Do Conspiracy Theorists Update Their Priors?

Evidence From COVID-19

Luke Frymire & Mikhael Gaster

Econ 594

Abstract

COVID-19 is a deadly pandemic, yet many individuals still doubt its veracity and severity. Laboratory evidence shows that conspiracy theorists update their beliefs in some scenarios, and anecdotal evidence suggests that many individuals revise their COVID-19 doubts once it seriously harms them or a loved one, however, to our knowledge, this is the first study to measure belief-updating in COVID-19 conspiracy theorists in a real-world setting. We study individual-level COVID-19 misinformation and beliefs on Twitter and examine how users' beliefs change in response to serious local outbreaks. We find some evidence that COVID-19 deniers become less conspiratorial after increased cases and deaths nearby, although we interpret these results with caution due to data concerns.

Introduction

While the COVID-19 pandemic captured the world's attention beginning in early 2020, another epidemic spread alongside it. Misinformation about COVID-19, its veracity, and its origins began to take hold in many communities, particularly through online social media networks. It is easy to laugh at conspiracy theories and dismiss the people who believe them, but these beliefs deserve more respect than their often comical presentations merit – they can and do cause serious harm. Tragically, there are many stories of individuals who did not believe that COVID-19 was real (or dangerous) until it seriously harmed them or a loved one (Srikanth, 2020). Understanding how and when conspiracy theorists change their beliefs is imperative to combatting COVID-19, but it is also important for facing other challenges which rely on social cohesion, such as climate change.

Given limited evidence suggesting that individuals who do not believe COVID-19 to be a serious threat often only change their mind upon being presented with first-hand evidence to the contrary, we investigate the following question empirically: Do COVID-19 conspiracy theorists (and those misinformed with respect to COVID-19)[1] change their inaccurate beliefs after a friend or loved one contracts COVID-19?

We focus on COVID-19 conspiracy theorists (rather than conspiracy theorists in general) for several reasons. First, events which may cause a conspiracy theorist to update their beliefs are indirectly observable in the case of COVID-19 conspiracies. As the frequency of COVID-19 cases or deaths increases in a given locale, it is reasonable to assume that individuals in the locale become more likely to observe first-hand evidence that COVID-19 is dangerous and real (either by catching the virus themselves, or through observing the sickness of a close family member or friend). We

[1] For brevity and convenience, we refer to both "conspiracy theorists" and "misinformed individuals" as "conspiracy theorists" in this paper.

therefore use local per-capita case and death rates as a proxy for the occurrence of these belief-updating events. This stands in contrast to other conspiracy theories – for instance, it is difficult to measure the occurrence of events where "flat-earthers" are exposed to contradictory evidence outside of a controlled experimental setting (if such events even exist).

Thesecond reason for focusing on COVID-19 conspiracy theories is that these belief-contradicting events occurred with considerable variation in timing across locations. Within the United States, per-capita case and death rates spiked at various times across counties and states. Crucially, we argue that the *timing* of exposure to a belief-contradicting event is more-or-less exogenous with respect to an individual's beliefs, creating a quasi-natural experiment which can be exploited for causal inference.[2]

Finally, COVID-19 conspiracy theories are popular, dangerous, and believed by a wide swath of individuals (Schaeffer, 2020), whereas many other conspiracy beliefs are held by relatively niche subpopulations and are considerably less consequential in their political and social implications. This both motivates the importance of understanding the phenomenon and provides a large and diverse sample.

Although there are possible several ways to measure individual-level beliefs regarding COVID-19, we believe that social media provides the ideal medium for our purposes. Compared to studies conducted by survey or within a laboratory setting, social media provides observational data about the real-world actions of conspiracy theorists. While our analysis could be conducted on other social

[2] Although COVID-19 conspiracy theorists are less likely to take necessary measures to avoid COVID-19, and therefore more likely to contract the disease, their beliefs are not the main determinant of when they contract the disease. We therefore argue that the timing is far more dependent on the number of cases around an individual than it is on their beliefs.

media platforms (such as Facebook or Reddit), Twitter has several advantages that place it firmly atop our list. First, Twitter users post or retweet frequently, so we get detailed data on users over time – this reduces measurement error when examining COVID-19 beliefs with respect to COVID-19 cases or deaths over time. Second, Twitter posts ("tweets") are short enough that users often favor hashtags and keywords over lengthy posts. This systematic language structure makes automated analysis much easier than it would be on the lengthier posts that we would find on other social media platforms. Finally, some Twitter users provide their geolocation, unlike those on many other social networks; this is a central feature for our research design.

Our research design is broadly structured as follows: first, we use a carefully selected list of hashtags, keywords, and URLs to classify some Twitter users as "potential" conspiracy theorists, and attempt to geolocate them using information from their tweets and user biographies ("bios"). For the subset of potential conspiracy theorists who have usable geolocations, we employ a web scraping tool to collect the entirety of their Twitter activity beginning in 2019. This data is used to make a final determination as to which "potential" conspiracy theorist make it to the list of "actual" conspiracy theorists. Next, we aggregate conspiracy theorist users and their Twitter activity over the course of the COVID-19 pandemic to a county-by-week granularity.[3] Finally, we examine how aggregate Twitter activity changes as local COVID-19 conditions change.

Analyzing multiple countries would introduce a host of problems – different national guidelines and lockdowns, as well as cultures, beliefs, and governments would make comparisons difficult at best, and misleading at worst. We therefore narrow our focus to one country – the United States – which has a few characteristics that make it an ideal candidate for our quasi-natural experiment: first, it has nearly four times more Twitter users than the next largest English-speaking country (Statista,

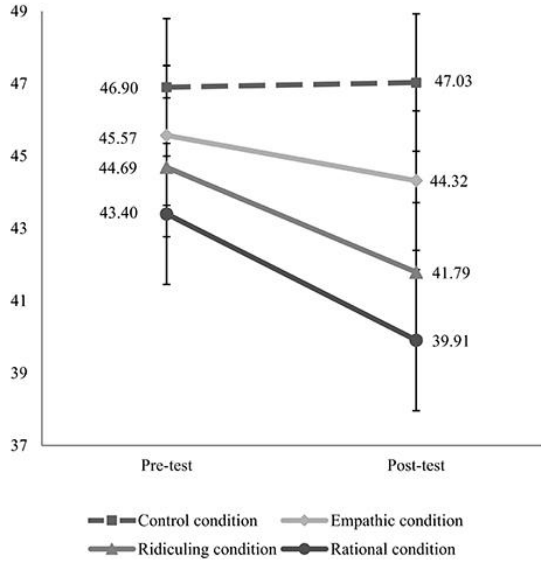
[3] County-by-week is used for our main specification, but we examine all permutations of county/state-by-week/month for robustness.

2021), granting us a large sample size. Second, it covers a large landmass, which is necessary since small countries see less variation in COVID-19 cases across their limited geography. Third, COVID-19 denial or skepticism is pervasive in the United States – as of mid-2020, roughly 25% of Americans believed in a COVID conspiracy theory, and 44% of Republicans believed that Bill Gates was using COVID vaccinations as a pretext to implant microchips (Sanders, 2020).

Literature Review

Our research draws from a broad cross-section of social and data science literature. First, there is a body of literature concerned with the psychology of conspiracy theory belief formation and dissolution. Orosz et al. (2016) examine conspiracy beliefs in a laboratory setting. After recruiting nearly 1,000 Hungarian subjects, the authors play the subjects an audio recording of a "super" conspiracy theory which is wide-ranging in scope (in other words, it linked together multiple areas of society, government, etc.); they subsequently poll subjects on their belief in the theory to get an initial measurement. Next, the authors split the subjects into four groups – three treatment and one control. Each treatment group is assigned one of three interventions – a *rational* counterargument against the conspiracy theory, a *ridiculing* counterargument against those who believe in the theory, and finally an *empathetic* counterargument which asked the subjects to empathize with the target of the conspiracy theory (for example, Jewish Hungarians, who were targeted in the "super" conspiracy theory). After this intervention, Orosz et al. poll the subjects' conspiracy theory beliefs once again and compare changes based on the treatment given. Orosz et al. find that rational and ridiculing counterarguments both significantly decrease the subjects' beliefs in the conspiracy theory, although the rational counterargument is slightly more effective (see: Figure **REPLACE W/ FIG #**).

The empathetic counterargument treatment also appears to decrease belief in the conspiracy theory, although the effect is not statistically significant.



As a controlled experiment in psychology, Orosz et al.'s research has merit, but it has considerable limitations. One concern is the Hawthorne effect (McCarney et al., 2007)– subjects are known to behave differently when they know they are being observed – so these laboratory findings may not accurately translate into real-world knowledge. A related concern is that the conspiracy theory that Orosz et al. played for their subjects might be "weaker" or less natural than a conspiracy theory found in the real world. This conclusion is reached through a simple thought experiment: if we post an abundance of conspiracy theories on Twitter, a few of them will be "stronger" (i.e., more believable or likely to be re-tweeted), and will propagate through social networks at a high rate, while the "weaker" conspiracy theories will spread slowly (if at all). Therefore, the conspiracy theories that we observe in the real world have faced a sort of natural selection in order to reach us, and are likely "stronger" than the average conspiracy theory. As there is no reason to believe that the conspiracy theory which Orosz et al. created was any "stronger" than average, it is probably "weaker" than those found in the real world. It is also possible that the source of conspiracy theories plays an important role in their credibility –recently presented theories (by strangers in a laboratory, no less) may have different characteristics than those acquired through natural interactions. One final concern with Orosz et al.'s research is that their sample was *random*: the selected participants were balanced by gender, age, level of education, and location of residence so as to be representative of Hungary's population. While one usually wants a balanced sample when running a laboratory experiment, in this case we believe that it is detrimental to the applicability of their research. This is because conspiracy theorists are *not* a random sample of the population – they self-select and have specific patterns and traits. Using a balanced sample allows Orosz et al. to answer the question "what will change conspiracy theory beliefs in the general population?" However, our research question, which we believe is more important and interesting, is "what will change conspiracy theory beliefs in the *conspiracy theorist* population".

Beyond this paper, there is an interdisciplinary literature surrounding Bayesian reasoning, crossing the boundaries of behavioural economics, psychology, and political science. The Bayesian framework provides a model for explaining the way that individuals update their beliefs using new information. In brief, this framework proposes that individuals hold prior beliefs about the state of the world, as well as their confidence that this belief is true. In the case of COVID-19, one potential state of the world could be that the virus is an elaborate hoax planned by political and societal elites. In the traditional formulation of the Bayesian framework, individuals update their priors upon encountering new information according to their belief that the information provides a credible signal. In this view, a news release detailing a large increase in case numbers need not significantly alter a conspiracy theorist's belief in a hoax if they assess the source to be uncredible. In recent decades, an alternative view known as "motivated reasoning" has become popular, in which individuals seek out information which supports their beliefs and disregard information contradicting their beliefs, regardless of the supposed "credibility" of the source (Kunda, 1990). Distinguishing between these competing conceptions is a difficult task; Druckman and McGrath (2019) review the available evidence with respect to beliefs about climate change and find it to be approximately equally consistent with each. Nevertheless, there is considerable interest in understanding the ways in which individuals may internalize information in “irrational” ways (Grether, 1980). While our research does not provide any additional evidence to distinguish between these frameworks, we keep each in mind while formulating our model and empirical strategy.

Our research also relates to the field of infodemiology, defined as the science of the distribution and evolution of online information, typically with the aim of understanding and informing public health and policy (Eysenbach, 2002). This field has grown enormously over the past decade as the amount of data generated online has exploded. Infodemiology has been used for a wide range of applications, from gathering information about levels of physical activity and drug usage (Liu et al., 2019; Yom-Tov & Lev-Ran, 2017), to detecting depression and suicidal ideation (Cheng et al., 2017; Ricard et al., 2018). Most relevant to our research, infodemiological studies have improved our understanding of the spread of information about COVID-19 as well as the spread of the disease itself. Mavragani (2020) uses data from Google Trends and the Baidu Index to show that certain search terms related to COVID-19 symptoms have positive correlations with future COVID-19 case counts, suggesting their use as a predictive tool. Several studies (Himelein-Wachowiak et al., 2021; Kouzy et al., 2020) have documented trends in COVID-19 misinformation sharing both through individuals and through automated bots. However, few if any of these studies attempt to find causal relationships, and to our knowledge no study has yet linked these patterns of misinformation with COVID-19 itself.

Finally, we make use of machine learning methodologies drawing from computer science literature. Most critically, we use a version of the BERT language model via HuggingFace’s “transformers” library, which uses state-of-the-art transfer learning algorithms to create accurate but lightweight text classification models requiring relatively little training (in comparison to traditional machine learning models). We discuss these technical details later in the Data section.

Data

Our final dataset is a novel longitudinal aggregation of scraped Tweets, Twitter data, and user metadata together with locales' COVID-19 statistics and demographics.[4] Each row corresponds to a geographical area (contiguous United States county or state) and time (week or month after January 1st, 2019) pairing, and reports the conspiracy theory activity in the area over the paired time period.

Twint

The core of our data comes from Twint – short for "Twitter Intelligence Tool" – a Python package which scrapes the front-end of Twitter's website. While it is possible to collect tweets using Twitter's internal API, the severe rate and quantity limitations make large scale data collection infeasible unless the data is collected in real time and stored locally. Conversely, Twint allows us to collect a large pool of over 5 billion individual tweets and their associated metadata within a relatively short time span.

Our research question is focused on understanding the behaviour of conspiracy theorists, so the first step is to obtain a list of presumed conspiracy theorists. Using algorithms to classify users as conspiracy theorists is a difficult task that inevitably introduces measurement error. Our baseline approach makes use of shared URLs, keywords, and hashtags (short tags used on Twitter to contextualize tweets and to allow them to be seen by a broader audience). We generate a list of conspiracy theorists by using Twint to search for all tweets containing any element from a list of conspiratorial links, keywords, and hashtags. These include several links to popular websites spreading COVID-19 misinformation and conspiracy theories, as well as 16 of the most popular hashtags and 14 of the most popular keywords related to COVID conspiracies (see Appendix for a list of all hashtags, keywords and links used). We use Twint to collect all tweets and retweets (an action

[4] An infographic which shows the full process of creating this dataset can be found in the appendix.

where users repost others’ tweets for their own followers to see) from any user who had tweeted at least one of the items on our list, beginning in January 2019 and ending in July 2021. This process leaves us with a dataframe which includes the text, user ID, and date for each tweet or retweet from each identified conspiracy theorist (as well as other less relevant columns). Next, we filter our list of users to include only those who have used any of the conspiratorial links, keywords, or hashtags at least three times. This filtering minimizes the probability of false positives, since users who believe that COVID-19 is a legitimate pandemic may still tweet these keywords sarcastically or in disbelief – e.g., “People really believe in a #CovidHoax? I had COVID myself!” While false positives remain a concern, our research design is structured in such a way that false positives attenuate our findings to insignificance, rather than producing a spurious significant result.

