Who Can Social Distance?
The Impact of Socioeconomic and Behavioral Factors on the Effectiveness of COVID-19 Mitigation Policy

Ciara Patterson
April 2021

TABLE OF CONTENTS

TABLE OF FIGURES

1. INTRODUCTION

When the first COVID-19 cases emerged in the United States, policymakers quickly moved to limit in-person interactions to curb the spread of the novel coronavirus. In the absence of a vaccine or effective antiviral treatment, these "social distancing" policies were some of the only methods available to protect people from COVID-19. Governors implemented shelter-in-place and stay-at-home orders in addition to a suite of non-pharmaceutical interventions (NPIs). These interventions include restrictions on business occupancy, mask mandates, contact tracing, etc.[1,2] Similar policies designed to encourage "social distancing" were also enacted in South Korea, Singapore, the United Kingdom, and the European Union.[3] The emerging consensus is that implementing those social distancing policies before the number of new cases begins to grow exponentially protects individuals from becoming infected with the virus, lessens the burden on the healthcare system, and subsequently saves lives.[4–6] Yet, some regions with social distancing policies in effect still observed significant numbers of COVID-19 cases. Washington state implemented relatively strict measures through early May, but the epidemic trajectory still varied between its counties. Yakima County recorded the highest case rate of any county on the West Coast while the nearby King County saw declines in COVID-19 transmission.[7] Thus, while social distancing policies are effective, their success may vary by region.

The virus's impact has varied between populations as well. Black and Latinx people and frontline workers have all been disproportionately infected by COVID-19. In July 2020, an analysis by the New York Times of 640,000 COVID-19 infections identified across 974 American counties found that Latinx and Black individuals were almost three times more likely to become infected with COVID-19 than their white neighbors.[8] Another study in the UK also found that individuals in "essential" occupations were more likely to become infected with

COVID-19 than people employed in “nonessential” occupations. These essential workers were deemed necessary to maintain vital societal functions, like providing people with food, healthcare, and transportation. They continued to go to work as their neighbors stayed home. Therefore, while social distancing policies may have been critical to mitigating the virus’s impact, there is a need to further understand how COVID-19 spread amongst these vulnerable populations and if social distancing policies adequately protected these people.

The complexity of human behavior and decision-making complicates our understanding of the effectiveness of COVID-19 mitigation policy. Some people, like essential workers, were unable to comply with social distancing recommendations.[9] Others may have chosen not to social distance due to a fear of isolation, a distrust of government messaging, or apathy towards the virus’s effects.[10,11] The success of any policy is limited by people’s ability and willingness to comply with that policy. We will hopefully aid the understanding of COVID-19 mitigation policy success through our investigation of the relationship between social factors and the COVID-19 epidemic trajectory in areas that implemented social distancing policies.

In this study, we examined the COVID-19 caseload burden in several US counties with social distancing policies in effect. By only examining counties with similar policy environments, we effectively controlled for at least some of the impact of policy on COVID-19 transmission. Then we examined which characteristics of the county’s residents would affect the policy’s success at reducing the caseload burden in those areas. To do so, we trained a random forest model to predict the average new COVID-19 cases observed during the policies’ enactment and then investigated which characteristics of the county’s residents best accounted for the difference in the number of COVID-19 cases between counties. Understanding the relative importance of each variable in making that prediction allows us to better understand how

COVID-19 continued to spread in areas with restrictions in place. The variables studied include the county residents' socioeconomic status, mobility, race, age, voting behavior, and occupation. While the spread of COVID-19 has been influenced by a variety of factors, we carefully reviewed the relevant literature to identify which characteristics most determine a person's ability and willingness to social distance.

1.1 SOCIAL DISTANCING POLICIES

The U.S. and international policy responses to the COVID-19 pandemic were based on a robust body of evidence, including scientific modeling and historical evidence from previous pandemics. In 2007, Markel et al. studied 48 US cities during the 1918 Spanish Influenza epidemic and found that an early and layered application of NPIs (e.g., school closures and public gathering restrictions in effect simultaneously) delayed the peak of the epidemic and lowered the total mortality burden of the disease.[12] The study's lead author, Dr. Markel, was part of the initial team of scientists that urged the Bush administration to make social distancing the official federal policy recommendation in the event of pandemic influenza.[13] In 2006, Glass et al. analyzed the impact of targeted social distancing on a simulation of the spread of an influenza virus in a modeled community, "representative of a small town in the US." The study concluded, "results for our stylized small town suggest that targeted social distancing strategies can be designed to effectively mitigate the local progression of pandemic influenza without the use of a vaccine or antiviral drugs."[14] Therefore, before the COVID-19 pandemic, there was substantial evidence that social distancing would mitigate the caseload burden of pandemic influenza in the United States before the COVID-19 pandemic.

Historical evidence also suggested that the effectiveness of certain containment strategies would vary between regions. For instance, during the 2003 SARS outbreak, health authorities in Vietnam credited the country's early containment of the outbreak to the Vietnamese government's early detection, isolation, and quarantine measures. However, officials in Taiwan now believe that Taiwan's aggressive use of quarantine was ultimately counterproductive and contributed to public panic surrounding SARS.[15] Rothstein et al. further theorized that if similar measures were implemented in the United States — "a heterogeneous society with a strong tradition of individualism and skepticism about government" — rates of compliance amongst the affected population may be lower than observed in the countries affected by the SARS outbreak.[15] Thus, while there was evidence that policies could contain novel viral threats, their exact effectiveness was unknown and subject to regional variation.

