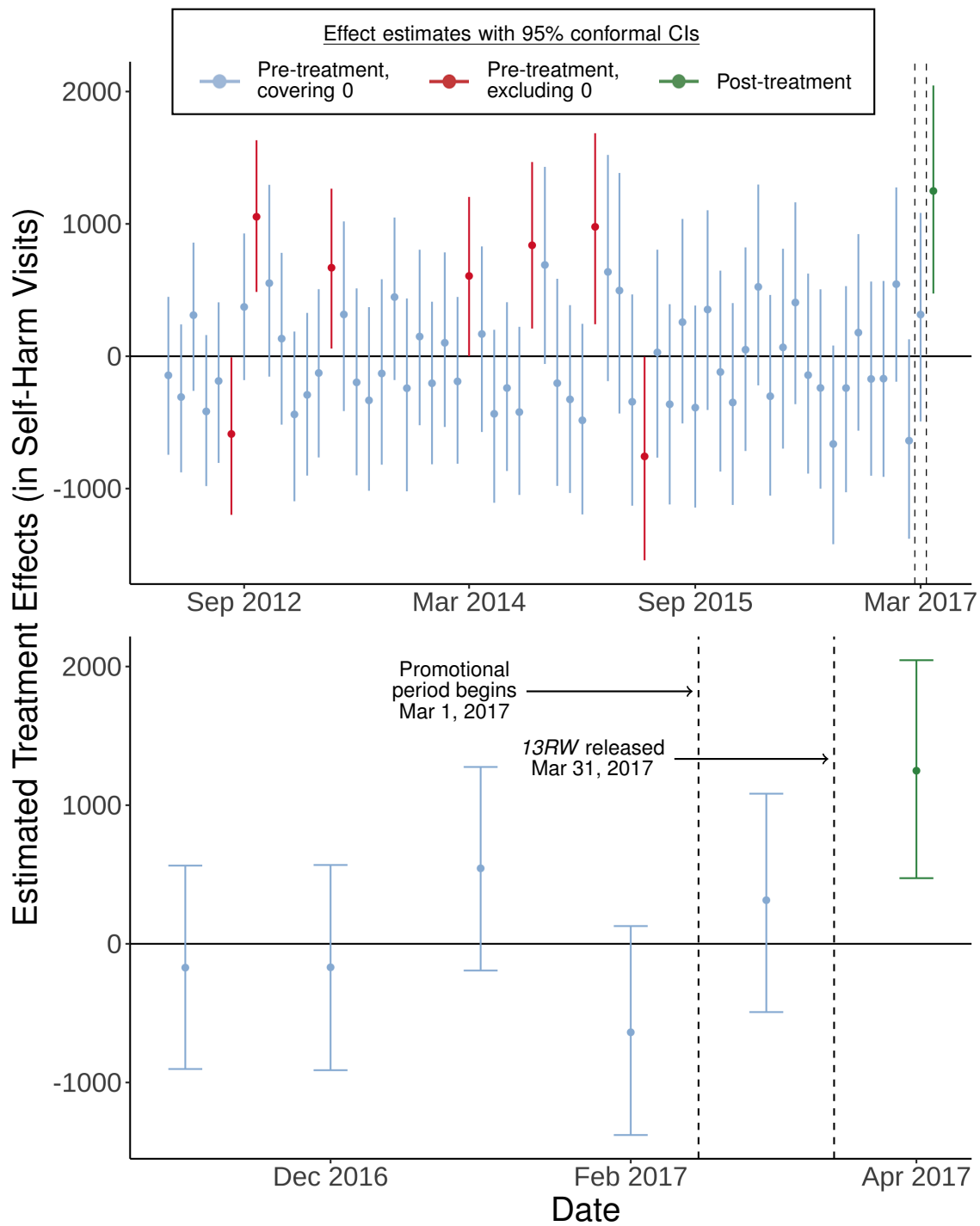


Figure 10: Repeated placebo test effect estimates. *Notes:* Treatment effect plots with 95 percent conformal confidence intervals. Effect estimates and confidence intervals are calculated using the forecasted outcomes and 95 percent conformal prediction intervals from Figure 9.

post-treatment period for comparison. Note that the black lines in [Figure 9](#) represent out-of-sample predictions—they are forecasted using models that were selected and fit using only previous time periods. The confidence intervals in [Figure 10](#) are computed by subtracting the upper and lower limits of the prediction intervals from the observed count of self-harm visits as in Chernozhukov et al. (2021).

The plots depict seven missed observations, but one of these occurs in August 2014, when Robin Williams died by suicide. Excluding this period, the conformal prediction intervals miss six out of 60 (10 percent) of observed values. This performance is neither stellar nor terrible. The intervals are likely invalid, but they do not severely undercover. Section B in the online supplement includes results with different block sizes. With a block size of 3, 90 percent conformal intervals appear to overcover, failing only 3 of 61 placebo tests. The 90 percent confidence interval in the post-treatment period still excludes 0 using this block size.

Three additional facts about the placebo test results are worth noting. First, the effect estimate for April 2017 is larger in magnitude than any placebo test estimate—including the estimate following Robin Williams's suicide. Thus, although the prediction intervals may undercover, the placebo test procedure suggests that the count of ER visits in April 2017 is the most unusual of the bunch.²³ Second, the absolute mean of the 60 placebo effect estimates is approximately 6.5—quite close to zero considering the estimated standard deviation of the placebo estimates is roughly 418. This provides evidence that treatment effects estimates suffer from little bias. Finally, the performance of the placebo tests seems to improve as the size of the training period grows, making us more confident in the main treatment effect estimate.