While Twint is free, fast, and (despite many bugs) a valuable tool, it has its limitations – tragically, it cannot access the “geolocation” value associated with some tweets. It can access users’ bios, where they can choose to fill a text “location” field, but these locations are notoriously inconsistent – some examples of locations are “å–µå–µç¼–è¾‘éƒ¨”, “Estonia”, and “San Francisco, USA”. Even a real, useful user location such as “San Francisco, USA” is difficult to match with a database of locations due to misspellings, abbreviations, and rearrangement-permutations of state, country, county, and region. In addition, users can set this “location” field to any location they wish, so the user responsible for the “San Francisco, USA” example might actually live in France (or be an automated bot). Given these constraints, we decided that the best approach to geolocating users was to cross-reference our list of users with massive online datasets of geolocated users. We found two useful datasets – “GeoCov19” and “US Election 2020 Tweets”**.**

Geo-Data

GeoCov19’s authors (Qazi, Imran, and Ofli, 2020) describe it as “A dataset of hundreds of millions of multilingual COVID-19 tweets with location information” and the metadata alone has a total size of

317 gigabytes.[5] It was collected by searching for all tweets containing certain COVID-19-related keywords or hashtags, such as "United States COVID-19" or "#covid19Canada" between February 1 and April 30, 2020. Each row of the GeoCov19 dataset corresponds to a single tweet, and we use the following five columns in our research: *User_id*, *user_location*, *geo*, *place*, and *Geo_Source*. *User_id* is the Twitter-generated user-ID of the user who posted the tweet from a given row, which we use to match users in the GeoCov19 database to our list of conspiracy theorist users. The *user_location* field denotes the location provided in the user's bio, if it exists. As previously mentioned, this text-based location can often be of low quality, but GeoCov19's methodology improves on this by attempting to resolve the text location in each user's bio to a specific location through reverse geocoding – essentially the same process that occurs when you search for a location in a map application such as Google Maps. This mitigates many of the aforementioned issues of abbreviation and rearrangement.

Beyond *user_location*, there are additional fields for tweet-specific locations. The *geo* and *place* fields contain the geolocation inferred from the tweet; this data can come from one of four sources recorded in *Geo_source*. The most accurate possible source occurs when a user tags their tweet with explicit coordinates. Because this tag must be in close proximity to their current location as given by the device they are tweeting from, users cannot falsify this information except through the use of a virtual private network. The *geo* field in GeoCov19 is only filled if users provide an exact location. A slightly less precise, but still highly valuable location source is provided if the user tags a *place* – for instance, they may tag "Washington, DC" rather than providing explicit GPS coordinates. This *place* field again relies on the user's device's internal location, providing the same barriers to falsification as *geo*. While these two location sources are the most precise and ideal ways to infer location, they are also infrequently used: approximately 0.85% of tweets feature one of these sources of location.

[5] By "metadata", we mean information about a tweet which does not include the tweet itself, such as tweet location, time, date, detected language, user, etc.

For the remaining majority of cases where these are not available, GeoCov19 fills *place* with *user_location* if it is available, or else attempts to impute the location of the tweet based on locations mentioned in the text of the tweet itself. Unfortunately, this final content-based imputation method – which is by far the most common source of location in GeoCov19 – is not reliable enough for use in our research. Consider the following tweet: "The UK has over 65,000 #COVID19 deaths. More than Qatar, Pakistan, and Norway." This tweet would be tagged with four locations (United Kingdom, Qatar, Pakistan, and Norway), none of which refer to the physical location of the person sending the tweet. Worse still, this form of geo-inference relies on toponym detection, resulting in the frequent inclusion of locations based on spurious acronyms – e.g., a tweet containing the phrase "does a**ny**body believe that?" may resolve to "New York" due to the "ny" in "anybody". A cursory sample of tweets with locations tagged in this manner suggests that the methodology produces far more spurious locations than potentially valid ones. Consequently, we omit tweets tagged via tweet content from our analysis.

Our second source of geo-inference is "US Election 2020 Tweets", a collection of tweets about the U.S. election in 2020 (specifically, all tweets with the hashtags "#joebiden" or "#donaldtrump"). Location in this dataset comes only from the *user_location* field from users' bios, as described above.

We process locations from GeoCov19 and "US Election 2020 Tweets" in a similar manner. We first collapse each dataset by its respective Twitter user ID field so that the same user does not have multiple entries. This process outputs a dataset (for each of the two original datasets) where each row corresponds to one unique user ID and holds a *list* of all locations that the user was recorded at. For example, if a Twitter user with ID "12" tweeted from Montana and then later tweeted from Maryland, this collapsed dataset would appear as follows "12": ["Montana", "Maryland"]. Next, we drop all users who do not have an identical inferred location for all tweets. For example, user "12" would be dropped at this point, since "Montana" and "Maryland" are different locations, but if their list of locations was instead ["Montana", "Montana"] they would not be dropped. Finally, we merge the processed GeoCov19 and "US Election Tweets 2020" datasets with our list of potential conspiracy theorists. Excluding all non-geolocated conspiracy theorists reduces our sample size from approximately 1.1 million users to approximately 95,000 users geolocated at the county level and 145,000 geolocated at the state level.

Our approach to geolocating users comes with some caveats. First is general measurement error which is endemic to any approach to geolocating Twitter users – aside from the rare coordinate-geotagged tweet, users' locations will not always be accurate or truthful, and this will introduce inaccuracies in our data. However, provided that users do not systematically falsely report being in one location while being in another (e.g., if many users claim to reside in a populous metropolis when they actually reside in a sparse rural area), this will only introduce *random* measurement error in our dependent variable (time-aggregated COVID-19 conspiracy activity on Twitter in a specific geographical area), which should increase standard errors but not attenuate our results toward zero. As a robustness check for *non*-random measurement error in user-reported locations, we re-run our main specifications on urban counties and then on rural counties.[6] Another caveat is the measurement error introduced by our atemporal approach – GeoCov19 covers January-April 2020, and "US Election 2020 Tweets" covers October-November 2020 – so it is possible that a user's location in either time period does not correspond to their location over the course of the COVID-19 pandemic. Again, we believe that this measurement error is uncorrelated with COVID-19 cases or conspiracy theory activity, since it is unlikely that users systematically misreport their location based on the time-period they are tweeting in.

[6] Counties are classified on a scale from 1-6, where 1 is the most urban and 6 is the most rural. For this robustness check, we group all counties with classifications between 1-3 as being urban, and those with classifications between 4-6 as being rural.

After merging our Twint and geolocation databases, we are able to track conspiratorial tweets over time for each of our geo-located conspiracy theorists. Our base estimation uses the number of conspiracy tweets as a share of the user's total tweets, aggregated to a weekly or monthly period. In this way, we can observe whether conspiratorial tweet frequency drops off in the aftermath of a sudden spike in COVID cases or deaths. However, this hashtag and link based methodology of identifying conspiracy beliefs has a considerable drawback. Our list is intentionally short, so as provide a focused and high-quality indication of COVID-19 conspiracy beliefs, but this also means that we are potentially missing a significant number of conspiratorial tweets which do not feature one of our chosen identifiers. To address this, we supplement our hashtag and link methodology with a machine learning methodology described in the next section.

BERT

BERT (Bidirectional Encoder Representations from Transformers) is a Google-developed neural network for natural language processing (Devlin et al., 2019). BERT is a flexible model which can be used for question answering, document summarization, and sequence categorization, which is the purpose we use it for. BERT is a transfer-learning algorithm, so it works by first generating a model of natural language through pre-training on a massive corpus of English books and articles, allowing it to recognize patterns and relationships between sentences, words, and sub-word tokens (i.e., prefixes and suffixes). This learning is then "transferred" to specific tasks, where BERT applies its understanding to more specific input text in order to generate a desired output. For example, Google uses the algorithm to identify sections of webpages which are likely to answer a question asked of their search engine, allowing them to present users with answers directly instead of requiring users to click through search results themselves. This transfer-learning enables us to train a purpose built model capable of distinguishing between conspiratorial and non-conspiratorial tweets in several orders of magnitude less time than it would take to do from scratch. Even still, machine learning is a highly time- and computation-intensive task, so we choose to use a lightweight implementation of BERT known as DistilBERT, created by the company HuggingFace. This implementation uses a process called distillation to reduce the number of parameters in the model by 40%, improving speeds by 60% while retaining 97% of the model's accuracy (Sanh et al., 2020).

In order to complete the fine-tuning training, it is necessary to compile labeled training and validation data. Our training dataset is made up of two kinds of tweets – COVID-19 conspiracy tweets, and non-conspiracy tweets, in roughly equal proportions. For the former, we use the set of tweets labeled as COVID-19-conspiratorial using our hashtag and link methodology. We compile non-conspiracy theory tweets by using Twint to search Twitter for all tweets with certain non-conspiracy keywords and/or hashtags.[8] This set of keywords and hashtags is constructed to provide a balanced sample of non-specific tweets combined with tweets about COVID-19 that are clearly not conspiratorial. The latter category is necessary because every tweet in the training set labeled as conspiratorial is related to COVID-19; failure to include enough COVID-19 related tweets that are *not* conspiratorial is likely to generate a model which associates any mention of COVID-19 with conspiracy. Before finalizing the set of non-conspiracy tweets, we cross reference each of their authors with our list of conspiracy users – if a user appears in both datasets (conspiracy tweets and non-conspiracy tweets), we remove their tweets from our sample to avoid mislabeling. We then combine the COVID-19-related conspiracy tweets and non-conspiracy tweets and clean them by keeping only English-language tweets (auto-detected by Twitter), removing all conspiracy or non-conspiracy hashtags and keywords we searched for, removing the "#" character from all remaining hashtags (ones we did not search for), removing all links, removing all mentions of other users, and only keeping tweets with at least three words. After compiling the DistilBERT training data, we split it to reserve 15% for validation.

[8] See: Appendix for a list of these non-conspiracy theory hashtags and keywords.

We train DistilBERT in three epochs on our collection of approximately 2 million labeled training tweets.[10] The resulting model was able to achieve an accuracy rate of 92.7% in the validation dataset, but it is important to recognize that this is only with respect to the labeled examples provided to the algorithm, which are themselves based off hashtags and keywords. This accuracy rate is comparable to benchmark examples of DistilBERT classification tasks (Sanh et al., 2020).

County & COVID-19 Data

Recall that our final dataset is created by merging Twitter conspiracy theory activity with COVID-19 statistics at US county and state levels. So far, we have only discussed the Twitter conspiracy theory data – now we explain how our county-level COVID-19 and demographics data are compiled. We first retrieve COVID-19 cumulative[11] cases and death counts (by county) from the New York Times (The New York Times, 2020/2021) and convert it to daily new cases and deaths by county.

We next clean and process this daily COVID-19 data by removing outliers – 504 (16.1%) of week-aggregated counties have COVID-19 spikes where one time period's mean daily case count is over 4 standard deviations[12] higher or lower than the previous week's,[13] and 4 (0.13%) counties have these erratic spikes when aggregated at the month level. We entirely remove these poorly recorded counties from our week-aggregated and month-aggregated datasets because leaving them in would introduce significant measurement error in our independent variable.

[10] Our training sample is large enough (approximately 0.9 million non-conspiracy tweets and 1-2 million conspiracy tweets – this represents every conspiracy tweet from approximately 20% of our conspiracy theorists) that the class-imbalance problem is not a major concern – there are enough observations in the "minority" class for BERT to pick up on the patterns which separate the two classes (Provost, 2000). However, we have so many more potential conspiracy theory tweets than potential non-conspiracy theory that we use a random sample of the potential conspiracy theory tweets in our BERT training dataset to keep the proportions within one order of magnitude.

[11] The New York Times daily case and death data only cover the first half of the pandemic, while their cumulative statistics are updated daily.

[12] Here, we calculate a standard deviation for each individual county by week, rather than one overall standard deviation for all counties by week.

[13] This usually occurs when reported cases are suddenly updated to reflect many weeks or months of unreported cases.

One important quirk of the New York Times COVID-19 data (and many COVID-19 datasets) is that some entries have negative cases and/or death counts to correct for previous overcounts. Not only are these negative tallies incorrect, they are indicative of inaccuracies in daily case and death counts in the prior weeks and months, and we have no way of discerning exactly which weeks or months are being corrected for (or the distribution of cases and deaths over those time periods). Faced with the options of discarding counties entirely or clipping values at zero, we chose a hybrid approach**.** We discard any counties which ever record a weekly case counts below -150 per 100,000 population or a death rate below -25 per 100,000, and then clip any remaining negative observations at zero. This minimizes the impact of very large and observable adjustments, while leaving a sufficiently large sample of counties to work with.

We then merge these county-level COVID-19 statistics with two other county-level datasets – one (the Census Bureau's "County Population Totals: 2010-2019" dataset) provides us with population estimates (2019), the other (the National Center for Housing Statistics' "Urban-Rural Classification Scheme for Counties", from 2013) classifies each county on an Urban-Rural spectrum. Finally, we collapse these daily county-level COVID-19 statistics and characteristics into weekly (and monthly) statistics.

Model

Our specification does not strictly follow the results of a theoretical model, but in this section we sketch a rough model to illustrate the intuition of our hypothesized effect.

Let the latent variable $\pi^*_{i,c,t}$ be the probability that individual *i* in county *c* at time *t* believes that COVID-19 is dangerous. We can model $\pi^*_{i,c,t}$ as a Bernoulli random variable which equals 1 if and only if person *i* contracts a serious case of COVID-19 or if one or more of person *i*'s friends or loved ones contracts a serious case.[14] [15]

$$\text{Let } \pi^{*}_{i,c,t} = Pr(any\ acquaintance\ or\ loved\ one\ has\ covid)$$

$$= 1 - Pr(no\ acquaintance\ or\ loved\ one\ has\ covid)$$

Assuming, for simplicity, that the probability of contracting COVID-19 in county *c* at time *t* is *i.i.d.* for all individuals, we can write

$$Pr(individual\ \mathrm{j}\ in\ \mathrm{c}\ at\ \mathrm{t}\ is\ \mathrm{not}\ infected\ with\ covid) = \frac{non{-}infected_{c,t}}{total\ pop_{c,t}}$$

Then, the probability that none of individual *i*'s *j*-many acquaintances or loved ones contract COVID-19 at time *t* in county *c* can is given by

$$\mathrm{Pr}(acquain.1\ not\ infected) * \mathrm{Pr}(acquain.2\ not\ infected)$$

$$* \cdots \mathrm{Pr}(acquain.2\ not\ infected)$$

$$= \prod_{j=1}^{n_i} \left(\frac{non - infected_{c,t}}{total\ pop_{c,t}} \right)$$

$$=^{i.i.d.} \left(\frac{non - infected_{c,t}}{total\ pop_{c,t}} \right)^{n_i}$$

Where n_i = *number of person i's acquaintances/loved ones + 1.*[16]

This allows us to re-write

$$\pi^{*}_{i,c,t} = 1 - (\frac{non - infected_{c,t}}{total\ pop_{c,t}})^{n_i}$$

[14] For brevity, we will refer to the group consisting of individual *i*, their acquaintances, and their loved ones as their acquaintances.