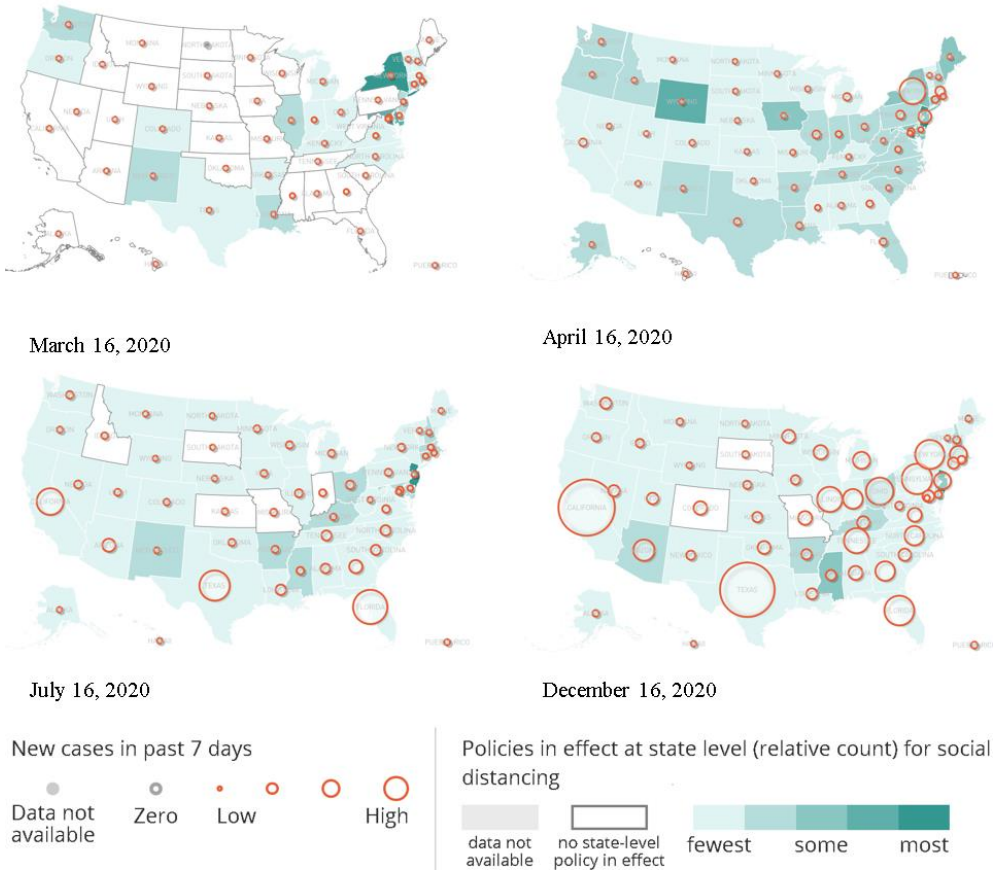

Figure 1. The number of adaptation and mitigation measures, private sector closures, and mass gathering restrictions implemented at the state-level throughout the COVID-19 pandemic.[16]

When the pandemic began, decision-makers began to act on this body of evidence, and researchers subsequently analyzed which strategies were most successful at containing SARS-CoV-2. Current research suggests that social distancing policies are effective at reducing COVID-19 caseload. In a study of 10 countries, Thu et al. found that daily confirmed cases and deaths showed signs of decreasing 1-4 weeks after the "highest level of social distancing measures" were enacted. However, the effectiveness of social distancing measures varied widely between the 10 countries studied by Thu et al. At the country level, these differences may be attributed to the different political systems in each country, the severity of the policies enacted, and the spread of the virus prior to the policies.[17] Other studies have analyzed the impact of mitigation policy on not just COVID-19 caseload but on the COVID-19 case fatality rate as well.

In a study of all business closures and related restrictions implemented in every county in the United States since March 2020, Spiegel and Tookes found that both mandatory mask mandates and closures of high-risk businesses, like restaurants and bars, were associated with fatality growth rates that were about 1% lower than the growth rates in areas without those mandates in place.[1] Therefore, not only are these policies effective in simulations and historical instances, they are effective at mitigating the burden of COVID-19 as well. However, conducting such research relies on successfully modeling COVID-19 and predicting population-level outcomes which have proven to be a complex and difficult task.

Modeling COVID-19 is difficult, in part, because the spread of this virus heavily depends on unknown parameters. In an assessment of the modeling problem, Subramanian and Kattan acknowledge that "the duration and degree of immunity conferred by exposure will determine the likelihood and intensity of recurrent outbreaks." However, such biological properties of SARS-CoV-2 are not fully understood yet.[18] Additionally, Subramanian and Kattan acknowledge that the spread of COVID-19 depends on two other factors that are not precisely known: the prevalence of COVID-19 in the population and the implementation of policy interventions. Studies of antibody tests suggest that the prevalence of COVID-19 may be underestimated.[18] Policy interventions change rapidly and are not easily quantifiable.[18] These are just a few of the variables that must be considered when designing an effective model. To predict population-level COVID-19 outcomes, models must either use proxies for or make assumptions about a variety of variables.

Therefore, researchers must carefully choose what assumptions to make and which models to employ. Many scientists use compartmental epidemic models, like the Susceptible-Exposed-Infectious-Removed (SEIR) model, to model the spread of infectious. As its name

suggests, The SEIR model captures how a virus spreads as susceptible populations interact with infectious individuals. While this model is standard in epidemiological research, its basic structure may be modified. Allison Hill, an assistant professor at Johns Hopkins University, separated the infectious compartment into mild infection, moderate, and severe compartments in her SEIR model.[19] Doing so better represents how people with differing degrees of COVID-19 severity interact less with the general population (in the case of severe or hospitalized infections) or are less contagious (in the case of asymptomatic infections). Eikenberry et al. employed a similarly structured SEIR model to analyze COVID-19 transmission in New York and Washington, but they further divided the compartments into masked and unmasked individuals[20] Deciding on how to structure an epidemiological model is just one example of the difficult decisions that researchers have to make as they attempt to capture complex population and viral dynamics. Many researchers may decide to use modeling techniques other than the SEIR model.