Could 2017 Data be Especially Noisy?

The NEDS collects data from a different sample of hospitals each year. Because the placebo effect estimate for January 2017 is somewhat high, and the estimate for February somewhat low, we might worry that the 2017 data is especially noisy due to the selected sample. If true, this might undermine our confidence in the accuracy of the April 2017 treatment effect estimate.

To explore this possibility, I first present the absolute placebo effect estimates by month in [Figure 11](#). The reason to present results by month is that the model appears to predict some months—like December and November—more accurately than others—like August and January. The pre-treatment 2017 estimates, highlighted in red, are not unusual relative to previous years' estimates for those months. We should exercise caution—we have only five to six estimates for each month, and there are other outliers, like October 2012 and January 2015—but the plot suggests that pre-treatment 2017 data is not exceedingly unusual.

One issue with the [Figure 11](#) is that it contains only four effect estimates from 2017. To further examine the 2017 data, I fit an ARIMA model to the entire time series up until December 2017 and plot the absolute residuals by month in [Figure 12](#). [Figure 13](#) plots the residuals sequentially. The residuals for April and May 2017 are noticeably higher than all others, but the remaining 2017 residuals are not especially large. The figures also provide evidence that predictive accuracy and model fit

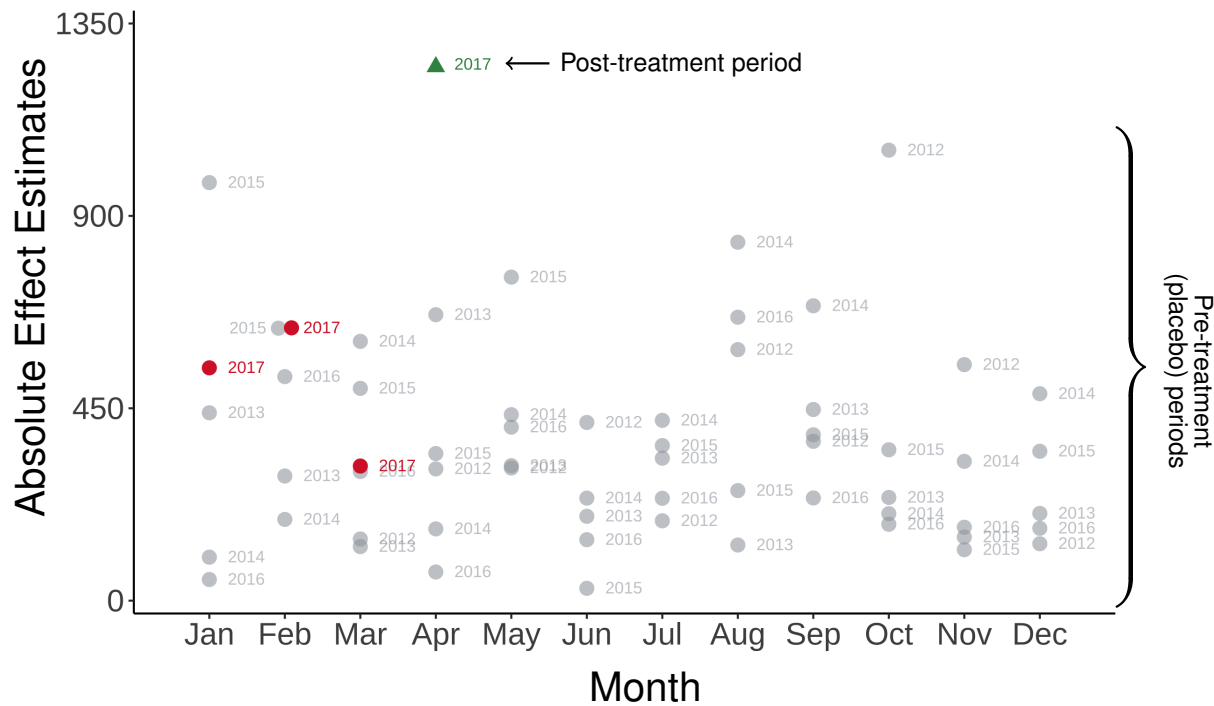


Figure 11: Repeated placebo test results by month. *Notes:* Placebo effect estimates in absolute value organized by month. Pre-treatment estimates from 2017 are red circles, and the post-treatment estimate is a green triangle. Although the estimates for January–March of 2017 are slightly elevated, they are not especially high for those months. The effect estimates for February 2015 and February 2017 are almost exactly equal and thus plotted side-by-side. The model appears to predict more accurately for the final three months of the year.

remain constant over time—they do not worsen as we approach the treatment period—which can make us more confident in the effect estimates for April 2017. The pattern is not especially model-dependent: Section B in the online supplement shows that other model specifications produce similar patterns.

Model-Free Plot Suggests a Causal Effect

The results of the main analysis remain virtually unchanged across multiple model specifications and model-selection strategies (see Section B in the online supplement). But we might still be concerned that are results are too model-dependent (Zoorob 2020). That is, there may be other plausible models that fit the data equally well but produce different causal estimates (King and Zeng 2006; King and Nielsen 2019).

Perhaps the simplest way to assuage concerns about model dependence is to present model-free results as in Figure 14. The plot shows the estimated number of ER visits for intentional self-harm among girls age 10–19 from October through May across several years.