[15] Note that since this is just to illustrate intuition, any model where $\pi^{*}_{i,c,t}$ is distributed with support $x \in N^{+}$ and a monotonically decreasing probability density function would serve the same purpose as our model. For example, we could also model $\pi^{*}_{i,c,t}$ as a random variable where individual *i* becomes more likely to believe that COVID-19 is dangerous as more people they know contract the disease (rather than believing it is dangerous after knowing one person who contracts a serious case).

[16] The "+1" comes from accounting for individual *i*, since it is also plausible that they will update their COVID-19 priors after contracting the disease themselves.

Re-writing, we see that

$$\pi^*_{i,c,t} = 1 - (1 - \frac{infected_{c,t}}{total\ pop_{c,t}})^{n_i}$$

$$= 1 - (1 - COV_{c,t})^{n_i}$$

Where $COV_{c,t} = \frac{infected_{c,t}}{total\ pop_{c,t}}$ is the variable from our final specification.[17]

We can use this framework to model individual i's beliefs as COVID-19 case numbers vary in their community over time:

As the entire community contracts COVID-19, we see that the probability that individual *i* believes that COVID-19 is dangerous approaches 1:

$$\lim_{COV_{c,t} \to 1^-} \pi^*_{i,c,t} \; = 1 - (1 - 1)^{n_i} = 1$$

Conversely, as community COVID-19 cases dwindle, we see that the probability that individual *i* believes that COVID-19 is dangerous approaches 0.

$$\lim_{COV_{c,t} \to 0^+} \pi^*_{i,c,t} \; = 1 - (1 - 0)^{n_i} = 0$$

Clearly, in this model, $\pi^*_{i,c,t}$ (the probability that individual *i* believes that COVID-19 is dangerous) is increasing in COVID-19 cases.

$\pi^*_{i,c,t}$ is not observable to us – it would take laboratories, surveys, and massive amounts of resources to directly estimate this latent variable from individuals (and even then, it may not be accurate[18]). The closest thing we have to $\pi^*_{i,c,t}$ is COVID-19-related conspiracy theory activity on Twitter – defined as $CT_{i,c,t}$. Specifically, we can simply model the two variables as perfectly inversely correlated:

$$CT_{i,c,t} = 1 - \pi^*_{i,c,t}$$

[17] Note that this model of $\pi^*_{i,c,t}$ as a function of $COV_{c,t}$ is equivalent to modeling $\pi^*_{i,c,t}$ as the cumulative distribution of a geometric random variable with probability parameter $COV_{c,t}$ and n_i trials.

[18] These inaccuracies would result from the same issues affecting Orosz et al.'s work.

In plain English, this equation means that as an individual begins to believe that COVID-19 is dangerous, they become less likely to post COVID-19-related conspiracy theory content on Twitter. While there are personal idiosyncrasies, differences, and entrenched beliefs, we believe that this is a reasonable way to model the average individual's relationship between tweeting and believing.

Empirical Specification

We have two main regression specifications, each intended to capture a slightly different effect. Here is Specification (1):

Specification 1: Cases (lags)

$$\ln\left(CT_{c,t}\right) - \ln\left(CT_{c,t-1}\right) = \alpha_0 + \sum_{j=0}^{k}\left(\beta_{-j}\ln\left(COV_{c,t-j}\right)\right) + \phi_c + \epsilon_{c,t}$$

Specification 1: Deaths (leads)

$$\ln\left(CT_{c,t}\right) - \ln\left(CT_{c,t-1}\right) = \alpha_0 + \sum_{j=0}^{k}\left(\beta_{j}\ln\left(COV_{c,t+j}\right)\right) + \phi_c + \epsilon_{c,t}$$

where the subscript *c* denotes county *c*, and *t* denotes time-period *t*.[19]

$CT_{c,t}$ is an aggregated measure of COVID-19-related conspiracy theory activity from all individuals *i* in county c. Specifically, $CT_{c,t} = \sum_{i\in c} CT_{i,c,t}$, where $CT_{i,c,t}$ is a measure of COVID-19-related conspiracy theory activity from individual *i*.[20] $CT_{c,t-1}$ is the same aggregated measure of COVID-19-related conspiracy theory activity, but from the previous time period. Thus, the left-hand side of specification equals the weekly or monthly percentage change in COVID-19 conspiracy activity

[19] In some robustness checks we aggregate users to State (rather than county) level and also to month (rather than week) time-periods.

[20] Specifically, $CT_{i,c,t}$ is either the number of COVID-19 related conspiracy theory tweets in time-period t, or the percentage of tweets from individual i in time-period t that are COVID-19 related conspiracy theories. When aggregating the percentage of tweets, we first tally the number of COVID-19 conspiracy tweets and non-conspiracy tweets, and only then compute the percentage (simply aggregating the percentages would mis-weight them).

$COV_{c,t}$ has two definitions, depending on the specification. As mentioned in the "Model" section, one definition $COV_{c,t} = \frac{infected_{c,t}}{total\ pop_{c,t}}$ is defined as the number of new *cases* per capita in the community during time period t. The other definition $COV_{c,t} = \frac{deaths_{c,t}}{total\ pop_{c,t}}$ is defined as the number of new deaths per capita in the community during time period t. We use both definitions of $COV_{c,t}$ to the same end – understanding how seriously COVID-19 afflicts a community during time period t.

Finally, ϕ_c is a county-level fixed effect. While it may seem strange to demean a left-hand side which is already time-differenced, we believe that including ϕ_c is necessary since it changes the interpretation of the left-hand side from "weekly change in log COVID-19-related conspiracy theory activity" to "weekly change in log COVID-19-related conspiracy theory activity, *relative to this county's average level of weekly change.*" This is a crucial distinction because COVID-19-related conspiracy theory activity decreased over the course of the pandemic, possibly irrespective of recent local COVID-19 cases and deaths (See Figure 1 in Results). With ϕ_c's inclusion, we now hope to measure the local changes in log COVID-19-related conspiracy theory activity which come from *recent* changes in local COVID-19 cases and deaths, rather than changes in log COVID-19-related conspiracy theory activity which come from *long-term* trends within a county.

The first empirical specification examines week-to-week changes in conspiracy theorist activity but does not tell us anything about the long-term persistence of any conspiracy theory reduction. This is where Specification (2) comes in – it is identical to Specification (1), except the outcome variable is slightly tweaked:

Specification 2: Cases (lags)

$$\ln\left(CT\_Perm_{c,t}\right) - \ln\left(CT\_Perm_{c,t}\right) = \alpha_0 + \sum_{j=0}^{k}\left(\beta_{-j}\ln\left(COV_{c,t-j}\right)\right) + \phi_c + \epsilon_{c,t}$$

Specification 2: Deaths (leads)

$$\ln\left(CT\_Perm_{c,t}\right) - \ln\left(CT\_Perm_{c,t}\right) = \alpha_0 + \sum_{j=0}^{k}\left(\beta_{j}\ln\left(COV_{c,t+j}\right)\right) + \phi_c + \epsilon_{c,t}$$

The new variable, $CT\_Perm_{c,t}$ represents the share of users in county *c* who cease tweeting COVID-19-related conspiracy theory content, starting in time-period *t*. We are careful with $CT\_Perm_{c,t}$, since many accounts (especially conspiracy theorists (BBC, 2021)) are likely to be banned over the course of the pandemic – to avoid false positives here, we only consider accounts who stop tweeting conspiracy theories but continue posting other content. An additional complication can be understood by imagining a scenario of two identical users (1 and 2) who each cease tweeting conspiracy theories for two months, except User 1 stops at the beginning of our observational period and User 2 stops toward the end. If we define $CT\_Perm_{c,t}$ as the share of users who *permanently* cease tweeting conspiracy theories, User 2 will be falsely labeled as meeting this criterion. To avoid this issue, we define $CT\_Perm_{c,t}$ as the share of users who *semi*-permanently cease tweeting conspiracy theories – specifically, those who do not post COVID-19 conspiracy theories for the next three months, starting in period *t*.[21] We also drop all observations from the last three months of our observational period, since they would be falsely labeled.

Interpretation

β_j is the parameter of interest in our specifications, and we include the lagged *j* case terms in both case-specifications because new COVID-19 cases usually take longer than one week to develop into a serious illness – the median elapsed time between *hospital-diagnosed* COVID-19 and dying is 18.5

[21] We also examine the share of users who stop tweeting COVID-19 conspiracy theories for the next 6 months, as a robustness check.

days (Fei Zhou et al., 2020)[22] – and we only expect conspiracy theorists to update their beliefs after the illness progresses to a serious stage.[23] We include lead terms in the death-specifications for the same reason – if an individual dies of COVID-19 in week *j*, it is likely that they were seriously ill in week *j+1*, and possibly also in weeks *j+2*, etc.

In Specification (1), β_j is the *week-to-week* percentage change in conspiracy theory activity due to an increase of local COVID-19 intensity *j* weeks ago, and in Specification (2), β_j is the *long-term* percentage change in conspiracy theory activity due to an increase of local COVID-19 intensity *j* weeks ago. For example, if we ran Specification (1) with new cases, β_2 would be the change (from last week to this week) in users who tweeted COVID-19 conspiracy theories, due to new cases per capita two weeks ago. We expect β_j to have a negative coefficient in Specification (1) – as we outline in our model, we expect to see a decrease in conspiracy theory activity as cases or deaths increase. Conversely, we expect β_j to have a positive coefficient in Specification (2) because we expect more users to cease tweeting conspiracy theories as cases or deaths increase.

One crucial detail of interpreting our research design is that we first identify a static list of potential conspiracy theorists, and only then aggregate their behaviour to a location. If we simply aggregated the behaviour of all users in a location, we would end up measuring the average conspiracy theory belief of the *entire population* of geo-locatable twitter users in that location. While this is still an interesting metric, it would not tell us whether the specific conspiracy theorist population has changed their beliefs at all – for example, it is entirely possible that the general population uses

[22] This study is from the extreme early stages of the pandemic (before Jan 31, 2020), when tests were slower and rarely available for the public. Since non-hospital testing was rare, and cases take some time to progress to a hospital-worthy stage, we believe that this 18.5-day estimate is an underestimate of the median time between contracting COVID-19 and dying. Harrison et al. (2020) also provide an estimate of the time between symptom onset and death, and conclude that it ranges from one to two weeks.

[23] Note that we do not include lagged COVID-19 *deaths* terms, since, unlike cases, we expect deaths to have immediate impact on family and loved ones.

more (or less) COVID-19 conspiracy related hashtags after a change in local case numbers (perhaps reflecting curiosity, rather than beliefs) but conspiracy theorists do not change their hashtag usage. Keeping a static sample is essential for our data gathering, since the imprecision of our conspiracy theory classification means that we can only compare conspiracy theory activity within a specific set of users.

Assumptions & Concerns

Our foremost concern is exogeneity – specifically, our research design requires the causation to run from COVID-19 cases or deaths to conspiracy theory activity and beliefs, and not the other way around. The obvious objection to our design is that COVID-19 conspiracy theorists are more likely to contract COVID-19 since they do not take defensive measures. We agree with this statement but do not believe that it is a valid objection to our research design due to its specific structure, since the timing of COVID-19 shocks is still nearly exogenous at the individual level, and we aggregate our data in such a way as to preserve that exogeneity.[24]

We believe that the timing of COVID-19 shocks is exogenous at the individual level because it is strongly dependent on the number of cases surrounding an individual at that time (i.e., in nearby geographies), which are not determined by the individual's beliefs regarding COVID-19. Put another way, while the conspiracy theorist's actions will partially determine the cases in their area, their actions will not determine the timing of those local cases and spikes.

[24] For example, when aggregating users in a county who stopped tweeting conspiracy theories but continued tweeting other content, we first assess whether each individual user stopped tweeting, and only then aggregate to the county average of users who stopped tweeting.

Results

First, we check our data to see whether our conspiracy theory identification methods appear to give reasonable results. Figure 1 shows the share of aggregate tweets that are identified as being conspiratorial by our hashtag and link methodology and by our DistilBERT methodology. The first thing to note is that the frequency of conspiracy tweets (as detected by hashtag, link, and keyword) appears to peak prior to the peak of COVID-19 cases and deaths, and it then slowly dwindles. There are multiple reasonable explanations for this – for one, depending on the average number of people in an individual’s circle of close family and friends, it is possible that a large proportion of American knew someone with COVID-19 well before cases peaked. It may also be the case that the hashtags, keywords, and links we use to identify conspiracy tweets were most popular during the early days of the pandemic, before losing favour to other, potentially still-conspiratorial tags. This is a possibility which could be explored in future research by utilizing a dynamic set of identifiers, but that is outside the scope of this paper. Finally, it is possible that people simply grew less interested in COVID-19 over time as they grew accustomed to it, and therefore tweeted about it less frequently.