Some scientists may choose to use analytical methods from statistics or econometrics instead of more traditional epidemiological approaches while others may combine these approaches with an SEIR model. Delen et al. fit a compartmental model to COVID-19 caseload and then used gradient boosted trees to identify the relationship between cellphone mobility data and the modeled COVID-19 transmission rates.[21] Kaur et al. used Bayesian change-point-analysis to analyze if a significant change in the epidemic trajectory occurred following a stay-at-home order.[6] Choosing what method to use depends on the goal of the study. If a scientist is interested in understanding COVID-19 transmission rates, then that scientist will need to use an SEIR model to compute transmission rates between the susceptible and the infected individuals. However, if that person is instead more interested in understanding the relationship between the

case count and mobility then a simpler statistical technique may be appropriate. In any case, researchers often must attempt to quantify variables that are not easily measured or understood.

1.2 MOBILITY AND COVID-19

Due to the difficulty of measuring "social distancing," many studies rely on cellphone location data to quantify the behavior of an area's residents in response to rising cases or policy implementation.[21–23] In a study of 25 US counties with the highest number of confirmed cases as of April 16th, 2020, Badr et al. found a strong correlation between a decrease in mobility (which dropped 35-63% relative to pre-pandemic conditions) and decreased COVID-19 case growth in the most affected counties ($r > .7$ in 20 of the 25 counties evaluated). That study relied on mobility data from Unacast, a provider of cell phone location data.[24] Chang et al. successfully built an SEIR model that predicted caseload in several metropolitan statistical areas in the U.S. using only mobility data provided by Safegraph, another company specializing in location data.[25]

The changes in mobility observed in the United States may occur in response to policy or independently of it. Engle et al. analyzed the relationship between the enactment of state-wide stay-at-home orders and reductions in daily mobility in 3,142 U.S. counties. Engle et al. concluded that official stay-at-home orders reduced daily mobility by 7.87%, but "a rise of local infection rate from 0% to 0.003% is associated with a reduction in mobility by 2.31%." The study's authors argue that this indicates that the perceived risk of contracting COVID-19 has a substantial impact on a person's behavior as well.[26] Gupta et al. examined the relationship between official emergency declarations and hours spent at home between March 1 and April 11. Gupta et al. conclude that "55% of the growth [in hours spent at home] comes from emergency declarations and 45% comes from secular (non-policy) trends."[27] Additionally, every state saw

increases in people staying at home before a stay-at-home order was implemented.[28] These conclusions align with evidence from the H1N1 epidemic as well. During that epidemic, measures to reduce interactions between people were not routinely recommended in the US.[29] However, polls found that 16 to 25% of Americans avoided "places where many people are gathered, like sporting events, malls, or public transportation," and 20% had "reduced contact with people outside household as much as possible."[29] While policy may be influential on the course of the COVID-19 epidemic in an area, people's behavior and decision-making are affected by factors independent of policy.

1.3 RESPONSIVENESS TO SOCIAL DISTANCING POLICIES

Some evidence suggests that people that live in denser areas where they are more likely to contact other individuals may be more responsive to social distancing policies and recommendations than people living in less dense areas. Lee et al. found that individuals in areas with a high population density traveled less than those in less-dense areas in mid-March. Lee et al. suggest that perhaps people in high-density areas practice social distancing more actively because of "the higher chances of contacting other people in the higher density area."[30] This finding aligns with Engle et al.'s study which found that residents of counties with a higher population density are more likely to reduce their mobility in response to both increased COVID-19 prevalence and stay-at-home orders.[26] A person's individual decision to social distance is thus substantially influenced by their surrounding environment.

People of different ages may also exhibit differing behavior in response to COVID-19. Engle et al. also found that "counties with larger shares of the population over age 65" were also more responsive to disease prevalence and restriction orders.[26] Monod et al. suggest that adults

age 20 to 49 have disproportionately contributed to the spread of COVID-19 relative to their size in the US population. This disproportionate impact may be partially attributable to "age-specific behavioral differences in, for example, consistent social distancing, mask use, duration of visits, or types of venues visited."[31] Therefore, age is another factor that can impact a person's decision-making and behavior.

Partisanship is yet another factor that might influence a person's willingness to social distance. Clinton et al. argue that in the United States, partisan preference informs the information that people collect, process, and respond to as well as the actions that they take"[32] Partisanship has been shown to impact a person's decisions regarding public health and healthcare. Evidence from the 2009 H1N1 pandemic suggests that partisan media viewing shaped Americans' perspectives on H1N1 vaccines.[33] During the COVID-19 pandemic, several researchers have found links between partisanship and mobility in the US.[32,34] Gollwitzer et al. found that counties that voted for the Republican candidate, Donald Trump, over the Democrat candidate, Hillary Clinton, in the 2016 presidential elections exhibited 14% less physical distancing than the majority Democrat counties studied. Similar partisan differences emerge in opinion polls on COVID-19 as well.[35] Liberals are often more worried that someone in their family will catch the virus.[36] Additionally, liberal individuals are also more likely to say that they have practiced social distancing in the past 24 hours and more likely to endorse the use of a variety of social distancing measures.[37] In conclusion, income, age, surrounding population density, and partisan preference are all factors that influence a person's likelihood of social distancing in response to COVID-19. At the population level, any one of these variables may impact the population's success at containing the spread of COVID-19.

1.4 ESSENTIAL WORKERS

While some people were less willing to stay home, other people were unable to stay at home. Essential occupations are defined as any occupation that is needed to maintain critical infrastructure.[38] Many of the people employed in these essential occupations were unable to conduct their work from home. Those essential workers who could not conduct their occupations from home are often referred to as frontline workers. Frontline workers were often exempt from broader social distancing policies, although the exact definition of an essential occupation varied by state.[38] Because of their inability to social distance and the potential for exposure in the workplace, frontline workers may have been more susceptible to COVID-19 infection. Zhang and Warner found that the COVID-19 growth rate increased in states with higher proportions of essential workers.[39] Not only are frontline workers unable to socially distance, but they are often more socioeconomically vulnerable than other populations.