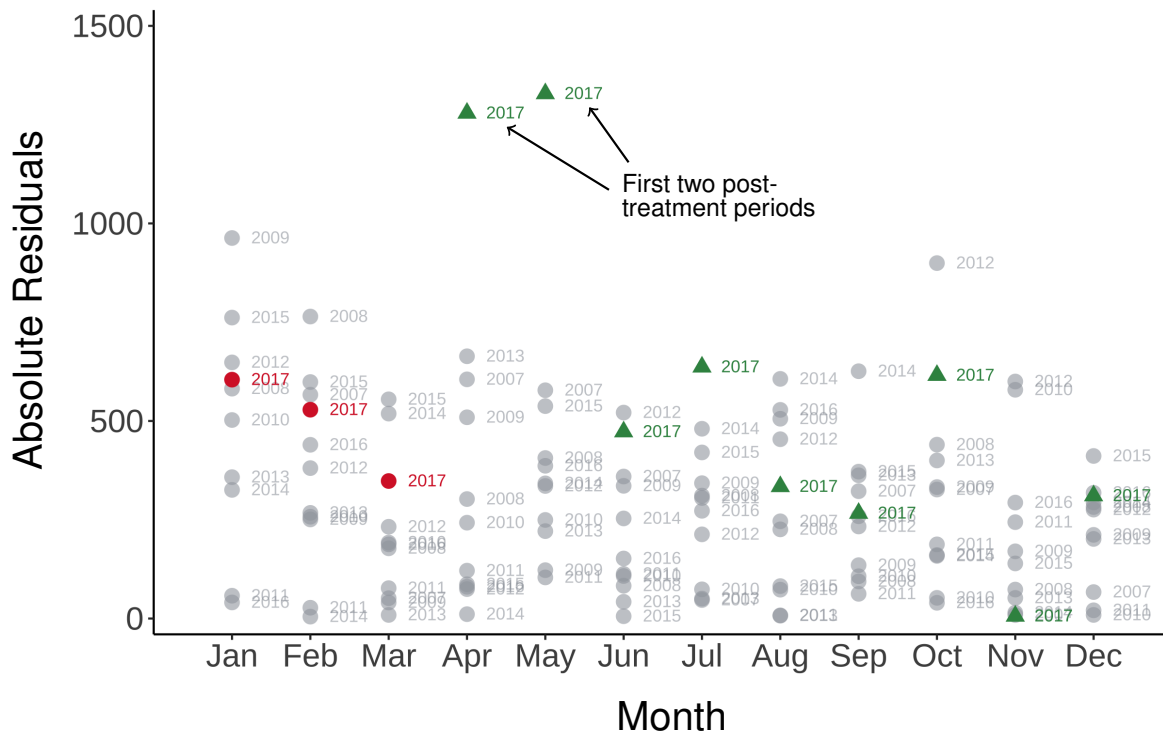


Figure 12: Absolute residuals by month. *Notes:* Absolute residuals from an ARIMA model fit to the entire differenced time series (February 2007 to December 2017), by month. Pre-treatment residuals in 2017 are red circles, and post-treatment residuals are green triangles. The residuals for April and May 2017 are much larger than other 2017 residuals. Especially noisy 2017 data cannot explain the post-treatment spike in self-harm visits.

The regular dips in December and spikes in January illustrate why simple before-and-after plots can be misleading in applications like this: suicide and self-harm rates exhibit strong seasonality and holiday effects, with particularly high suicide mortality on New Year’s Day (Phillips and Wills 1987; Beauchamp et al. 2014).²⁴ Comparing trajectories by year provides a clearer picture of how unusual a particular month’s count of ER visits is.

At a first glance, it looks like 2017’s unusual trajectory begins *before* the onset of treatment. Indeed, it is plausible that *13 Reason Why’s* promotional period had at least some effect on intentional self-harm. But it is worth examining the plot more carefully: the increase from February 2017 to March 2017 is not especially unusual relative to other February–March shifts. It is typical for the count to rise in March. The increase from March 2017 to April 2017, in contrast, is much more striking. In only one other year—2013—does the count increase between March and April (including years not shown on the plot). It is also worth noting that, because *13 Reasons Why* came out on March 31st, some of the increase in March could have been caused by the show’s release.