Figure 1

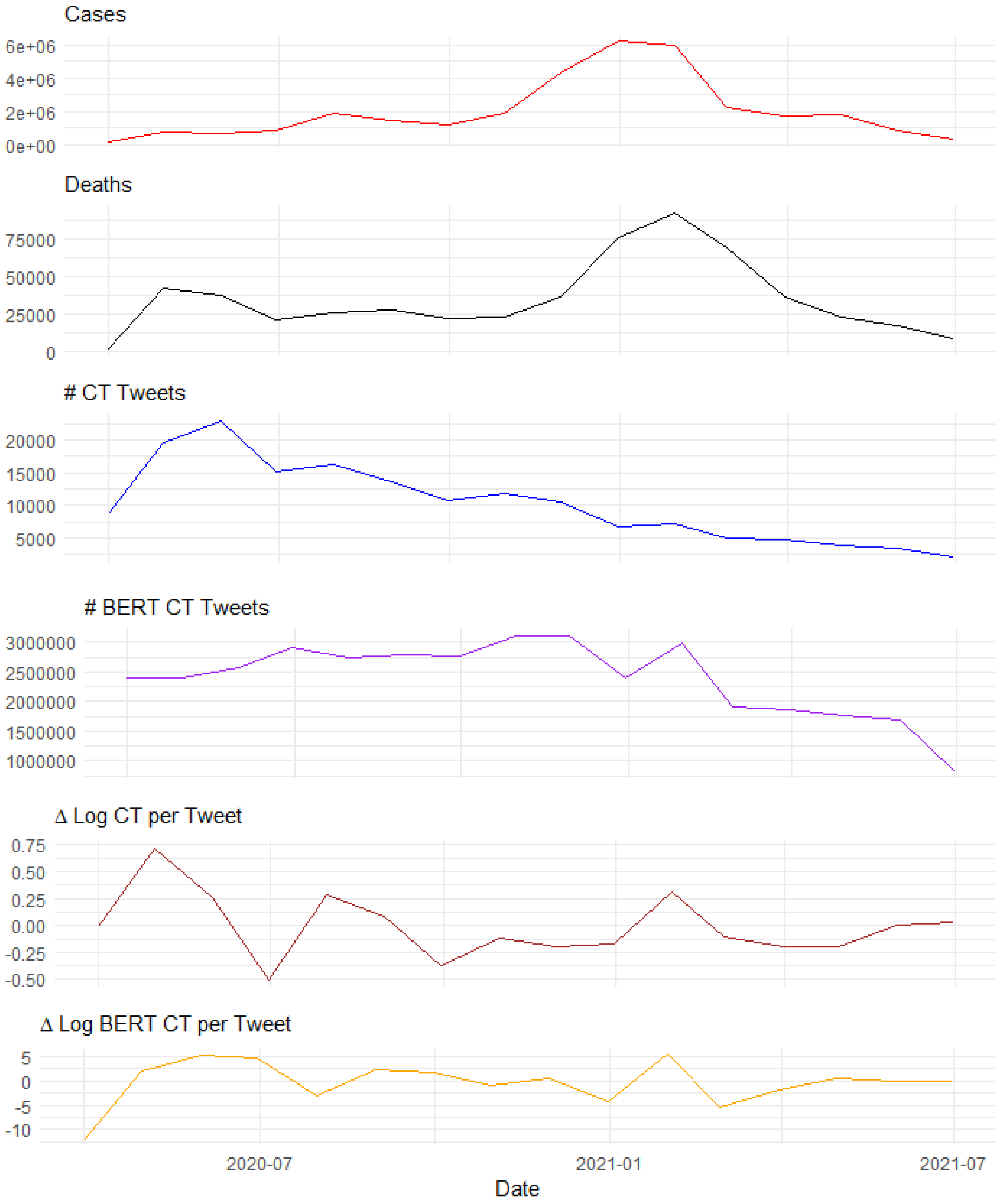


Table 1 shows the results of our first specification for county-by-week data – our most granular specification. These results may constitute evidence that an increase in COVID-19 cases causes a decrease in conspiracy behaviour on Twitter – a 100% increase in the number of COVID-19 cases is associated with a 4.5% decrease in the number of conspiracy theory tweets 4 weeks later, as measured by our hashtag and link methodology. Our DistilBERT specifications (3) and (4) find a larger effect: The same doubling in cases produces a 14.2% to 14.4% decrease in conspiracy activity after 3 weeks. Interestingly, these specifications also find weak evidence for an apparent dip in conspiracy theory behaviour associated with a 1-week lead of COVID-19 cases, which is followed by a comparable increase caused by current COVID-19 cases.

Table 1: Specification 1 (County-by-Week)

Dependent Variables:	Diffed Log CT		Diffed Log BERT CT		Diffed Log CT		Diffed Log BERT CT	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Intercept	0.00015		0.00338**		0.00010		0.00163**	
	(0.00070)		(0.00147)		(0.00032)		(0.00069)	
Log cases per 100k (4 wk lag)	-0.00045*	-0.00045*	-0.00009	-0.00010				
	(0.00024)	(0.00024)	(0.00058)	(0.00057)				
Log cases per 100k (3 wk lag)	0.00046	0.00046	-0.00142*	-0.00144*				
	(0.00030)	(0.00035)	(0.00073)	(0.00083)				
Log cases per 100k (2 wk lag)	-0.00020	-0.00017	0.00083	0.00079				
	(0.00030)	(0.00033)	(0.00075)	(0.00085)				
Log cases per 100k (1 wk lag)	-0.00006	-0.00008	-0.00033	-0.00034				
	(0.00031)	(0.00035)	(0.00078)	(0.00087)				
Log cases per 100k	0.00042	0.00040	0.00138*	0.00138				
	(0.00031)	(0.00036)	(0.00078)	(0.00090)				
Log cases per 100k (1 wk lead)	-0.00021	-0.00019	-0.00112*	-0.00113*				
	(0.00027)	(0.00031)	(0.00064)	(0.00065)				
Log deaths per 100k (1 wk lag)					-0.00035	-0.00037	-0.00049	-0.00058
					(0.00029)	(0.00029)	(0.00068)	(0.00064)
Log deaths per 100k					0.00046	0.00049	0.00014	0.00014
					(0.00030)	(0.00031)	(0.00070)	(0.00076)
Log deaths per 100k (1 wk lead)					-0.00009	-0.00007	0.00026	0.00024
					(0.00030)	(0.00035)	(0.00073)	(0.00076)
Log deaths per 100k (2 wk lead)					-0.00020	-0.00021	-0.00022	-0.00025
					(0.00031)	(0.00038)	(0.00072)	(0.00073)
Log deaths per 100k (3 wk lead)					0.00069**	0.00065**	0.00081	0.00073
					(0.00028)	(0.00032)	(0.00069)	(0.00083)
Log deaths per 100k (4 wk lead)					-0.00059**	-0.00061**	-0.00184***	-0.00190***
					(0.00028)	(0.00028)	(0.00065)	(0.00067)
Fixed-effects								
County		Yes		Yes		Yes		Yes
Fit statistics								
Observations	43,355	43,355	43,355	43,355	43,941	43,941	43,941	43,941
R^2	0.00016	0.00816	0.00035	0.00607	0.00028	0.00831	0.00029	0.00486
Within R^2		0.00015		0.00036		0.00029		0.00031

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

The second feature of note in Figure 1 is that DistilBERT identifies a far greater share of tweets as being conspiratorial than our link and hashtag methodology – nearly 23,594%. While this seems unlikely, it does not necessarily invalidate our DistilBERT-results; to begin with, our hashtag, keyword, and link-labeling criteria are excessively strict to avoid false-positives and mislabeling, so DistilBERT should be labeling far more tweets as conspiracy theories. Ideally, we would fine-tune the algorithm to create a model more in line with our original, narrow conception, but given time and resource constraints we choose to use the model as is. It should be stressed, however, that these numbers still reflect real changes in tweet behaviour identified by the model.

More significantly, Table 1 appears to show modest-to-strong evidence of a decrease in conspiracy theory Twitter activity caused by the 4-week *lead* of COVID-19 deaths – a decrease of around 6% for a doubling of deaths as measured by our hashtag and link methodology, and around 18-19% for our DistilBERT methodology. Though this could potentially be related to reverse causality, we believe this is more likely to be due to the lagged nature of death statistics. As previously mentioned, the median time between the onset of symptoms and death is approximately 2 weeks (Harrison, 2020), but this period is both longer and higher in variance for individuals under 70, who make up the large majority of Twitter users.[26] We also run similar specifications with additional lags and leads. Leads of more than 1 week are insignificant for cases, and lags of more than 1 week are insignificant for deaths, as expected. Specifications with additional lags for cases or leads for deaths appears to create overfitting issues, so we omit them.

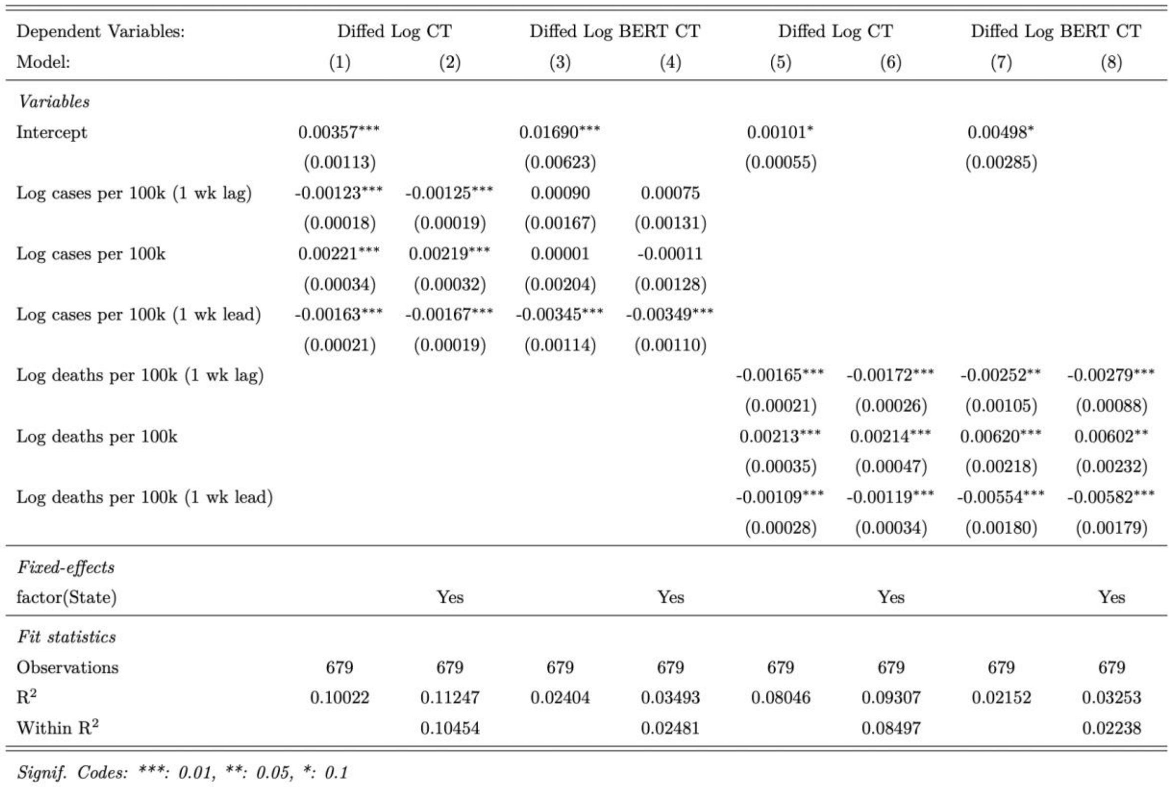
While Table 1 gives us a view of our data at the most granular level, it is also instructive to look at more aggregated regressions. Using these coarser-grained data mitigates measurement error over

[26] In addition, this is an estimate of the median time between confirmed onset of symptoms and death, so it is an underestimate of the time between contracting COVID-19 and death.

both location and time since, as previously mentioned, a large proportion of users' locations are imputed from the stated location in their bio, which is subject to considerable measurement error. Given that we expect that users may choose to locate themselves in the nearest city rather than their exact location, it is likely that location measurement error is much smaller when data is aggregated to the state level. Similarly, issues of misreporting and subsequent revisions over time may be partially mitigated by aggregating to monthly data. These aggregations come at the cost of forgoing local variation between counties and from week-to-week, which can be significant. Table 2 shows the results of our first specification using state-by-month data. Interestingly, these regressions tell a different story – increases in cases and deaths from both the preceding and following months appear to decrease Twitter conspiracy behaviour, relative to the previous month, while cases and deaths in the current month appear to *increase* the number of conspiracy tweets, relative to the previous month. One potential explanation for this pattern is that conspiratorial tweets are lower in the month preceding a spike in COVID due to lower levels of interest. In the month during which COVID numbers spike, it may be the case that conspiracy theorists spend some time coming to terms with the new information, potentially tweeting justifications which are captured as conspiratorial. Finally, in the month following a spike, the information becomes difficult to ignore, and conspiratorial tweets become less frequent.

While this narrative provides one potential explanation, the true underlying explanation is far from clear. Potential future research could clarify the causal pathway at play by creating several categorizations of conspiratorial tweets to separate mental states at a finer grain. For instance, a list of keywords related to common justifications for COVID-19 cases and deaths might include mentions of falsified data and co-morbidities. This could help to distinguish a "justification" stage from more general-purpose conspiracy tweets regarding the veracity or origins of the virus.

Table 2: Specification 1 (State-by-Month)