Frontline workers typically earn lower wages and have less education than the rest of the U.S. employed population. The average wages of a frontline worker are lower than those of all other workers and other essential workers who work in critical occupations but may conduct their work from home.[9] Frontline workers are also less likely to obtain a high school education or to have a four-year degree than employees in other sectors.[9] Both educational attainment and income are associated with a variety of adverse health outcomes.[40] Thus, frontline workers may be vulnerable on two fronts. They are both unable to protect themselves from exposure to COVID-19 and potentially more susceptible to adverse outcomes in the event of infection.

Not only are frontline workers vulnerable due to working conditions, but they often reside in households with other vulnerable individuals. An estimated 13% of essential workers live in high-risk households, where a high-risk household was defined as having “(1) low

household income (below $40 000); (2) uninsurance (at least 1 person in the household is uninsured); and (3) household presence of 1 or more persons aged 65 years or older." 48% of essential workers lived in a household with at least one risk factor.[41] If an essential worker becomes exposed to COVID-19, they may bring the virus home to other vulnerable household members.

COVID-19 spreads rapidly in both households and the workplace. A meta-review of COVID-19 spread in several settings found that households were associated with the highest transmission rates.[42] A recent report from the King County Department of Public Health found that 27% of all confirmed COVID-19 cases had a potential exposure in their household. Workplaces are sites of transmission as well. 18% of all COVID-19 cases had a COVID-19 exposure at a non-healthcare workplace. 20% of cases had an exposure in a healthcare setting, but that statistic includes healthcare workers who may have been exposed on the job.[43] Therefore, frontline workers may be exposed to the virus at work then subsequently infect other members of their household.

1.5 RACIAL DISPARITIES IN COVID-19 INFECTIONS

Members of racial minorities, especially Black and Hispanic Americans, have been disproportionately impacted by COVID-19. Many studies have shown that people of color are experiencing disproportionate COVID-19 cases, hospitalizations, and deaths.[44] A study conducted by the US Centers for Disease Control and Prevention (CDC) found that age-adjusted COVID-19 hospitalization rates for Black and Hispanic Americans were 4.5 and 3.5 times higher respectively than those of white Americans, as of May 30, 2020.[45] A county-level study found that higher shares of Black people living in the area were associated with increased shares of

COVID-19 cases and deaths in that county. These disparities persisted even after the authors controlled for characteristics such as age, poverty, comorbidities, and epidemic duration.[46]

These inequities are emerging because of a history of systemic racism and segregation in the United States. Although segregation based on racial identity has been illegal since the 1960s, its legacy has persisted through "an interlocking set of individual actions, institutional practices, and governmental policies."[47] Such long-standing structural inequalities affect a wide range of outcomes including a person's income, health care, employment, physical health, and living circumstances.[48] As a result, people of color more likely to live and work in environments where they may be exposed to COVID-19. People of color are more likely to live in crowded areas and more likely to work in frontline industries. Additionally, people of color are more likely to or live in a household with at least one frontline worker.[9] Black and Hispanic adults at high risk for severe illness are also more likely to live with at least one worker who was unable to work from home than white adults.[49] Thus, longstanding inequities in the US have caused a reality in which people of color are more likely to live and work in environments where they are more susceptible to COVID-19.

2. PROBLEM STATEMENT

We have thus outlined a variety of factors that may impact an individual's ability or willingness to social distance. Even when social distancing policies are implemented and enforced, some individuals may continue to venture outside their homes to work essential jobs. Other individuals may choose to not social distance for reasons unrelated to occupation. Their behavior may be informed and influence by their income, age, living environment, and partisan

preference. The ability or willingness of any one person to social distance may then impact the COVID-19 caseload burden.

To understand the relationship between these factors and the number of COVID-19 cases in a region, we controlled for social distancing policy and built a predictive model that estimated a region's caseload burden given a variety of social factors. The goal of this study is to contribute to a growing body of literature that attempts to identify who can and will change their behavior in response to social distancing policies and rising COVID-19 cases. This information can help policymakers better understand how their residents will behave during a pandemic and shape policy accordingly.

We collected data on 360 US counties that implemented certain social distancing policies during the pandemic. Counties were identified using policy data obtained from the COVID Analysis and Mapping of Policies (AMP) database, a publicly available and comprehensive database of policies and plans implemented in response to COVID-19.[16] Data related to the socioeconomic status, voting behavior, mobility patterns, etc. of the population was retrieved from the American Community Survey, the MIT Election and Data Science Lab, Safegraph's Social Distancing Dataset, and the New York Time's COVID-19 Github repository.[50–53] We then obtained information on confirmed COVID-19 cases at the county level from the New York Times.[54] Using that information, we constructed a random forest model that accounted for variance in the caseload burden using data on the demographics, socioeconomic status, and behavior of the county's residents. The random forest model then produced a list of the variables that were the most "significant" or "important" in generating its prediction of the COVID-19 caseload burden in the studied areas.

3. LITERATURE REVIEW

To our knowledge, very few if any previous studies have attempted to control for the effects of policies before studying the relationships between other social factors and COVID-19 infections. As previously discussed, measuring the impact of social distancing policies is a difficult endeavor. By controlling for policies in this way, we could simply assume that policies would have some effect on the COVID-19 caseload without attempting to quantify that exact impact. This approach inverts another approach employed by Spiegel and Tookes. Spiegel and Tookes examined the different case fatality rates in counties that enacted different policies while holding other independent variables such as the weather, age, mobility, voting behavior, etc. as control.[1] By contrast, we held social distancing policy as control and sought to then identify those non-policy related variables that might explain the remaining variance in COVID-19 caseload.