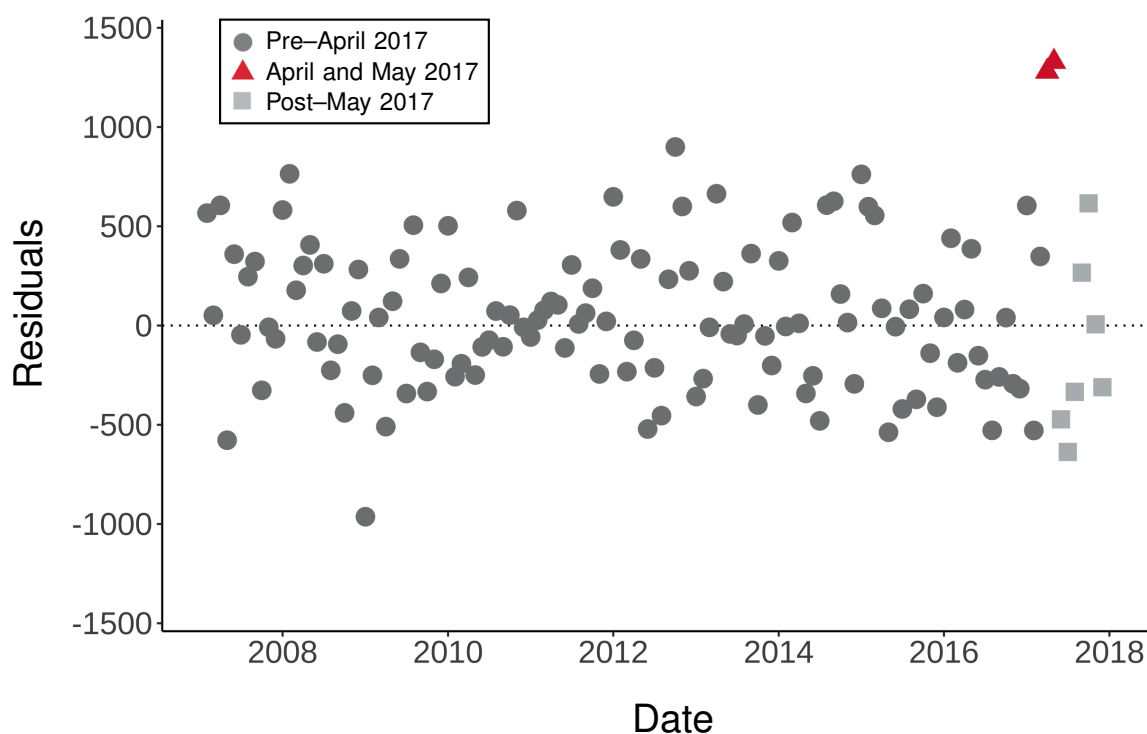


Figure 13: Residuals across time. *Notes:* Residuals from an ARIMA model fit to the entire differenced time series (February 2007 to December 2017). Model fit remains constant over time, and April and May 2017 are noticeable outliers.

Difference-in-Differences Estimates Provide Evidence of Causation

The ITS analyses for teen boys, men age 40–65, and women age 40–65 suggest the show’s release had negligible effects on self-harm visits for these groups. Below, I use these groups as control groups in difference-in-differences (DID) analyses.

DID depends on a parallel trends assumption (Angrist and Pischke 2009). Because this assumption likely fails, I present results in the form of an *event-study plot* in Figure 15, which helps us gauge what typical deviations from parallel trends look like in this setting (Ben-Michael et al. 2021). In this plot, each point represents an estimate from a simple, two-period, two-group DID analysis. For instance, a point for November 2011 represents the point estimate of a DID analysis using October and November 2011. The points for all pre-treatment periods thus serve as placebo tests: there was no treatment assigned during these periods, so the point estimates should be close to zero in expectation. We can also view the pre-treatment estimates as a null distribution—that is, the distribution of estimates we would expect to see if the treatment effect were zero.

Figure 15 shows that the effect estimate in April 2017, the post-treatment month, is unusually large relative to the distribution of pre-treatment effect estimates. Across all three control groups, the April 2017 estimate is near the 95th percentile of the pre-treatment point estimates.

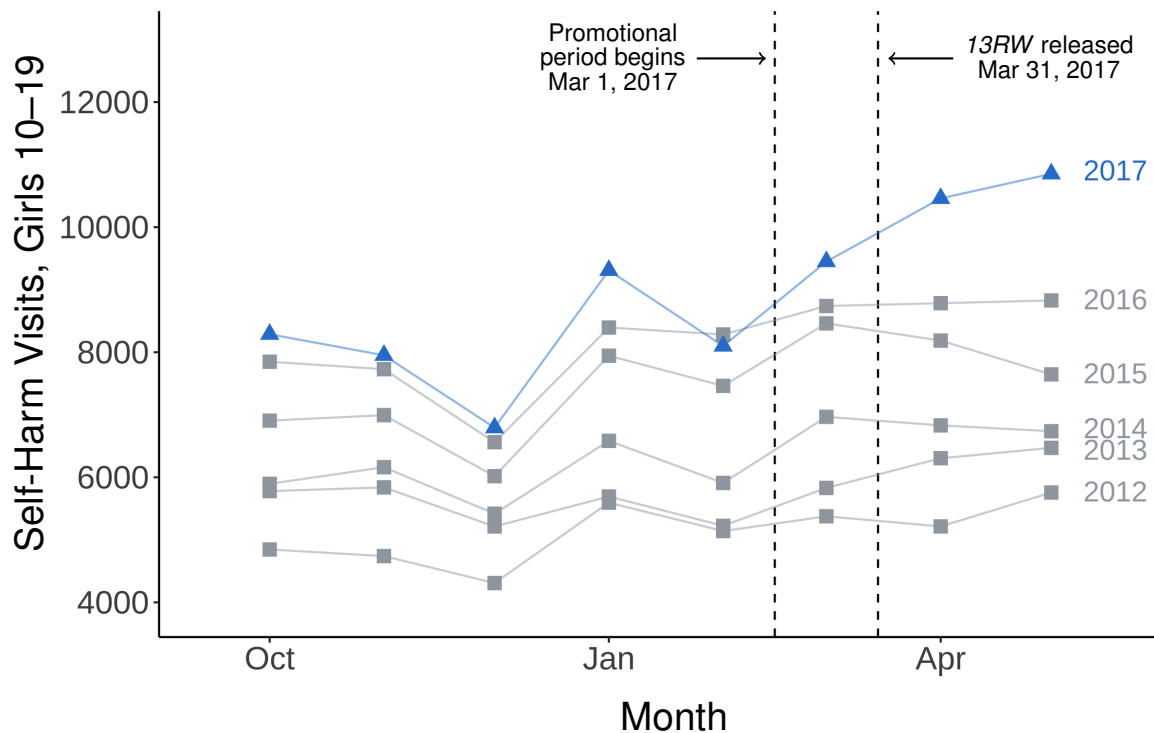


Figure 14: Model-free plot of the time series. *Notes:* Estimated number of ER visits for intentional self-harm among girls aged 10–19 from October through May. The trend for October 2016–May 2017 is colored blue; previous series are colored gray. Like suicides, self-harm visits spike in January due to New Year’s Day and fall in February. The number of ER visits for intentional-self harm rises following the beginning of the promotional period for *13 Reasons Why*, although an increase from February to March is typical for the time series. The increase following the show’s release is much more unusual.