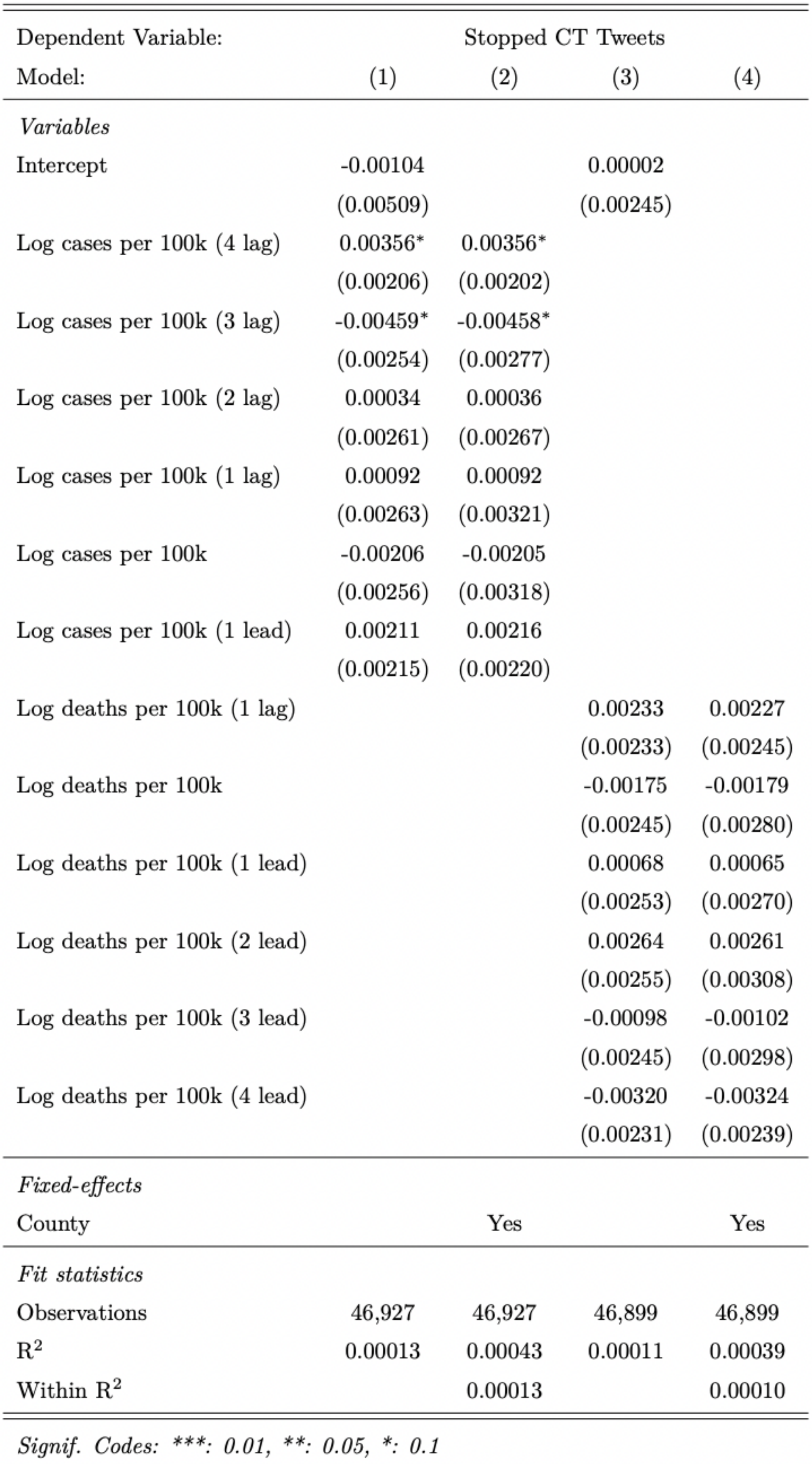


Dependent Variables:	Diffed Log CT		Diffed Log BERT CT		Diffed Log CT		Diffed Log BERT CT	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Intercept	0.00357***		0.01690***		0.00101*		0.00498*	
	(0.00113)		(0.00623)		(0.00055)		(0.00285)	
Log cases per 100k (1 wk lag)	-0.00123***	-0.00125***	0.00090	0.00075				
	(0.00018)	(0.00019)	(0.00167)	(0.00131)				
Log cases per 100k	0.00221***	0.00219***	0.00001	-0.00011				
	(0.00034)	(0.00032)	(0.00204)	(0.00128)				
Log cases per 100k (1 wk lead)	-0.00163***	-0.00167***	-0.00345***	-0.00349***				
	(0.00021)	(0.00019)	(0.00114)	(0.00110)				
Log deaths per 100k (1 wk lag)					-0.00165***	-0.00172***	-0.00252**	-0.00279***
					(0.00021)	(0.00026)	(0.00105)	(0.00088)
Log deaths per 100k					0.00213***	0.00214***	0.00620***	0.00602**
					(0.00035)	(0.00047)	(0.00218)	(0.00232)
Log deaths per 100k (1 wk lead)					-0.00109***	-0.00119***	-0.00554***	-0.00582***
					(0.00028)	(0.00034)	(0.00180)	(0.00179)
Fixed-effects								
factor(State)		Yes		Yes		Yes		Yes
Fit statistics								
Observations	679	679	679	679	679	679	679	679
R^2	0.10022	0.11247	0.02404	0.03493	0.08046	0.09307	0.02152	0.03253
Within R^2		0.10454		0.02481		0.08497		0.02238

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

We provide appendix tables containing the results of similar regressions which use county-by-month data (Table A.1), state-by-week data (Table A.2), as well as a set of tables showing results for our first specification when data split is split between rural and urban counties (Tables A.3 – A.4). Of these latter tables, it is interesting to note that the relationships shown in Tables 1 and 2 are replicated or even stronger in rural counties, but are mostly insignificant in urban counties. There may be several reasons for this: first, it is much more common for individuals to tag themselves in large cities, whether on vacation or because it is more convenient to say you are from a nearby city than to explain the location of your rural town. This likely leads to a larger rate of location mismeasurement in urban counties than in rural ones, where geotagged individuals are more likely to be true residents. Additionally, automated bots may tend to place themselves disproportionally in larger cities, again increasing measurement in urban counties. Both phenomena would have the effect of biasing results towards zero.

Table 3: Specification 2 (County-by-Week)



Dependent Variable:	Stopped CT Tweets			
Model:	(1)	(2)	(3)	(4)
Variables				
Intercept	-0.00104		0.00002	
	(0.00509)		(0.00245)	
Log cases per 100k (4 lag)	0.00356*	0.00356*		
	(0.00206)	(0.00202)		
Log cases per 100k (3 lag)	-0.00459*	-0.00458*		
	(0.00254)	(0.00277)		
Log cases per 100k (2 lag)	0.00034	0.00036		
	(0.00261)	(0.00267)		
Log cases per 100k (1 lag)	0.00092	0.00092		
	(0.00263)	(0.00321)		
Log cases per 100k	-0.00206	-0.00205		
	(0.00256)	(0.00318)		
Log cases per 100k (1 lead)	0.00211	0.00216		
	(0.00215)	(0.00220)		
Log deaths per 100k (1 lag)			0.00233	0.00227
			(0.00233)	(0.00245)
Log deaths per 100k			-0.00175	-0.00179
			(0.00245)	(0.00280)
Log deaths per 100k (1 lead)			0.00068	0.00065
			(0.00253)	(0.00270)
Log deaths per 100k (2 lead)			0.00264	0.00261
			(0.00255)	(0.00308)
Log deaths per 100k (3 lead)			-0.00098	-0.00102
			(0.00245)	(0.00298)
Log deaths per 100k (4 lead)			-0.00320	-0.00324
			(0.00231)	(0.00239)
Fixed-effects				
County		Yes		Yes
Fit statistics				
Observations	46,927	46,927	46,899	46,899
R^2	0.00013	0.00043	0.00011	0.00039
Within R^2		0.00013		0.00010

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 3 shows the results of our second specification using county-by-week data. While there is some weak significance on the 3rd and 4th lags of COVID-19 case numbers, there appears to be no strong patterns of significant coefficients when comparing across tables using other permutations of county or state by weekly or monthly data (see Tables A.5 – A.7 in the appendix). This suggests that while our first specification shows some evidence of patterns of decreased conspiratorial Twitter activity surrounding increases in COVID-19 cases and deaths, these changes may not be permanent. Note here that we omit BERT data from our second specification estimations, as the high frequency of conspiracy theory categorizations leave very few individuals who are deemed to not tweeted a single conspiratorial tweet for an extended period.

Conclusion

Our regressions show a wide range of results which do not lend themselves to a clear-cut conclusion. Our county-by-week specification (1) in Table 1, which we believe to be the most reliable due to its high granularity, constitutes weak evidence that increased COVID-19 cases or deaths cause a decrease in conspiracy theory activity on Twitter. On the other hand, our county-by-week specification (2) in Table 3 appears to suggest that reductions in conspiracy theory belief are at best temporary, as increases in cases or deaths do not lead to an increase in the number of individuals who permanently stop their conspiratorial activity.

One possible interpretation of our many non-significant coefficients is that conspiracy theorists are “stubborn” and do not change their minds when they see the disease’s impact first-hand. However, this conclusion relies on accurate data and minimal measurement error, and due to the imprecision of our data sources, we believe there to be significant measurement error in our data. This attenuates our results to insignificance and make the “stubborn” conclusion premature and potentially incorrect.

Despite our ambiguous results, we believe that our research design is valid, and that this research question can be answered with improved data (and a slightly improved methodology). First, we need to reduce measurement error in conspiracy theory tweet classifications – we could accomplish this by spending the requisite time fine-tuning DistilBERT into a more accurate and sensitive classification algorithm; by providing DistilBERT with hand-labeled training and validation data that is 100% accurate; and by finding better methods for mechanically labeling tweets (e.g., coming up with a better list of hashtags, links, and keywords, or splitting them into more detailed categories). Another important area of improvement is our geolocation. Our methodology of cross-referencing identified conspiracy theorists with massive databases of geolocated users yielded more matches than we expected (over 10%), but this methodology was only used because Twitter denied us access to their Academic API, which would have allowed us to geolocate far more users. The Academic API would also grant us access to much more of a user’s location data over time, so we could more reliably deduce their locations.

Third, there is an emerging field of adversarial machine learning dedicated to detecting automated twitter accounts (“bots”) and automatically generated text. By applying these methods, we could reduce the share of bots in our sample of conspiracy theory users, leading to less measurement error.

Finally, it may be possible to identify additional conspiracy theorists by analyzing the networks of individuals they follow and interact with. One approach could be to search for tightly connected users in the vicinity of popular COVID-19 conspiracy accounts using graph clustering algorithms. This approach would strengthen our sample size as well as its quality, by ensuring that identified conspiracy theorists are part of a conspiracy theorist community rather than a bot with no legitimate followers, or a non-conspiracy theorist who occasionally tweets one of our keywords in jest or protest.

Ultimately, the jury is still out on whether COVID-19 conspiracy theorists update their prior beliefs. However, we believe (and hope) that with the appropriate data, tools, and resources, this important question can be answered.

References

Bureau, U. C. (n.d.). *County Population Totals: 2010-2019*. The United States Census Bureau. Retrieved July 31, 2021, from https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html

Cheng, Q., Li, T. M., Kwok, C.-L., Zhu, T., & Yip, P. S. (2017). Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study. *Journal of Medical Internet Research*, *19*(7), e7276. https://doi.org/10.2196/jmir.7276

CO-CIN. (2020). CO-CIN: COVID-19 - Time from symptom onset until death in UK hospitalised patients, 7 October 2020. *GOV.UK*. https://www.gov.uk/government/publications/co-cin-covid-19-time-from-symptom-onset-until-death-in-uk-hospitalised-patients-7-october-2020

Data Access—Urban Rural Classification Scheme for Counties. (2019, December 2). https://www.cdc.gov/nchs/data_access/urban_rural.htm

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423

Druckman, J. N., Link to external site, this link will open in a new window, McGrath, M. C., & Link to external site, this link will open in a new window. (2019). The evidence for motivated reasoning in climate change preference formation. *Nature Climate Change*, *9*(2), 111–119. http://dx.doi.org/10.1038/s41558-018-0360-1

Eysenbach, G. (2002). Infodemiology: The epidemiology of (mis)information. *The American Journal of Medicine*, *113*(9), 763–765. https://doi.org/10.1016/S0002-9343(02)01473-0

Eysenbach, G. (2009). Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *Journal of Medical Internet Research*, *11*(1), e1157. https://doi.org/10.2196/jmir.1157

Grether, D. M. (1980). Bayes Rule as a Descriptive Model: The Representativeness Heuristic*. *The Quarterly Journal of Economics*, *95*(3), 537–557. https://doi.org/10.2307/1885092

Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A., Epstein, D. H., Leggio, L., & Curtis, B. (2021). Bots and Misinformation Spread on Social Media: Implications for COVID-19. *Journal of Medical Internet Research*, *23*(5), e26933. https://doi.org/10.2196/26933

Imran, M. (2020). *GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information* [Data set]. IEEE. https://ieee-dataport.org/open-access/geocov19-dataset-hundreds-millions-multilingual-covid-19-tweets-location-information

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., & Baddour, K. (2020). Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus*, *12*(3), e7255. https://doi.org/10.7759/cureus.7255

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Liu, S., Chen, B., & Kuo, A. (2019). Monitoring Physical Activity Levels Using Twitter Data: Infodemiology Study. *Journal of Medical Internet Research*, *21*(6), e12394. https://doi.org/10.2196/12394

Many COVID-19 patients insist 'it's not real' until they die, nurse says—National | Globalnews.ca. (n.d.). Global News. Retrieved July 31, 2021, from https://globalnews.ca/news/7467283/coronavirus-denier-deaths-nurse-hoax/

Mavragani, A. (2020). Tracking COVID-19 in Europe: Infodemiology Approach. *JMIR Public Health and Surveillance*, *6*(2), e18941. https://doi.org/10.2196/18941

McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne Effect: A randomised, controlled trial. *BMC Medical Research Methodology*, *7*, 30. https://doi.org/10.1186/1471-2288-7-30

NW, 1615 L. St, Washington, S. 800, & Inquiries, D. 20036 U.-419-4300 | M.-857-8562 | F.-419-4372 | M. (n.d.). A look at the Americans who believe there is some truth to the conspiracy theory that COVID-19 was planned. *Pew Research Center*. Retrieved July 31, 2021, from https://www.pewresearch.org/fact-tank/2020/07/24/a-look-at-the-americans-who-believe-there-is-some-truth-to-the-conspiracy-theory-that-covid-19-was-planned/

Orosz, G., Krekó, P., Paskuj, B., Tóth-Király, I., Bőthe, B., & Roland-Lévy, C. (2016). Changing Conspiracy Beliefs through Rationality and Ridiculing. *Frontiers in Psychology*, *0*. https://doi.org/10.3389/fpsyg.2016.01525

Provost, F. (2000). *Machine Learning from Imbalanced Data Sets 101*. 3.

Ricard, B. J., Marsch, L. A., Crosier, B., & Hassanpour, S. (2018). Exploring the Utility of Community-Generated Social Media Content for Detecting Depression: An Analytical

Study on Instagram. *Journal of Medical Internet Research*, *20*(12), e11817. https://doi.org/10.2196/11817

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv:1910.01108 [Cs]*. http://arxiv.org/abs/1910.01108

Srikanth, A. (2020, July 13). *"I thought this was a hoax": A COVID-party guest learned coronavirus was real on their deathbed* [Text]. TheHill. https://thehill.com/changing-america/well-being/prevention-cures/507046-i-thought-this-was-a-hoax-a-covid-party-guest

The difference between what Republicans and Democrats believe to be true about COVID-19 | YouGov. (n.d.). Retrieved July 31, 2021, from https://today.yougov.com/topics/politics/articles-reports/2020/05/26/republicans-democrats-misinformation

The New York Times. (2021). *Coronavirus (Covid-19) Data in the United States*. The New York Times. https://github.com/nytimes/covid-19-data (Original work published 2020)

Twitter: Most users by country. (n.d.). Statista. Retrieved July 28, 2021, from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Twitter suspends 70,000 accounts linked to QAnon. (2021, January 12). *BBC News*. https://www.bbc.com/news/technology-55638558

US Election 2020 Tweets. (n.d.). Retrieved July 31, 2021, from https://kaggle.com/manchunhui/us-election-2020-tweets

Yom-Tov, E., & Lev-Ran, S. (2017). Adverse Reactions Associated With Cannabis Consumption as Evident From Search Engine Queries. *JMIR Public Health and Surveillance*, *3*(4), e8391. https://doi.org/10.2196/publichealth.8391

Appendix

General Conspiracy Hashtags:

- Plandemic
- Scamdemic
- Covidhoax
- Nwo
- Covid1984
- Plandemia
- Agenda21
- Thegreatreset
- Agenda2030
- Openamericanow
- Firefauci
- Wwg1wga
- Qanon
- Coronahoax

General Conspiracy Keywords

- Plandemic
- Scamdemic
- Covidhoax
- Covid hoax
- Covid1984
- Plandemia
- New world order
- Wake up america
- Open america now
- Fire fauci
- Wwg1wga
- Qanon
- Coronahoax
- Corona hoax

General Conspiracy Links[27]

- Zerohedge.com
- Infowars.com
- Principia-scientific.com
- Tx.voice-truth.com
- Humansarefree.com
- Activistpost.com
- Gnews.org

[27] Source: https://tweets.covid19misinfo.org/#

- Wakingtimes.com
- Brighteon.com
- Thewallwillfall.org
- Sott.net

COVID-19-Specific Hashtags

- Plandemic
- Scamdemic
- Covidhoax
- Covid1984
- Plandemia
- Firefauci
- Coronahoax