No other researchers may have attempted to control for social distancing policies in part because they lacked policy data with the granularity and breadth of the COVID AMP database and because much of the initial research in COVID-19 focused only on statewide stay-at-home orders. Engle et al. examined the relationship between age, partisanship, population density, and mobility, but they relied on policy data from the New York Times that only included information on statewide stay-at-home orders. Their study did not account for how social distancing measures varied by county and only considered one broad category of policy.[26] By contrast, the policy data in COVID AMP used in this study contains data on multiple categories of policy and policies implemented at the federal, state, and local level.[16]

Similar studies have been conducted which attempt to model the relationship between several socioeconomic factors and the effectiveness of COVID-19 mitigation policy. Gao et al.

proceeded by first examining the relationship between statewide stay-at-home orders and rates of COVID-19 transmission. The authors then constructed a multiple linear regression model that predicted changes in median home dwell time (as reported by Safegraph) from data about socioeconomic variables (obtained from the American Community Survey). However, while the authors concluded that socioeconomic variables explained approximately 69% percent of the variance in median home dwell time, they could not differentiate between each variable's impact on that measure of mobility.[22] A multiple linear regression does not perform well on input variables that are correlated with one another. Variables like "Proportion of Population under Age 18" and "Proportion of Population between Age 18 to 44" are necessarily dependent since they are sampled from the same population and mutually exclusive. Thus, if we wish to examine the impact of socioeconomic variables on an outcome, we must be aware of the relationships between such variables and select a model that is robust to multicollinearity. Doing so is one of the difficulties in examining the relationship between COVID-19 cases and more than one socioeconomic or behavioral factor.

Another difficulty in answering our question is determining where exactly essential workers live and work. No nationwide central repository on where essential workers live and work exists, so researchers must estimate this value using a variety of other methods. Methods may vary according to whether the study's authors are interested in all people employed in occupations that have been deemed critical, or just employees in frontline occupations, who are at a higher risk of COVID-19 exposure. Reitsma et al. used a list based on the Department of Homeland Security's guidance on critical industries, which includes some occupations that may be conducted from home. Reitsma et al. then used the American Community Survey 2014-2018 5-year estimate in conjunction with that list of Standard Occupational Classification codes to

identify households with more people than rooms and at least one essential worker in the residence with the assumption that residents of these households were at a high risk of COVID-19 infection.[55] Mongey et al. used the Occupational Information Network (O*NET), to construct a measurement of the likelihood that an occupation could be conducted from home and then combined the resulting dataset with employment information from the Bureau of Labor Statistics to approximate the number of people that could work from home. They validated this measure by verifying that it was "(i) uncorrelated with pre-epidemic mobility as measured using cellphone data from SafeGraph, but (ii) strongly correlated with the change in mobility during the epidemic."[56] These studies informed our approach to constructing and validating an estimate of the number of frontline workers living in a given area.

We selected our variables of interest according to the research outlined in the introduction and surveyed the literature to ensure that our data sources and methods of analysis aligned with other researchers. For instance, we used 2016 voting behavior as a proxy for partisanship at the county level because Parker et al. found correlations between partisanship and individual mobility using the same county-level election data from the Massachusetts Institute of Technology's Election Data and Science Lab.[34] Data from Safegraph and the American Community Survey was also used in the literature to determine population-level characteristics that might affect a person's susceptibility to COVID-19.[22,57]

Therefore, while our selection of predictor variables was heavily influenced by existing literature, to our knowledge no other study has attempted to hold policy variables as control while studying the relationship between multiple socioeconomic and behavioral factors and COVID-19 cases.

4. METHODS

4.1 RESEARCH APPROACH

To better understand the relationship between social determinants of health – factors unrelated to medical care that nonetheless affect health outcomes – and COVID-19 infections, we constructed a random forest regression model using data related to individuals' mobility, socioeconomic status, partisanship, etc. and the number of confirmed COVID-19 cases in those same locations. This analysis was conducted at the county level. Counties are the most granular geographical unit at which COVID-19 case data is widely available and collected into a single national repository.[54] We only modeled counties that had similar social distancing policies in place in response to the pandemic and chose to control for policies to emphasize the amount that other factors, especially social determinants of health, contribute to the COVID-19 caseload observed in an area. In this section, we will describe how we identified counties with similar policy environments, the datasets that we used to collect information on each of the counties identified, and the random forest regression model that we constructed from that information.

4.2 SELECTING SIMILAR POLICY ENVIRONMENTS

We obtained information on social distancing policies from the COVID AMP database, a continually updated (as of March 2021) database of policies implemented in response to COVID-19. The database is published and maintained by Talus Analytics and the Georgetown Center for Global Health Science and Security. The COVID AMP research team defines a policy as "government-issued and backed by legal authority or precedent."[16] The COVID AMP

database includes policies implemented by the U.S. federal government, the District of Columbia, U.S. state governments, some U.S. counties and cities, and other national governments. To catalog the policies in the database, the COVID AMP team has developed a custom data ontology to represent each policy's content, impact, and legal authority. The data is collected and curated by researchers at the Georgetown University Center for Global Health Science and Security and Talus Analytics. As of December 2020, data is complete for most U.S. states but is only available for select U.S. counties and cities.[16] However, the data within COVID AMP is robust and well-documented at the local level, even if it is incomplete for some areas.

We theorized that we could effectively control for an area's policy environment by searching the COVID AMP database for policies with certain attributes and then checking if, for any time period and area, these policies were simultaneously in effect. While COVID AMP captures a variety of policy types, we focused only on social distancing policies since previous studies have substantiated the effects of those policies on the transmission of infectious viral diseases.[1,3,5,25]

To ensure that our dataset would not be affected by incomplete data, we first analyzed the COVID AMP database to determine which states had a high proportion of counties recorded within the policy database. Only states for which over 95% of counties had been recorded as an "Affected Local Area" in the COVID AMP policy database were included. 95% was selected as a threshold to ensure that states with a significant number of counties were not excluded from the analysis because a single county had not been recorded completely. Additionally, we confirmed that at least 95% of their population was represented by the included counties to ensure that the major metropolitan areas in the state were represented in the recorded counties. The omission of a major metropolitan area might signal that the included policy data for that state was missing

substantial information. The following states and the District of Columbia met the above criteria: California, Illinois, Indiana, Iowa, Mississippi, Nebraska, New Mexico, New York, Oregon, Pennsylvania, Utah, and Washington. Once states with substantial policy data had been identified, we searched for counties within those states that met certain criteria.