A Causal Effect of 13 Reasons Why Remains the Most Plausible Explanation for the Post-Release Spike in Self-Harm Visits

Before concluding, it is worth briefly reviewing the evidence presented in this study. Although the ITS estimator is likely biased, placebo tests on plausible control groups—and on pre-treatment data for teen girls—indicate the bias is probably small (“Results”, “Prediction Intervals Show Good but Imperfect Coverage in Repeated Placebo Tests”). ER visits for intentional cutting are particularly high in April and May 2017, matching the method of suicide depicted in *13 Reasons Why*, and the treatment effect persists for two months, consistent with show-related social media activity (“Understanding the *13 Reasons Why* Effect”). Self-harm visits for teen girls show slight elevation prior to the show’s release, but this was plausibly caused by Netflix’s promotion of the series. Skeptics of such a promotional effect, however, need not worry: an unusually high count of self-harm visits in March 2017 should prompt our model to forecast an even higher value for April 2017, not lower. Furthermore, both model residuals and careful inspection of the raw data suggest

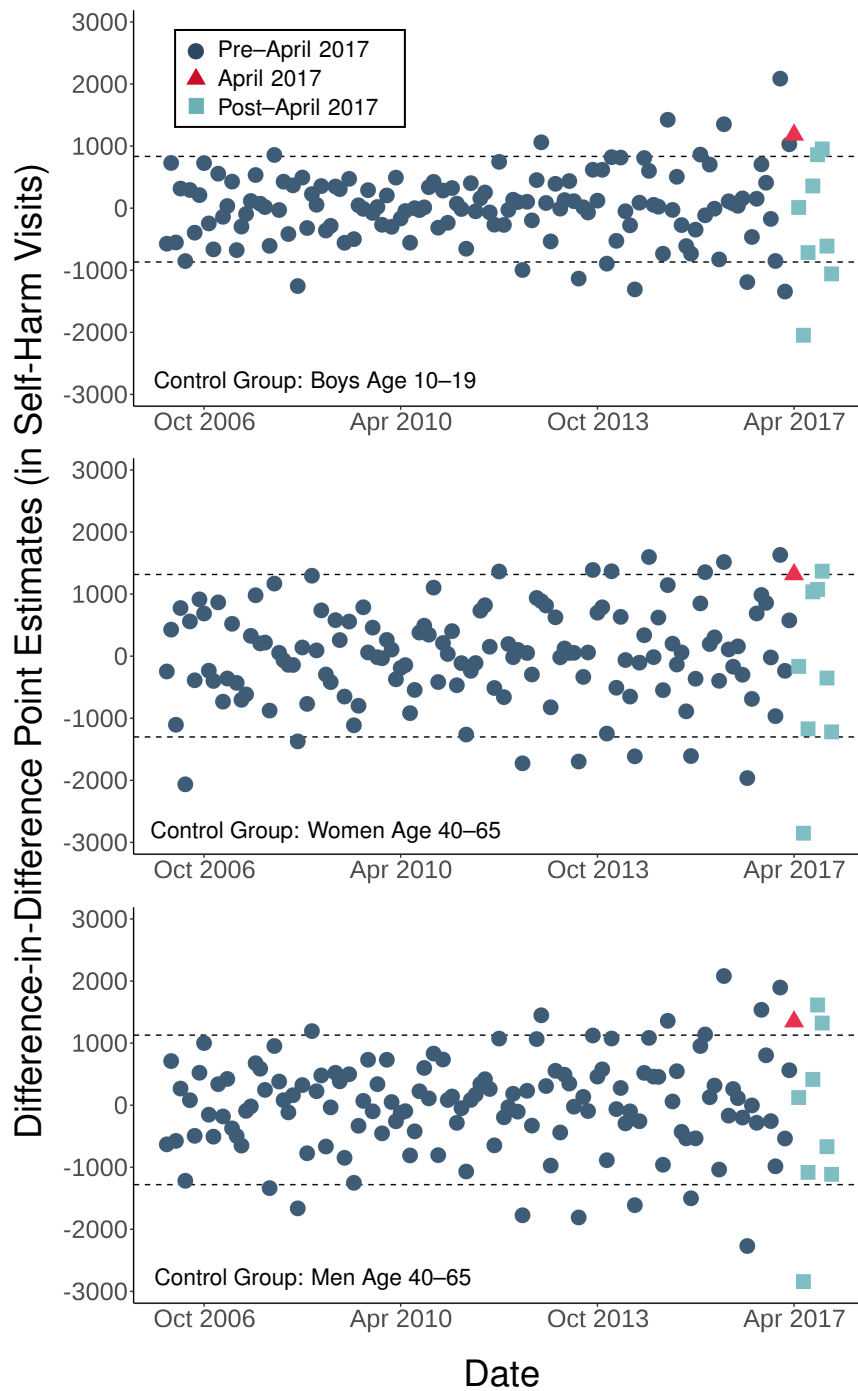


Figure 15: Difference-in-differences point estimates across time. *Notes:* Dashed lines represent 5th and 95th percentiles of the distribution of pre-treatment point estimates. Across all three plots, the point estimate for April 2017 is unusually high, hovering around the 95th percentile of pre-treatment estimates.

that April and May's counts are far more unusual than March's. Finally, self-harm visits return to expected levels by June 2017 ("Understanding the 13 Reasons Why Effect"). If the elevated counts in April and May merely reflect secular trends in the outcome, such a return would represent an extraordinary coincidence.