COVID-19-Specific Keywords

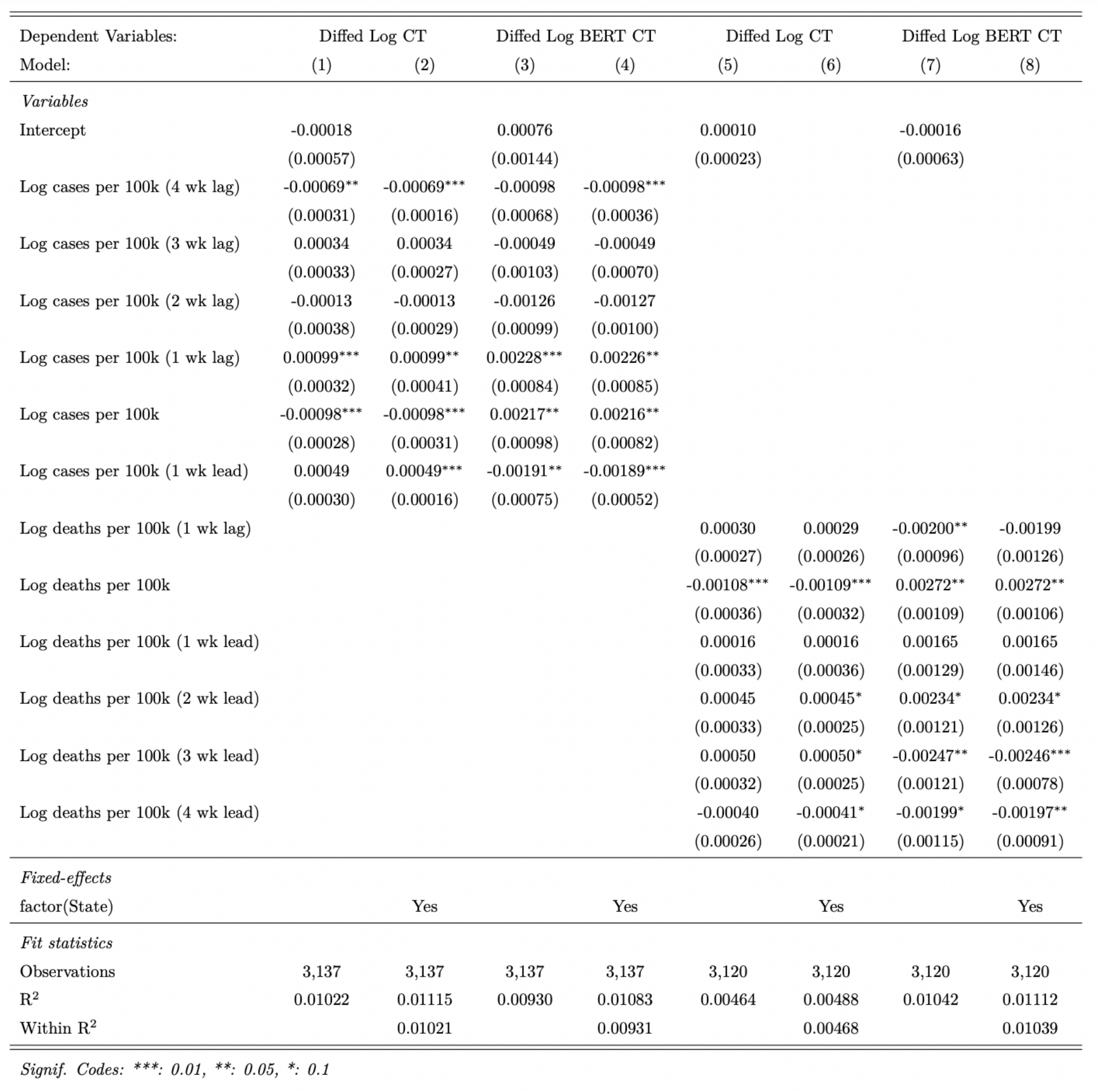
- Plandemic
- Scamdemic
- Covidhoax
- Covid hoax
- Covid1984
- Plandemia
- Fire Fauci
- Coronahoax
- Corona hoax

Table A.1: Specification 1 (County-by-Month)

Dependent Variables:	Diffed Log CT		Diffed Log BERT CT		Diffed Log CT		Diffed Log BERT CT	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Intercept	0.00516***		0.01971***		0.00065		0.00640***	
	(0.00120)		(0.00239)		(0.00057)		(0.00115)	
Log cases per 100k (1 lag)	-0.00040	-0.00047*	-0.00086*	-0.00093*				
	(0.00025)	(0.00025)	(0.00051)	(0.00050)				
Log cases per 100k	0.00020	0.00015	-0.00203***	-0.00216***				
	(0.00034)	(0.00036)	(0.00073)	(0.00071)				
Log cases per 100k (1 lead)	-0.00076***	-0.00080***	-0.00024	-0.00041				
	(0.00023)	(0.00021)	(0.00052)	(0.00050)				
Log deaths per 100k (1 lag)					-0.00084***	-0.00096***	-0.00108	-0.00139*
					(0.00024)	(0.00026)	(0.00066)	(0.00082)
Log deaths per 100k					0.00075**	0.00065*	-0.00079	-0.00082
					(0.00030)	(0.00035)	(0.00085)	(0.00109)
Log deaths per 100k (1 lead)					-0.00051**	-0.00059**	-0.00078	-0.00123*
					(0.00026)	(0.00028)	(0.00064)	(0.00073)
Fixed-effects								
County		Yes		Yes		Yes		Yes
Fit statistics								
Observations	11,811	11,811	11,811	11,811	11,811	11,811	11,811	11,811
R^2	0.00219	0.01616	0.00448	0.02325	0.00154	0.01546	0.00191	0.02074
Within R^2		0.00256		0.00482		0.00186		0.00227

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

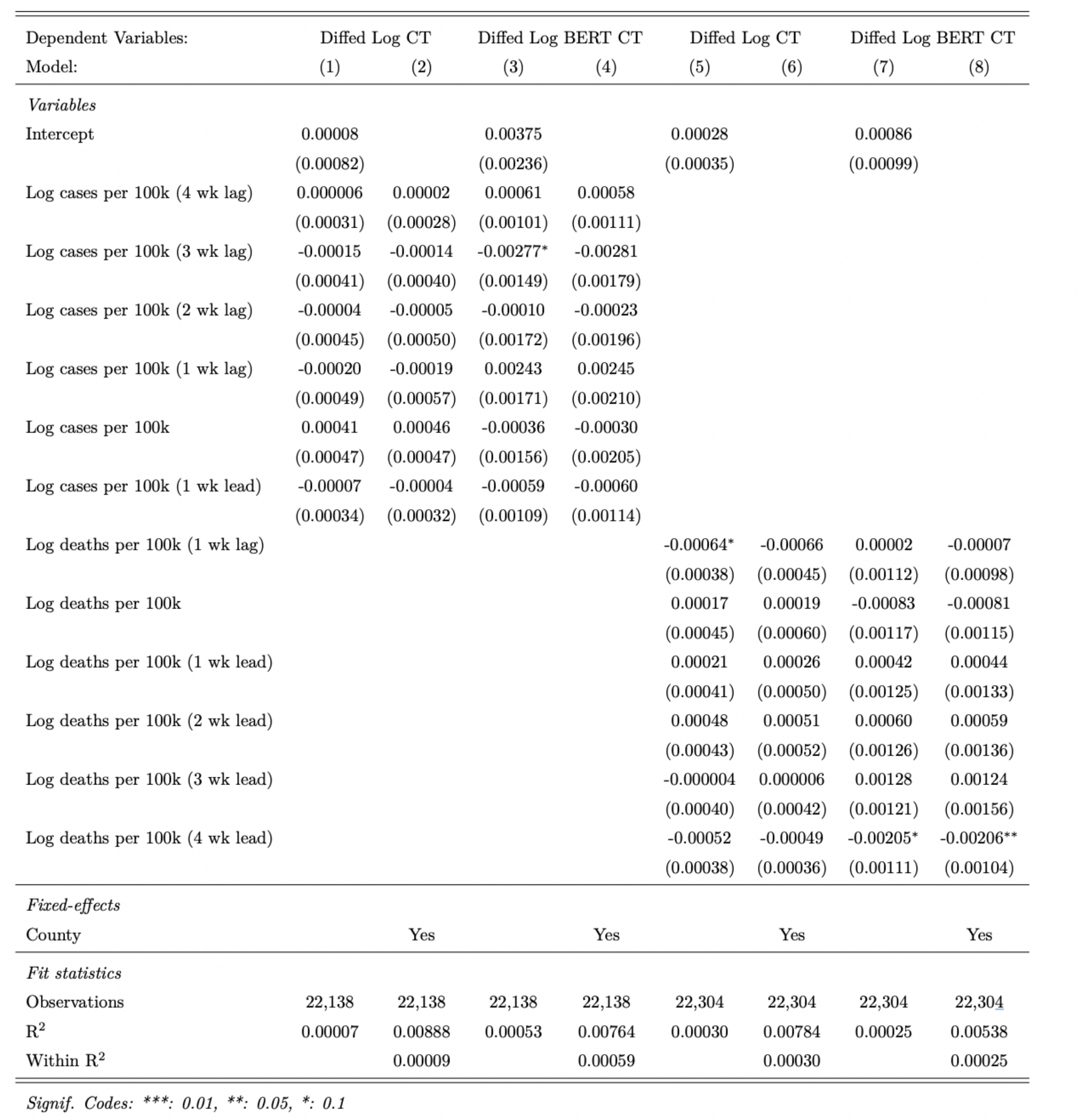
Table A.2: Specification 1 (State-by-Week)



Dependent Variables:	Diffed Log CT		Diffed Log BERT CT		Diffed Log CT		Diffed Log BERT CT	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Intercept	-0.00018		0.00076		0.00010		-0.00016	
	(0.00057)		(0.00144)		(0.00023)		(0.00063)	
Log cases per 100k (4 wk lag)	-0.00069**	-0.00069***	-0.00098	-0.00098***				
	(0.00031)	(0.00016)	(0.00068)	(0.00036)				
Log cases per 100k (3 wk lag)	0.00034	0.00034	-0.00049	-0.00049				
	(0.00033)	(0.00027)	(0.00103)	(0.00070)				
Log cases per 100k (2 wk lag)	-0.00013	-0.00013	-0.00126	-0.00127				
	(0.00038)	(0.00029)	(0.00099)	(0.00100)				
Log cases per 100k (1 wk lag)	0.00099***	0.00099**	0.00228***	0.00226**				
	(0.00032)	(0.00041)	(0.00084)	(0.00085)				
Log cases per 100k	-0.00098***	-0.00098***	0.00217**	0.00216**				
	(0.00028)	(0.00031)	(0.00098)	(0.00082)				
Log cases per 100k (1 wk lead)	0.00049	0.00049***	-0.00191**	-0.00189***				
	(0.00030)	(0.00016)	(0.00075)	(0.00052)				
Log deaths per 100k (1 wk lag)					0.00030	0.00029	-0.00200**	-0.00199
					(0.00027)	(0.00026)	(0.00096)	(0.00126)
Log deaths per 100k					-0.00108***	-0.00109***	0.00272**	0.00272**
					(0.00036)	(0.00032)	(0.00109)	(0.00106)
Log deaths per 100k (1 wk lead)					0.00016	0.00016	0.00165	0.00165
					(0.00033)	(0.00036)	(0.00129)	(0.00146)
Log deaths per 100k (2 wk lead)					0.00045	0.00045*	0.00234*	0.00234*
					(0.00033)	(0.00025)	(0.00121)	(0.00126)
Log deaths per 100k (3 wk lead)					0.00050	0.00050*	-0.00247**	-0.00246***
					(0.00032)	(0.00025)	(0.00121)	(0.00078)
Log deaths per 100k (4 wk lead)					-0.00040	-0.00041*	-0.00199*	-0.00197**
					(0.00026)	(0.00021)	(0.00115)	(0.00091)
Fixed-effects								
factor(State)		Yes		Yes		Yes		Yes
Fit statistics								
Observations	3,137	3,137	3,137	3,137	3,120	3,120	3,120	3,120
R^2	0.01022	0.01115	0.00930	0.01083	0.00464	0.00488	0.01042	0.01112
Within R^2		0.01021		0.00931		0.00468		0.01039

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table A.3: Specification 1, Urban only (County-by-Week)



Dependent Variables:	Diffed Log CT		Diffed Log BERT CT		Diffed Log CT		Diffed Log BERT CT	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Intercept	0.00008		0.00375		0.00028		0.00086	
	(0.00082)		(0.00236)		(0.00035)		(0.00099)	
Log cases per 100k (4 wk lag)	0.000006	0.00002	0.00061	0.00058				
	(0.00031)	(0.00028)	(0.00101)	(0.00111)				
Log cases per 100k (3 wk lag)	-0.00015	-0.00014	-0.00277*	-0.00281				
	(0.00041)	(0.00040)	(0.00149)	(0.00179)				
Log cases per 100k (2 wk lag)	-0.00004	-0.00005	-0.00010	-0.00023				
	(0.00045)	(0.00050)	(0.00172)	(0.00196)				
Log cases per 100k (1 wk lag)	-0.00020	-0.00019	0.00243	0.00245				
	(0.00049)	(0.00057)	(0.00171)	(0.00210)				
Log cases per 100k	0.00041	0.00046	-0.00036	-0.00030				
	(0.00047)	(0.00047)	(0.00156)	(0.00205)				
Log cases per 100k (1 wk lead)	-0.00007	-0.00004	-0.00059	-0.00060				
	(0.00034)	(0.00032)	(0.00109)	(0.00114)				
Log deaths per 100k (1 wk lag)					-0.00064*	-0.00066	0.00002	-0.00007
					(0.00038)	(0.00045)	(0.00112)	(0.00098)
Log deaths per 100k					0.00017	0.00019	-0.00083	-0.00081
					(0.00045)	(0.00060)	(0.00117)	(0.00115)
Log deaths per 100k (1 wk lead)					0.00021	0.00026	0.00042	0.00044
					(0.00041)	(0.00050)	(0.00125)	(0.00133)
Log deaths per 100k (2 wk lead)					0.00048	0.00051	0.00060	0.00059
					(0.00043)	(0.00052)	(0.00126)	(0.00136)
Log deaths per 100k (3 wk lead)					-0.000004	0.000006	0.00128	0.00124
					(0.00040)	(0.00042)	(0.00121)	(0.00156)
Log deaths per 100k (4 wk lead)					-0.00052	-0.00049	-0.00205*	-0.00206**
					(0.00038)	(0.00036)	(0.00111)	(0.00104)
Fixed-effects								
County		Yes		Yes		Yes		Yes
Fit statistics								
Observations	22,138	22,138	22,138	22,138	22,304	22,304	22,304	22,304
R^2	0.00007	0.00888	0.00053	0.00764	0.00030	0.00784	0.00025	0.00538
Within R^2		0.00009		0.00059		0.00030		0.00025