Ensuring that policies had similar stringency and intent was essential to effectively control for the impact of social distancing policies on COVID-19 caseload. We only considered policies that were coded as having a "restricting" intent. Restricting is a relative term that refers to any policy that "was introduced to decrease the overall interactions and activity among individuals." Alternatively, the policy could be coded as "relaxing" which indicates that the policy's aim was "to increase activity and interactions towards pre-outbreak levels."[16] Restricting policies are thus often implemented in response to a rise in COVID-19 cases. They both impose restrictions and can serve as a public messaging that cases as rising. What exactly constitutes a restricting policy is determined by the affected state's previous policies. For instance, a policy that imposes 75% occupancy limits on all private businesses may be seen as restricting or relaxing depending on the state's previous occupancy limit. Thus, all restricting policies are not of the same severity. This is a limitation of our method, but we manually reviewed relevant policies to ensure that there were no significant differences in the severity of the restrictions.

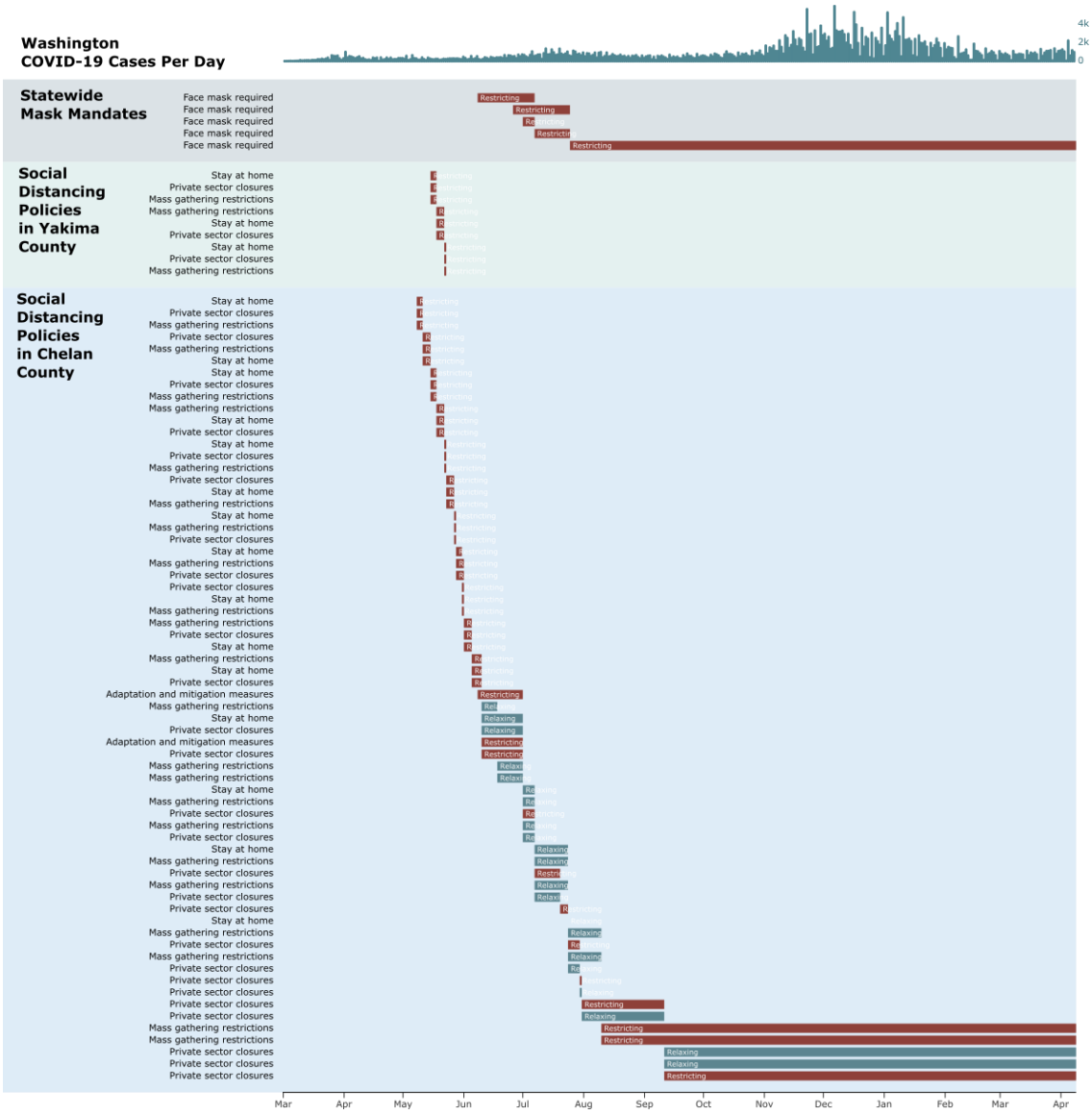

Figure 2. A timeline of relevant policies in two counties, Yakima County and Chelan County, within Washington state. The length of the bar represents the policies' duration.[16]

For a county to be included in the study the following policies had to be in effect simultaneously: a face mask requirement, the closure of all non-essential business or specifically the closure of restaurants, bars, and entertainment venues, and restrictions on the number of people allowed to gather in private or public. These policies could be effective statewide or county-level. Every policy – excluding those that dealt with private sector closures – had to be aimed at the general population (as indicated by the "Policy subtargets" attribute in the COVID

AMP database). These criteria were developed with members of the COVID AMP research team and align with emerging evidence that these specific policy types mitigate COVID-19 cases in the affected area.[1,3,5,25] Figure 3 shows the specific values that corresponded to this criteria in the COVID AMP database. Policies were understood to be in effect from the "Actual Start Date" to the "Actual End Date" recorded in COVID AMP. If the actual end date attribute was empty, then it was assumed that the policy was still in effect. This assumption was corroborated by the COVID AMP data collection team.