Conclusion

In this study, I presented evidence that the release of *13 Reasons Why* led to a temporary spike in the number of ER visits for intentional self-harm among teenage girls. The study provides new evidence that fictional stories can affect self-harm rates, contributing to literatures on both suicide contagion and the effects of fiction on behavior. It provides additional support for theories of suicide that emphasize the importance of imitation, a view Durkheim famously dismissed (Abrutyn and Mueller 2014b). In contrast to work on celebrity suicides, this study shows that the deceased need not be high-status or widely admired to prompt imitation effects. Furthermore, the imitation of method suggests that cognitive accessibility may partly explain suicide contagion.

The study also showcases how to make credible causal claims when data is imperfect and identification assumptions are violated. ITS requires implausibly strong assumptions, and measurement error likely affects the time series of self-harm visits. Accordingly, I make no claim that treatment effect estimates are exactly unbiased or that confidence intervals have perfectly valid coverage. Instead, I conduct a series of supplementary analyses that help diagnose how severe the bias and undercoverage might be.²⁵ The results show that—despite the limits of the data and identification strategy—a causal effect of *13 Reasons Why's* release remains the most plausible explanation for the large post-release spike in self-harm visits.

Notes

- 1 Why the focus on celebrity deaths? Suicide is a rare event, so we are typically forced to rely on aggregate suicide mortality data to identify causal effects, usually using some variant of the interrupted time series design. This design yields low statistical power, so the only effects we can reliably detect are fairly large ones. It makes sense, then, to focus primarily on extremely widely-reported suicide deaths to estimate treatment effects, which usually means relying on celebrity suicides.
- 2 I should emphasize, however, that an ER visit for self-harm does not necessarily indicate a suicide attempt.
- 3 The status explanation is related to what Abrutyn and Mueller (2014b) call Tarde's *law of prestige imitation*, and the personal-identification explanation fits with the *law of propinquity*.
- 4 The stylized interactions between *J* and *K* help illustrate a subtler point: even if we had person-level data on who watched the show and who self-harmed, interference across people would pose problems for causal inference—i.e., *J's* treatment status (and outcome) might affect *K's* outcome, violating the stable-unit-treatment-value assumption.
- 5 The resulting counts can be found at <https://github.com/cmfelton/13rw>, along with code for replicating the counts from the original NEDS data.

- 6 In April 2017, the first post-treatment month, only 0.28 percent of self-harm visits among teen girls were coded by the NEDS as fatal. The average monthly proportion in 2017 is 0.46 percent. Curiously, the average proportion in 2006 is 5.2 percent but falls to 1.2 percent by 2008, with an abrupt drop occurring in July 2007. This almost certainly reflects changes to the NEDS rather than a genuine trend.
- 7 In the former setting, difference-in-differences will be impossible to use, and in the latter setting, it would provide poor effect estimates. Synthetic control methods are feasible in the latter scenario but require many control units, which are unavailable in the case of *13 Reasons Why*.
- 8 I avoid describing this hypothetical celebrity death as a “confounder” because this death would not affect the release of *13 Reasons Why* (the primary treatment).
- 9 This problem is not unique to the ITS setting—it is relevant in all settings where we use the same data to both select and fit a statistical model (see, e.g., Freedman (1983), Faraway (1992), or Kuchibhotla et al. (2022)).
- 10 Box (1976) famously states that “all models are wrong,” and Cox (1995) adds that “the very word ‘model’ implies simplification and idealization.” White (1982), Aronow and Miller (2019), and Buja et al. (2019) express similar sentiments, to name just a few.
- 11 Resulting forecasts are then de-differenced for the plots. See Section E of the online supplement for more details.
- 12 This procedure produces intervals that are only *approximately* exact. Exactness is guaranteed in the i.i.d. setting.
- 13 More carefully, if model misspecification is the source of the bias, conformal inference can provide valid confidence intervals. We still have to assume the model specification remains unchanged in the post-treatment period.
- 14 High-variance estimators like ITS can systematically exaggerate effect estimates conditional on statistical significance (Gelman and Carlin 2014). I explore this possibility using simulations in Section D of the online supplement and conclude the risk is probably small in this setting.
- 15 First- and seasonal-differencing is performed on all three time series. For teen boys, however, the Hyndman–Khandakar algorithm chooses a specification that omits first-differencing (Hyndman and Khandakar 2008). I estimate treatment effects with this alternative specification in Section B of the online supplement. Estimated treatment effects are virtually identical.
- 16 One complication worth mentioning is that Chris Cornell, frontman of Soundgarden and Audioslave, died by suicide on May 18, 2017. This plausibly had *some* effect on self-harm visits in May, but we would not expect teen girls in 2017 to be strongly affected by this death. Given the elevated counts in April—the month before his suicide—a causal effect of *13 Reasons Why* remains the most plausible explanation.
- 17 Note that the time series of total self-harm visits is not plagued by these problems—see Section B in the online supplement.
- 18 If we omit post–May 2017 months, the percentiles are 95th and 85th for April and May 2017, respectively.
- 19 Alternatively, the show’s release may have caused physicians to *accurately* classify self-harm behavior that, in the absence of the show’s release, they would have *misclassified* as accidental.
- 20 More carefully, we would expect a decline in such poisonings *relative to the counterfactual trend* we would observe in the absence of *13 Reason Why’s* release.