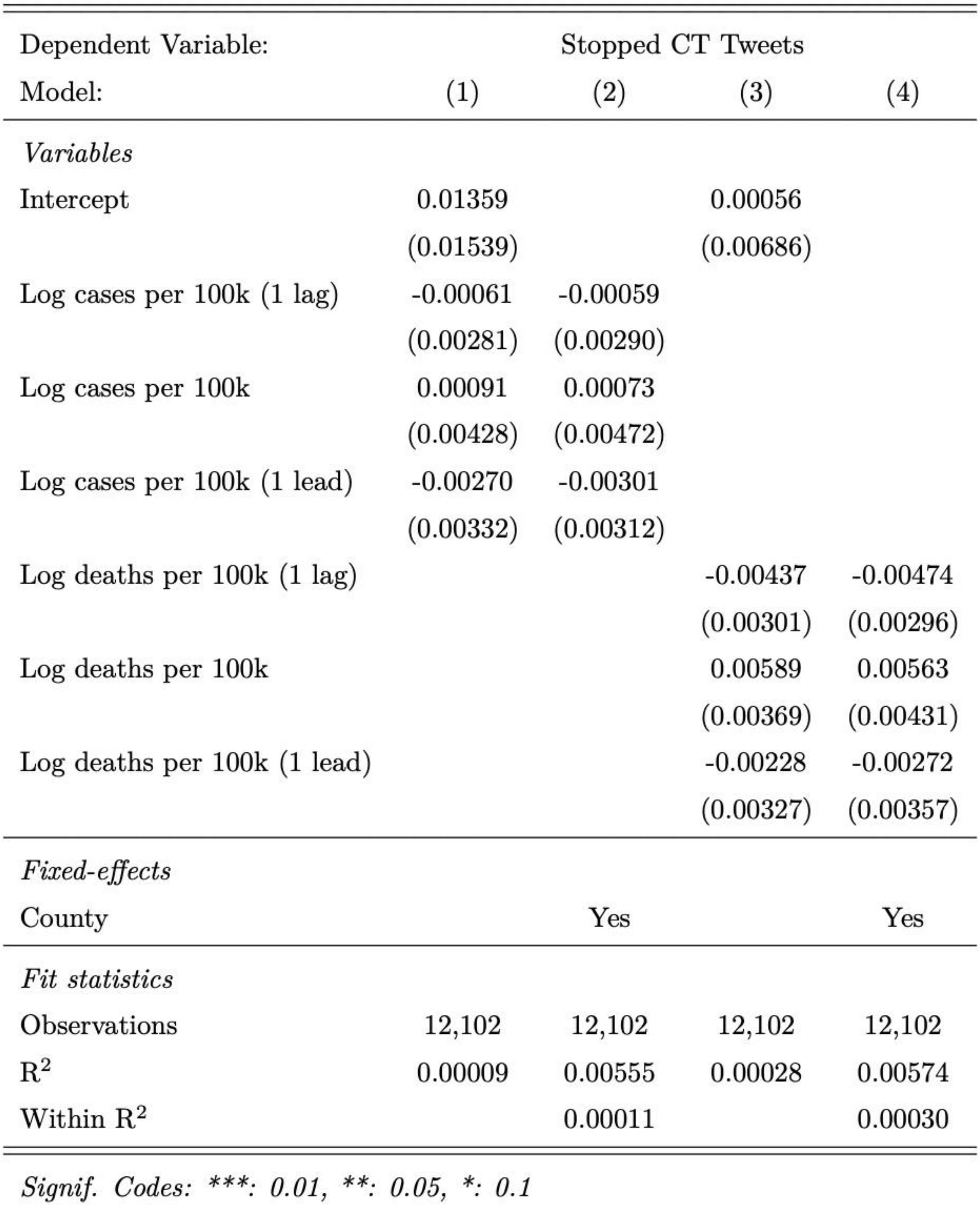
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table A.4: Specification 1, Rural only (County-by-Week)

Dependent Variables:	Diffed Log CT		Diffed Log BERT CT		Diffed Log CT		Diffed Log BERT CT	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Intercept	0.00019		0.00318*		1.65×10^{-7}		0.00209**	
	(0.00099)		(0.00187)		(0.00049)		(0.00095)	
Log cases per 100k (4 lag)	-0.00061*	-0.00061*	-0.00023	-0.00021				
	(0.00032)	(0.00032)	(0.00073)	(0.00068)				
Log cases per 100k (3 lag)	0.00064*	0.00064	-0.00100	-0.00099				
	(0.00037)	(0.00045)	(0.00084)	(0.00094)				
Log cases per 100k (2 lag)	-0.00023	-0.00019	0.00113	0.00113				
	(0.00038)	(0.00041)	(0.00082)	(0.00093)				
Log cases per 100k (1 lag)	-0.000005	-0.00003	-0.00119	-0.00121				
	(0.00039)	(0.00042)	(0.00088)	(0.00093)				
Log cases per 100k	0.00042	0.00038	0.00188**	0.00185*				
	(0.00039)	(0.00045)	(0.00090)	(0.00099)				
Log cases per 100k (1 lead)	-0.00026	-0.00025	-0.00132*	-0.00134*				
	(0.00035)	(0.00041)	(0.00078)	(0.00078)				
Log deaths per 100k (1 lag)					-0.00021	-0.00022	-0.00069	-0.00079
					(0.00039)	(0.00036)	(0.00086)	(0.00081)
Log deaths per 100k					0.00058	0.00061	0.00050	0.00047
					(0.00038)	(0.00037)	(0.00087)	(0.00096)
Log deaths per 100k (1 lead)					-0.00019	-0.00018	0.00015	0.00010
					(0.00038)	(0.00044)	(0.00090)	(0.00092)
Log deaths per 100k (2 lead)					-0.00046	-0.00050	-0.00058	-0.00064
					(0.00039)	(0.00049)	(0.00088)	(0.00087)
Log deaths per 100k (3 lead)					0.00095***	0.00089**	0.00055	0.00045
					(0.00036)	(0.00041)	(0.00084)	(0.00098)
Log deaths per 100k (4 lead)					-0.00062*	-0.00068*	-0.00182**	-0.00192**
					(0.00037)	(0.00038)	(0.00081)	(0.00084)
Fixed-effects								
County		Yes		Yes		Yes		Yes
Fit statistics								
Observations	21,217	21,217	21,217	21,217	21,637	21,637	21,637	21,637
R^2	0.00025	0.00796	0.00050	0.00534	0.00043	0.00863	0.00042	0.00465
Within R^2		0.00024		0.00049		0.00044		0.00046

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

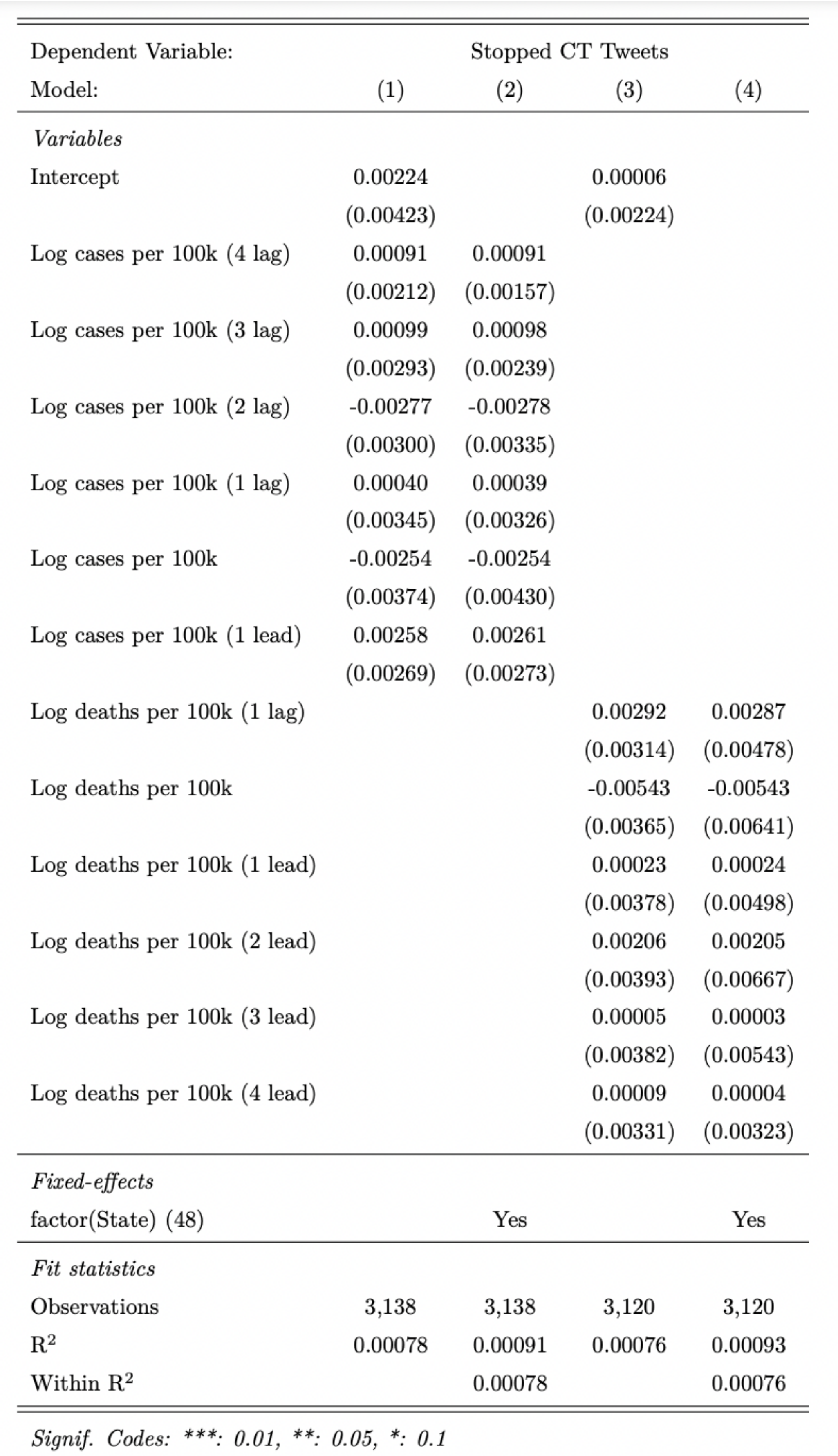
Table A.5: Specification 2 (County-by-Month)



Dependent Variable:	Stopped CT Tweets			
Model:	(1)	(2)	(3)	(4)
Variables				
Intercept	0.01359		0.00056	
	(0.01539)		(0.00686)	
Log cases per 100k (1 lag)	-0.00061	-0.00059		
	(0.00281)	(0.00290)		
Log cases per 100k	0.00091	0.00073		
	(0.00428)	(0.00472)		
Log cases per 100k (1 lead)	-0.00270	-0.00301		
	(0.00332)	(0.00312)		
Log deaths per 100k (1 lag)			-0.00437	-0.00474
			(0.00301)	(0.00296)
Log deaths per 100k			0.00589	0.00563
			(0.00369)	(0.00431)
Log deaths per 100k (1 lead)			-0.00228	-0.00272
			(0.00327)	(0.00357)
Fixed-effects				
County		Yes		Yes
Fit statistics				
Observations	12,102	12,102	12,102	12,102
R^2	0.00009	0.00555	0.00028	0.00574
Within R^2		0.00011		0.00030

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

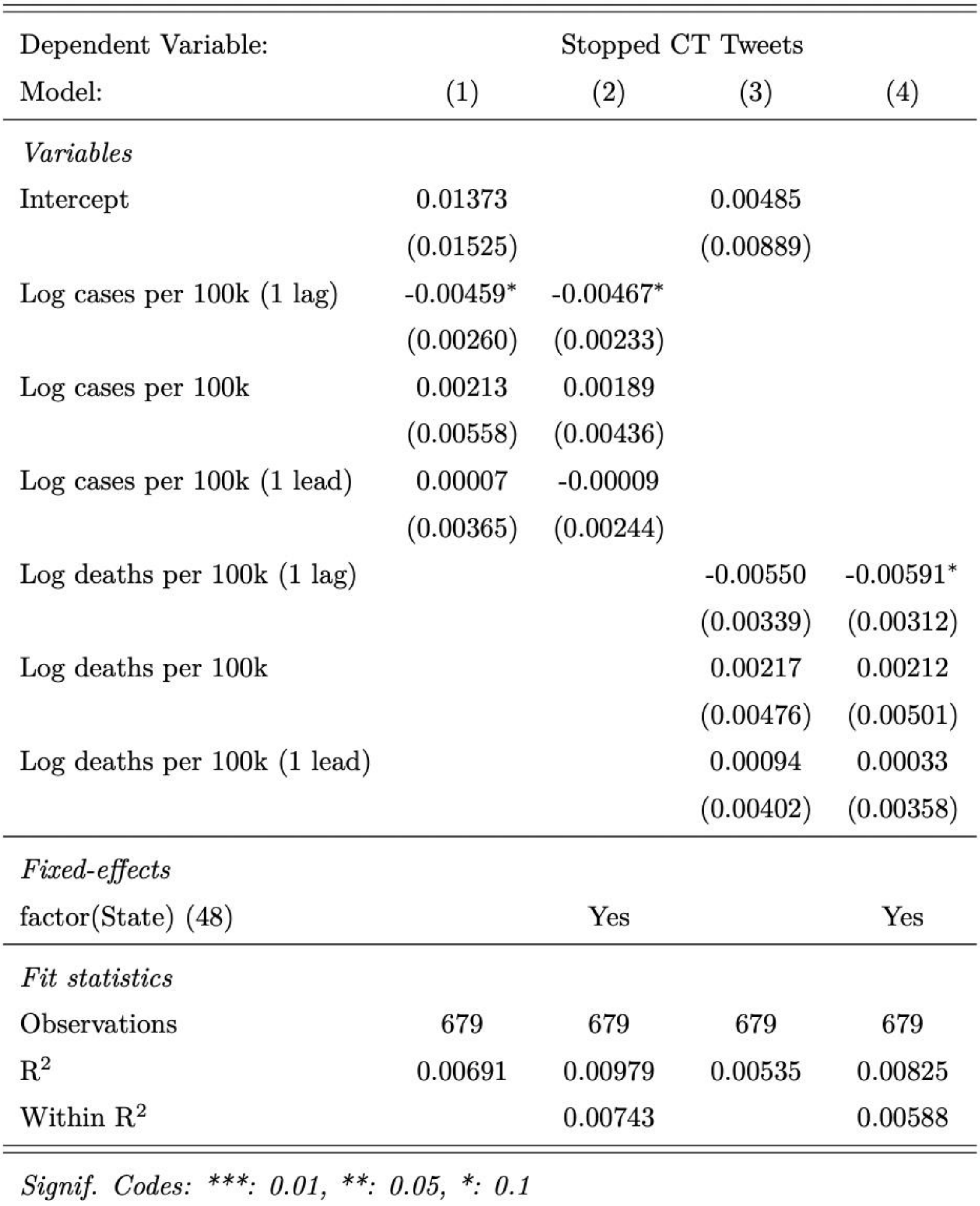
Table A.6: Specification 2 (State-by-Week)