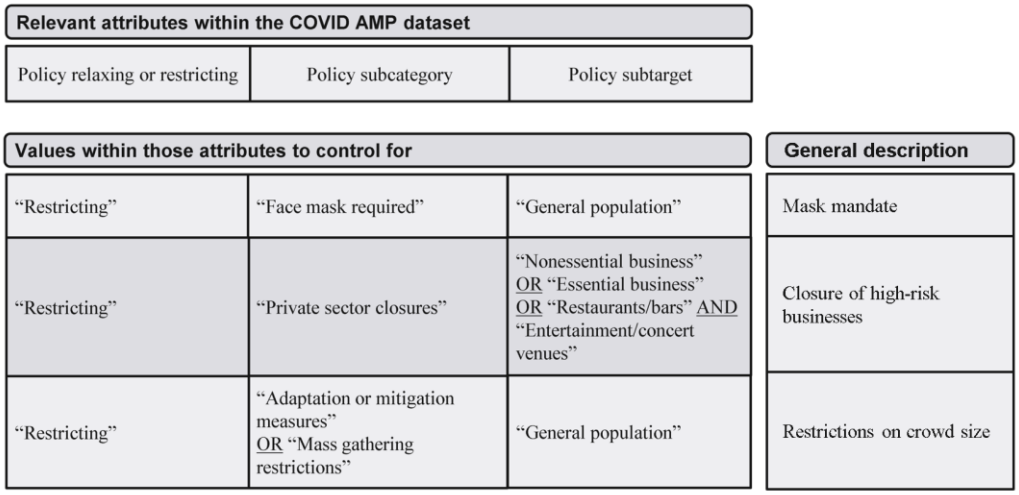

Figure 3. Outline of the queries that were used to filter the COVID AMP database.

In total, 360 counties were identified that met the policy criteria between April 19th, 2020 and January 10th, 2020, the day on which we searched the COVID AMP database. No policies that met the criteria were identified in March in those states which is unsurprising since many state governors and public officials first implemented stay-at-home orders before transitioning to a more specific policy response later in the pandemic.[16,58] Table 1 shows the U.S. states with the number of counties that met the criteria.

Table 1. The number of counties included in the study by state.

State	Number of counties
Illinois	101

Pennsylvania	67
New York	57
Washington	39
Oregon	36
New Mexico	33
California	25
Indiana	1
District of Columbia	1

4.3 EXPERIMENTAL VARIABLES STUDIED

Once counties with social distancing policies were identified, we collected data on other factors that have been shown to correlate with COVID-19 infection rates, such as the socioeconomic status and overall mobility of a county's residents. This data was collected from a variety of sources which we describe below.

Safegraph Social Distancing Dataset. Safegraph – a location data company – temporarily offered a Social Distancing metrics dataset to help researchers better understand how and if people were changing their behavior in response to COVID-19 and associated policies. Safegraph collects foot traffic data from "a panel of GPS pings from anonymous mobile devices," and uses that information to quantify how often people are leaving their homes. In this dataset, home refers to "the common nighttime location of each mobile device over a 6-week period to a Geohash-7 granularity (~153m x ~153m)."[59] That data is collected daily and aggregated at the census block group level by Safegraph.

We used Safegraph's daily census-block-group-level data to create county-level features. Data were aggregated to the county level by taking the median value of the census block groups within a county on a given day. We then computed the 7-day rolling average of this aggregated measure to identify broader trends in the data and reduce the impact of noise on the analysis. Additionally, we wanted to reduce the daily data to a single feature for each county, so that the data could be used in our regression algorithm. To reduce the location data to a single feature in the dataset, we then computed the percent increase between the first week of January and the first week of April for the weekly average of the metrics included in the Safegraph Social Distancing dataset.

New York Times and Dynata Survey. Dynata, a global survey and data firm, conducted a survey of mask use at the behest of the New York Times. The firm obtained 250,000 survey responses between July 2, 2020 and July 14, 2020. Participants were asked to respond to the question, "How often do you wear a mask in public when you expect to be within six feet of another person?" Responses were weighted by age, sex, and location to produce a county-level estimate of how often residents of that county would respond "never", "rarely", "sometimes", "always", or "frequently" to the above question. We then represented this information as a single feature by aggregating the estimate for "always" and "frequently" for each county in the dataset.

Massachusetts Institute of Technology (MIT) Election Data + Science Lab. The MIT Election Data + Science Lab has published county-level presidential results data dating back to 1976. For each county in the study, we collected information on the percent of voters that voted for the Republican candidate in 2016. This data point was intended to serve as a proxy for partisanship in the studied area.

American Community Survey. The American Community Survey is an ongoing survey conducted by the Census Bureau. The survey is designed to capture the social, economic, demographic, and housing characteristics of the U.S. population, and the resulting data is published annually. In addition to publishing the annual results, the Census Bureau also publishes 5-year estimates. The data within those multiyear estimates is much more reliable for less populated areas and small population subgroups. The more current 1-year estimates are only available for areas with populations of 20,000 people or more.

CDC Social Vulnerability Index 2018 dataset. The U.S. Center for Disease Control uses 15 variables from the American Community Survey to identify communities that may need "support before, during, or after public health disasters." These variables capture the degree to

which a community exhibits "certain social conditions, including high poverty, low percentage of vehicle access, or crowded households" that might make a community particularly vulnerable to "human suffering and financial loss in the event of a disaster."[60] The CDC uses these social factors to rank counties and census tracts according to their relative social vulnerability. However, we will only be using the raw data estimates and percentages for these variables rather than the CDC's ranking of each county. This is to ensure that the impact of each variable on the number of COVID-19 infections in an area can be better discerned from one another. Additionally, not every variable from this dataset was used in constructing our model. For a full list of the variables used in the model from this dataset see Figure 4.