- 21 The plot also speaks to the worry that the show may have caused teen girls who would have self-harmed regardless to be more likely to seek help for their injuries. If this were the case, we would not expect a rise in accidental cuttings: such an increase indicates that at least some girls were concealing the intentionality of their behavior. The observed increase in suicide mortality for boys and girls should also shift our expectations about the likelihood of a genuine change in self-harm behavior (Niederkrontenthaler et al. 2019). It remains possible, however, that a change in help-seeking behavior explains some of the total rise.
- 22 See Section E in the online supplement for explanation of the assumptions.
- 23 This line of reasoning is close in spirit to the *inductive* conformal inference approach of Papadopoulos (2008).
- 24 The spikes in January could *partly* result from changing hospital samples in the NEDS, but population-level suicide mortality data from the CDC show very similar spikes in January.
- 25 In the spirit of, e.g., Cornfield et al. (1959) or Rosenbaum (2015).

References

- Abrutyn, Seth and Anna S. Mueller. 2014a. "Are Suicidal Behaviors Contagious in Adolescence? Using Longitudinal Data to Examine Suicide Suggestion." *American Sociological Review* 79:211–227.
- Abrutyn, Seth and Anna S. Mueller. 2014b. "Reconsidering Durkheim's Assessment of Tarde: Formalizing a Tardian Theory of Imitation, Contagion, and Suicide Suggestion." *Sociological Forum* 29:698–719.
- Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Aronow, P.M. and Benjamin T Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press.
- Baicker, Katherine and Theodore Svoronos. 2019. "Testing the Validity of the Single Interrupted Time Series Design." Technical report, National Bureau of Economic Research.
- Beauchamp, Gillian A., Mona L. Ho, and Shan Yin. 2014. "Variation in Suicide Occurrence by Day and During Major American Holidays." *The Journal of Emergency Medicine* 46:776–781.
- Ben-Michael, Eli, Avi Feller, and Elizabeth A. Stuart. 2021. "A Trial Emulation Approach for Policy Evaluations with Group-Level Longitudinal Data." *Epidemiology* 32:533.
- Berk, Richard A. 2021. "Post-Model-Selection Statistical Inference with Interrupted Time Series Designs: An Evaluation of an Assault Weapons Ban in California." *arXiv preprint arXiv:2105.10624*.
- Box, George E.P. 1976. "Science and Statistics." *Journal of the American Statistical Association* 71:791–799.
- Bridge, Jeffrey A., Joel B. Greenhouse, Donna Ruch, Jack Stevens, John Ackerman, Arielle H. Sheftall, Lisa M. Horowitz, Kelly J. Kelleher, and John V. Campo. 2020. "Association Between the Release of Netflix's 13 Reasons Why and Suicide Rates in the United States: An Interrupted Time Series Analysis." *Journal of the American Academy of Child & Adolescent Psychiatry* 59:236–243.

- Brito, Christopher. 2019. "Netflix Deletes Graphic Suicide Scene from First Season of '13 Reasons Why'." *CBS News* <https://www.cbsnews.com/news/13-reasons-why-suicide-scene-hannah-baker-season-finale-death-katherine-langford-season-1-episode-13/>.
- Buja, Andreas, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. 2019. "Models as Approximations I." *Statistical Science* 34:523–544.
- Canetto, Silvia Sara and Isaac Sakinofsky. 1998. "The Gender Paradox in Suicide." *Suicide and Life-Threatening Behavior* 28:1–23.
- Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu. 2021. "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls." *Journal of the American Statistical Association* 116:1849–1864.
- Cornfield, Jerome, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L. Wynder. 1959. "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions." *Journal of the National Cancer Institute* 22:173–203.
- Cox, D.R. 1995. "Comment on 'Model Uncertainty, Data Mining and Statistical Inference'." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 158:455–456.
- Domaradzki, Jan. 2021. "The Werther Effect, the Papageno Effect, or No Effect? A Literature Review." *International Journal of Environmental Research and Public Health* 18:2396.
- Durkheim, Emile. 1897. *Le Suicide: Étude de Sociologie*. Alcan.
- D'Zurilla, Christie. 2019. "'13 Reasons Why' Influenced the Suicide Rate? It's Not That Simple." *The Los Angeles Times* <https://www.latimes.com/entertainment/tv/la-et-st-13-reasons-why-suicide-study-netflix-20190501-story.html>.
- Faraway, Julian J. 1992. "On the Cost of Data Analysis." *Journal of Computational and Graphical Statistics* 1:213–229.
- Fink, David S., Julian Santaella-Tenorio, and Katherine M. Keyes. 2018. "Increase in Suicides the Months After the Death of Robin Williams in the US." *PLoS One* 13:e0191405.
- Freedman, David A. 1983. "A Note on Screening Regression Equations." *The American Statistician* 37:152–155.
- Gelman, Andrew and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9:641–651.
- Gould, Madelyn S. and David Shaffer. 1986. "The Impact of Suicide in Television Movies." *New England Journal of Medicine* 315:690–694.
- Grady, Constance. 2019. "Why It's So Hard to Prove That '13 Reasons Why' Caused an Increase in Suicide." *Vox* <https://www.vox.com/culture/2019/5/3/18522559/13-reasons-why-netflix-youth-suicide-rate>.
- Griffith, Alan. 2022. "Name Your Friends, but Only Five? The Importance of Censoring in Peer Effects Estimates Using Social Network Data." *Journal of Labor Economics* 40:779–805.
- Healthcare Cost and Utilization Project. 2021. "Suicidal Ideation, Suicide Attempt, or Self-Inflicted Harm: Pediatric Emergency Department Visits, 2010–2014 and 2016."