Dependent Variable:	Stopped CT Tweets			
Model:	(1)	(2)	(3)	(4)
Variables				
Intercept	0.00224		0.00006	
	(0.00423)		(0.00224)	
Log cases per 100k (4 lag)	0.00091	0.00091		
	(0.00212)	(0.00157)		
Log cases per 100k (3 lag)	0.00099	0.00098		
	(0.00293)	(0.00239)		
Log cases per 100k (2 lag)	-0.00277	-0.00278		
	(0.00300)	(0.00335)		
Log cases per 100k (1 lag)	0.00040	0.00039		
	(0.00345)	(0.00326)		
Log cases per 100k	-0.00254	-0.00254		
	(0.00374)	(0.00430)		
Log cases per 100k (1 lead)	0.00258	0.00261		
	(0.00269)	(0.00273)		
Log deaths per 100k (1 lag)			0.00292	0.00287
			(0.00314)	(0.00478)
Log deaths per 100k			-0.00543	-0.00543
			(0.00365)	(0.00641)
Log deaths per 100k (1 lead)			0.00023	0.00024
			(0.00378)	(0.00498)
Log deaths per 100k (2 lead)			0.00206	0.00205
			(0.00393)	(0.00667)
Log deaths per 100k (3 lead)			0.00005	0.00003
			(0.00382)	(0.00543)
Log deaths per 100k (4 lead)			0.00009	0.00004
			(0.00331)	(0.00323)
Fixed-effects				
factor(State) (48)		Yes		Yes
Fit statistics				
Observations	3,138	3,138	3,120	3,120
R^2	0.00078	0.00091	0.00076	0.00093
Within R^2		0.00078		0.00076

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

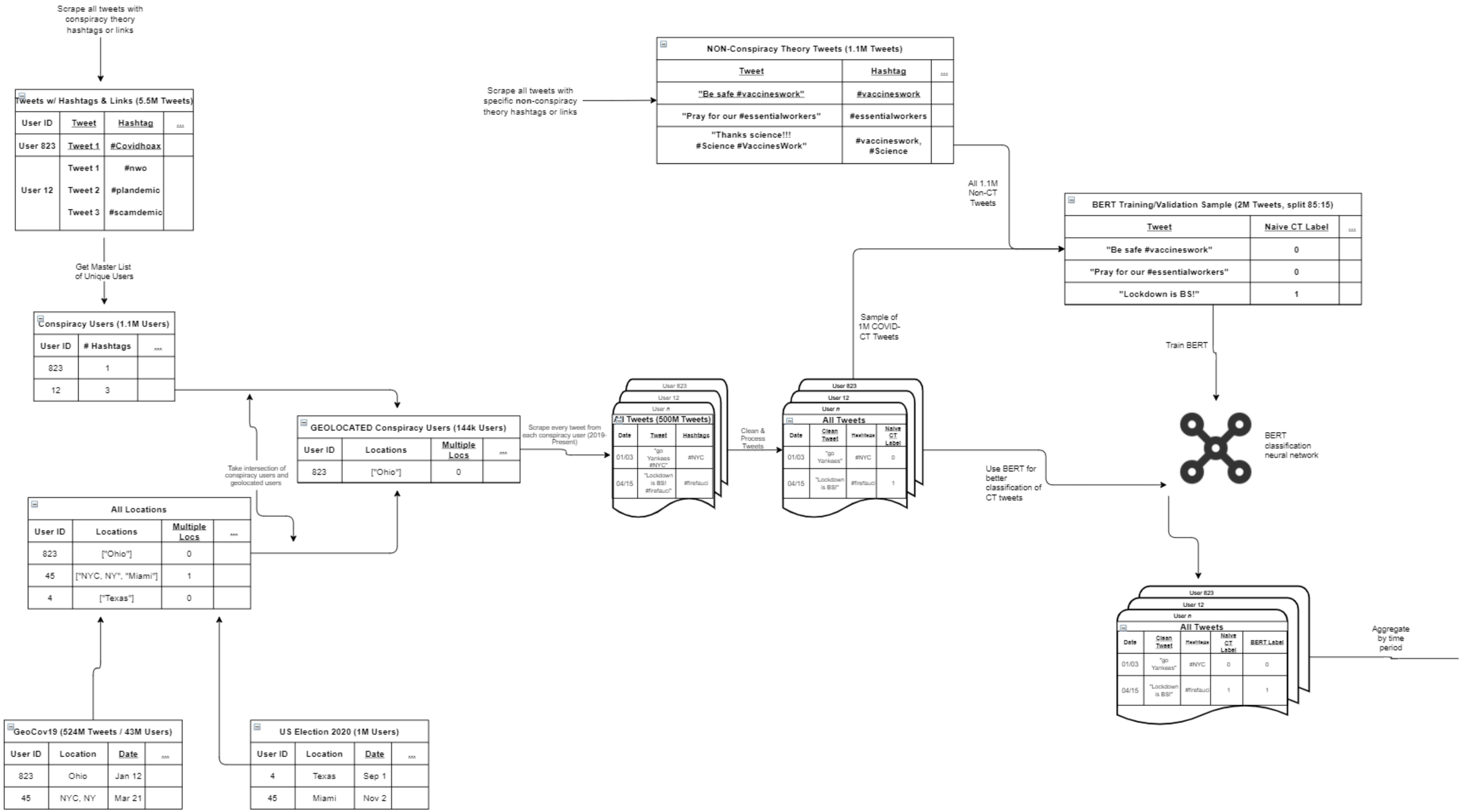
Table A.7: Specification 2 (State-by-Month)



Dependent Variable:	Stopped CT Tweets			
Model:	(1)	(2)	(3)	(4)
Variables				
Intercept	0.01373		0.00485	
	(0.01525)		(0.00889)	
Log cases per 100k (1 lag)	-0.00459*	-0.00467*		
	(0.00260)	(0.00233)		
Log cases per 100k	0.00213	0.00189		
	(0.00558)	(0.00436)		
Log cases per 100k (1 lead)	0.00007	-0.00009		
	(0.00365)	(0.00244)		
Log deaths per 100k (1 lag)			-0.00550	-0.00591*
			(0.00339)	(0.00312)
Log deaths per 100k			0.00217	0.00212
			(0.00476)	(0.00501)
Log deaths per 100k (1 lead)			0.00094	0.00033
			(0.00402)	(0.00358)
Fixed-effects				
factor(State) (48)		Yes		Yes
Fit statistics				
Observations	679	679	679	679
R^2	0.00691	0.00979	0.00535	0.00825
Within R^2		0.00743		0.00588

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Flowchart 1


Scrape all tweets with conspiracy theory hashtags or links
Get Master List of Unique Users
Take intersection of conspiracy users and geolocated users
Scrape every tweet from each conspiracy user (2019-Present)
Clean & Process Tweets
Scrape all tweets with specific non-conspiracy theory hashtags or links
All 1.1M Non-CT Tweets
Sample of 1M COVID-CT Tweets
Train BERT
BERT classification neural network
Use BERT for better classification of CT tweets
Aggregate by time period

Tweets w/ Hashtags & Links (5.5M Tweets)

User ID	Tweet	Hashtag	...
User 823	Tweet 1	#Covidhoax	
User 12	Tweet 1	#nwo	
	Tweet 2	#plandemic	
	Tweet 3	#scamdemic	

Conspiracy Users (1.1M Users)

User ID	# Hashtags	...
823	1	
12	3	

All Locations

User ID	Locations	Multiple Locs	...
823	["Ohio"]	0	
45	["NYC, NY", "Miami"]	1	
4	["Texas"]	0	

GeoCov19 (524M Tweets / 43M Users)

User ID	Location	Date	...
823	Ohio	Jan 12	
45	NYC, NY	Mar 21	

US Election 2020 (1M Users)

User ID	Location	Date	...
4	Texas	Sep 1	
45	Miami	Nov 2	

GEOLOCATED Conspiracy Users (144k Users)

User ID	Locations	Multiple Locs	...
823	["Ohio"]	0	

User 823
User 12
User *n*

All Tweets (500M Tweets)

Date	Tweet	Hashtags
01/03	"go Yankees #NYC"	#NYC
04/15	"Lockdown is BS! #firefauci"	#firefauci

User 823
User 12
User *n*

All Tweets

Date	Clean Tweet	Hashtags	Naive CT Label
01/03	"go Yankees"	#NYC	0
04/15	"Lockdown is BS!"	#firefauci	1

NON-Conspiracy Theory Tweets (1.1M Tweets)

Tweet	Hashtag	...
"Be safe #vaccineswork"	#vaccineswork	
"Pray for our #essentialworkers"	#essentialworkers	
"Thanks science!!! #Science #VaccinesWork"	#vaccineswork, #Science	

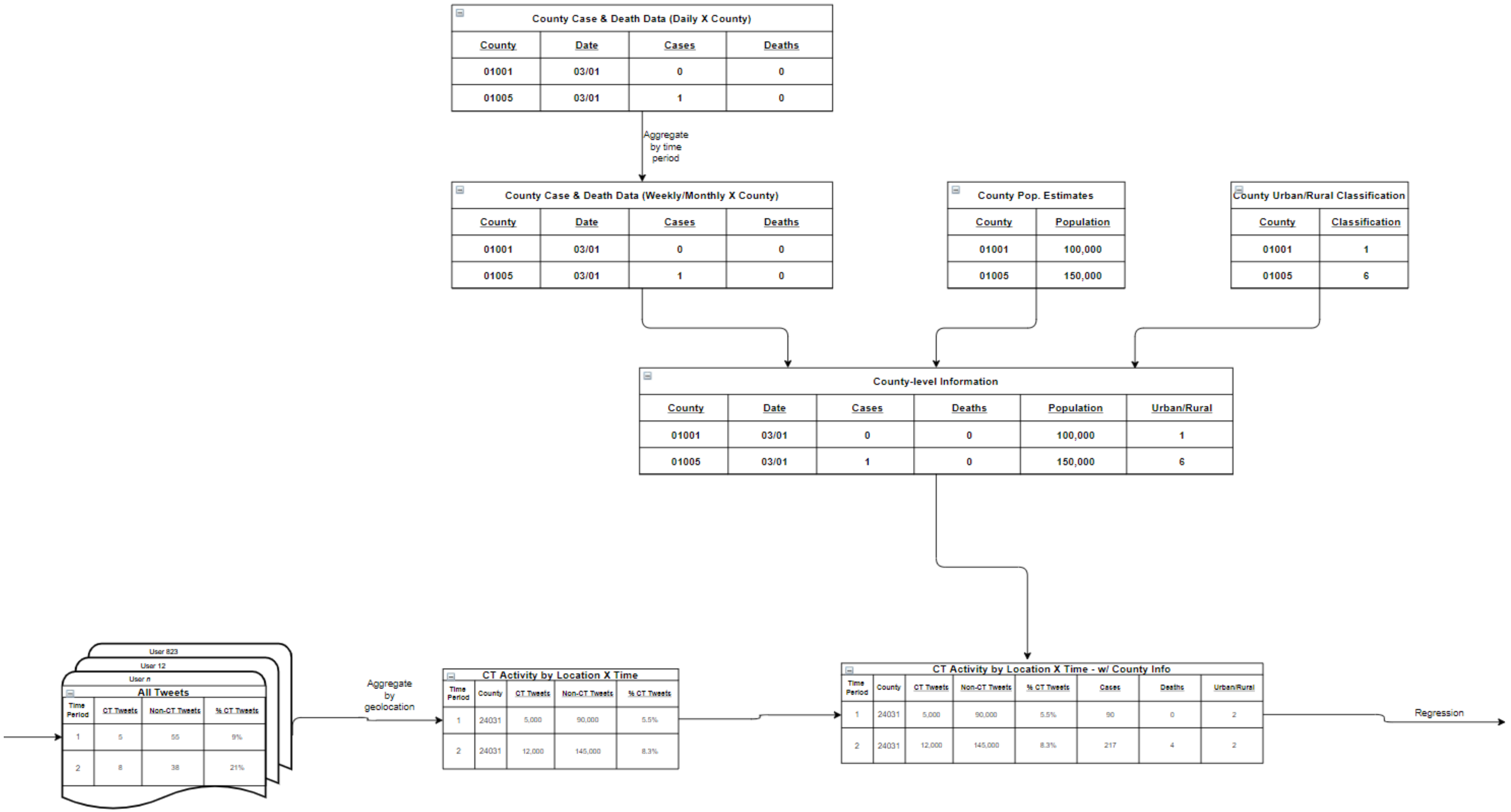
BERT Training/Validation Sample (2M Tweets, split 85:15)

Tweet	Naive CT Label	...
"Be safe #vaccineswork"	0	
"Pray for our #essentialworkers"	0	
"Lockdown is BS!"	1	

User 823
User 12
User *n*

All Tweets

Date	Clean Tweet	Hashtags	Naive CT Label	BERT Label
01/03	"go Yankees"	#NYC	0	0
04/15	"Lockdown is BS!"	#firefauci	1	1



County Case & Death Data (Daily X County)

County	Date	Cases	Deaths
01001	03/01	0	0
01005	03/01	1	0

County Case & Death Data (Weekly/Monthly X County)

County	Date	Cases	Deaths
01001	03/01	0	0
01005	03/01	1	0

County Pop. Estimates

County	Population
01001	100,000
01005	150,000

County Urban/Rural Classification

County	Classification
01001	1
01005	6

County-level Information

County	Date	Cases	Deaths	Population	Urban/Rural
01001	03/01	0	0	100,000	1
01005	03/01	1	0	150,000	6

All Tweets

Time Period	CT Tweets	Non-CT Tweets	% CT Tweets
1	5	55	9%
2	8	38	21%

CT Activity by Location X Time

Time Period	County	CT Tweets	Non-CT Tweets	% CT Tweets
1	24031	5,000	90,000	5.5%
2	24031	12,000	145,000	8.3%

CT Activity by Location X Time - w/ County Info

Time Period	County	CT Tweets	Non-CT Tweets	% CT Tweets	Cases	Deaths	Urban/Rural
1	24031	5,000	90,000	5.5%	90	0	2
2	24031	12,000	145,000	8.3%	217	4	2