- Hyndman, Rob J. and Yeasmin Khandakar. 2008. "Automatic Time Series Forecasting: the forecast Package for R." *Journal of Statistical Software* 27:1–22.
- King, Gary and Richard Nielsen. 2019. "Why Propensity Scores Should Not be Used for Matching." *Political Analysis* 27:435–454.
- King, Gary and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14:131–159.
- Kuchibhotla, Arun K., John E. Kolassa, and Todd A. Kuffner. 2022. "Post-Selection Inference." *Annual Review of Statistics and Its Application* 9:505–527.
- Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. "Distribution-Free Predictive Inference for Regression." *Journal of the American Statistical Association* 113:1094–1111.
- Libbey, Peter. 2018. "Netflix Adds a Warning Video to '13 Reasons Why'." *The New York Times* <https://www.nytimes.com/2018/03/22/arts/television/netflix-warning-video-13-reasons-why.html>.
- Miratrix, Luke W. 2022. "Using Simulation to Analyze Interrupted Time Series Designs." *Evaluation Review* 46:750–778.
- Morgan, Stephen L. and Christopher Winship. 2015. *Counterfactuals and Causal Inference*. Cambridge University Press.
- Niederkrotenthaler, Thomas, Steven Stack, Benedikt Till, Mark Sinyor, Jane Pirkis, David Garcia, Ian R.H. Rockett, and Ulrich S. Tran. 2019. "Association of Increased Youth Suicides in the United States with the Release of '13 Reasons Why'." *JAMA Psychiatry* 76:933–940.
- Owens, Pamela L., Kimberly W. McDermott, Rachel N. Lipari, and Megan M. Hambrick. 2020. "Emergency Department Visits Related to Suicidal Ideation or Suicide Attempt, 2008–2017." Technical Report 263, Healthcare Cost and Utilization Project.
- Paluck, Elizabeth Levy. 2009. "Reducing Intergroup Prejudice and Conflict Using the Media: A Field Experiment in Rwanda." *Journal of Personality and Social Psychology* 96:574.
- Papadopoulos, Harris. 2008. "Inductive Conformal Prediction: Theory and Application to Neural Networks." In *Tools in Artificial Intelligence*. Citeseer.
- Patterson, Orlando. 2014. "Making Sense of Culture." *The Annual Review of Sociology* 40:1–30.
- Phillips, David P. 1974. "The Influence of Suggestion on Suicide: Substantive and Theoretical Implications of the Werther effect." *American Sociological Review* pp. 340–354.
- Phillips, David P and John S Wills. 1987. "A Drop in Suicides Around Major National Holidays." *Suicide and Life-Threatening Behavior* 17:1–12.
- Pirkis, Jane, R. Warwick Blood, Annette Beautrais, Philip Burgess, and Jaelea Skehan. 2006. "Media Guidelines on the Reporting of Suicide." *Crisis* 27:82–87.
- Romer, Daniel. 2020a. "Reanalysis of the Bridge et al. Study of Suicide Following Release of 13 Reasons Why." *PLoS One* 15:e0227545.
- Romer, Daniel. 2020b. "Reanalysis of the Effects of '13 Reasons Why': Response to Bridge et al." *PLoS One* 15:e0239574.

- Rosenbaum, Paul. 2015. "How to See More in Observational Studies: Some New Quasi-Experimental Devices." *Annual Review of Statistics and Its Application* 2:21–48.
- Shafer, Glenn and Vladimir Vovk. 2008. "A Tutorial on Conformal Prediction." *Journal of Machine Learning Research* 9.
- Shalizi, Cosma Rohilla and Andrew C Thomas. 2011. "Homophily and Contagion are Generically Confounded in Observational Social Network Studies." *Sociological Methods & Research* 40:211–239.
- Stack, Steven, Michael Kral, and Teresa Borowski. 2014. "Exposure to Suicide Movies and Suicide Attempts: A Research Note." *Sociological Focus* 47:61–70.
- Tankard, Margaret E. and Elizabeth Levy Paluck. 2016. "Norm Perception as a Vehicle for Social Change." *Social Issues and Policy Review* 10:181–211.
- Tarde, Gabriel de. 1903. *The Laws of Imitation*. H. Holt.
- White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica: Journal of the Econometric Society* pp. 1–25.
- Zoorob, Michael. 2020. "Do Police Brutality Stories Reduce 911 Calls? Reassessing an Important Criminological Finding." *American Sociological Review* 85:176–183.

Acknowledgments: For helpful discussions and feedback relevant to this project, I thank (in reverse-alphabetical order) Brandon Stewart, Varun Satish, Momoko Nishikido, Ian Lundberg, Marielle Côté-Gendreau, Dalton Conley, members of the Stewart Lab, the editor, and the anonymous referees. Replication data and code can be found at <https://github.com/cmfelton/13rw>. All errors are my own.

Chris Felton: Postdoctoral Fellow, Graduate School of Education, Harvard University. This study was completed while the author was a PhD student in the Department of Sociology and Office of Population Research at Princeton University.
E-mail: christopher_felton@gse.harvard.edu.