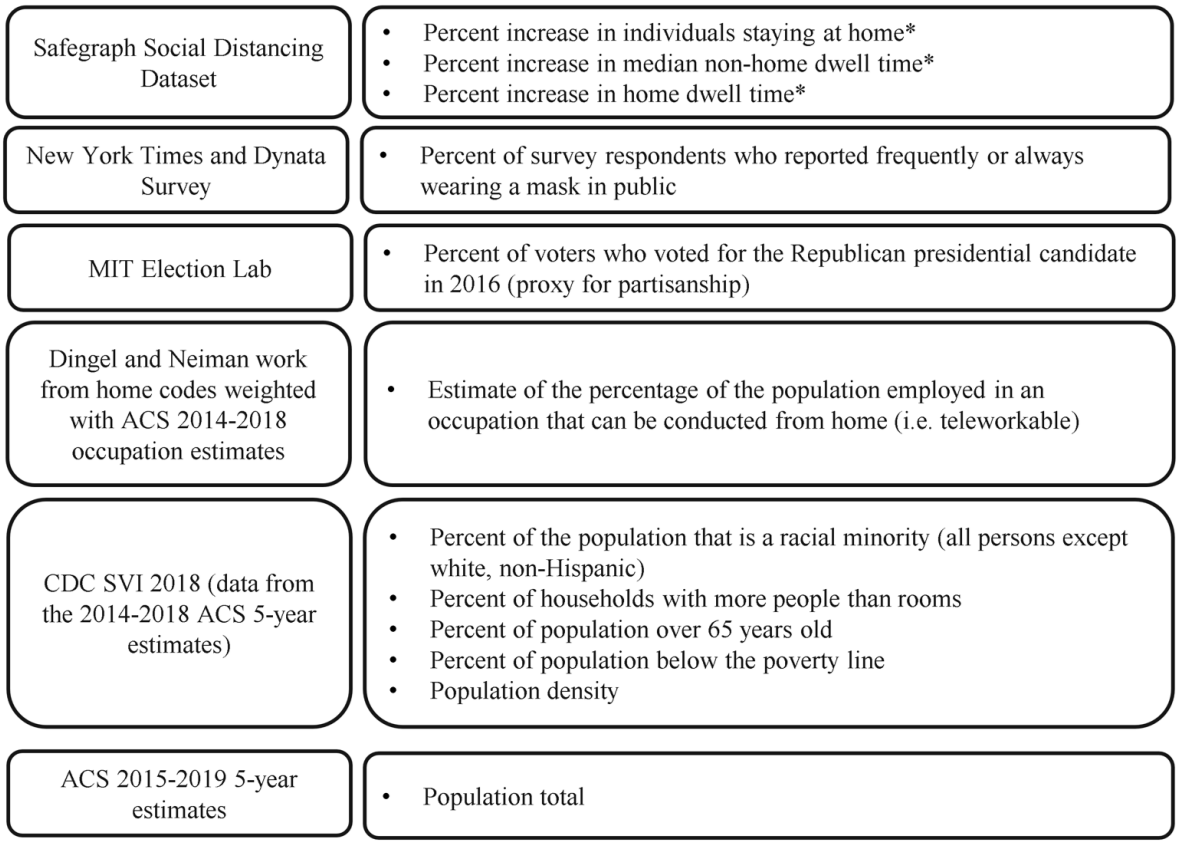

Figure 4. List of the experimental variables used in this study and the corresponding source of information. *Percent increases are measured between the average during the first week of January 2020 and the first week of April 2020.

Estimating essential workers in an area. To examine the relationship between essential workers and COVID-19 caseload, we attempted to estimate what proportion of workers in each county were employed in occupations that could be conducted from home. In a 2020 paper, Dingel and Neiman classify the feasibility of working at home for all occupations, as designated by Standard Occupational Classification (SOC) system, the federal government's standard system for classifying workers by occupation. Dingel and Neiman utilize the responses to two surveys – the Work Context Questionnaire and the Generalized Work Activities Questionnaire – to identify the tasks and activities associated with each occupation. Occupations are then coded as "teleworkable" or not according to the nature of the occupation and th physical and social context in which those tasks are performed.[61] Whether an occupation is teleworkable or not is coded as a binary flag. The occupation-level results of this classification are publicly available.[62]

We then combined these occupation-level classifications with occupation-level employment data from the American Community Survey 2014-2018 5-year estimates. However, a weighting technique had to be used to match the information in both datasets because the information in Dingel and Neiman's dataset is more granular than the information included in the American Community Survey. The SOC codes published by the Bureau of Labor Statistics designate major groups (e.g., Management Occupations) that encompass more detailed occupation designations (e.g., Marketing Managers). The Dingel and Neiman classifications are available for each of the detailed occupation designations, but the ACS only publishes employment data by major occupation group. To accommodate for this difference, for each SOC major group, we computed the proportion of detailed occupations within the broader group that had been designated as teleworkable. Actual numbers of employees in each occupation group were then multiplied by the weights. The sum of those weighted numbers was then divided by

the total number of employees in the county to get an estimate of the proportion of teleworkers in the area.

We then verified that this teleworker estimate correlated with other metrics associated with essential workers. This estimation correlated with the percent increase in residents staying at home all day between the first week of January and the first week of April ($r = .35$, $P = .06$) in the counties studied. The estimate also inversely correlated with both the percentage of the people in the area without a high school diploma ($r = -.56$, $P < .001$) and the percentage of people below the poverty line for all U.S. counties ($r = -.46$, $P < .001$). As previously discussed, frontline workers typically earn lower wages and have less educational attainment than other employees.[9] These strong correlations thus indicate that our estimate is capturing the relative numbers of frontline workers in each county.

Handling missing data. Relevant data could be found for all 360 counties except for one, Rio Arriba, New Mexico. Rio Arriba's occupation data was recorded as "null" in the ACS 2014-2018 5-year estimates, so we were unable to estimate the proportion of people employed in teleworkable occupations in that area. Since this was the only county with missing data, we opted to exclude Rio Arriba from our analysis. Therefore, the final size of the sample was 359 counties.

Policy environment variables. Since we attempted to control for social distancing policies in this study, we did not include any variables related to policy environment in our model except for one. We included the "Start date of the policy environment" as a predictor variable. The pandemic in the U.S. has had distinctive waves with varying case counts, so we assumed that the model would require information on the time period studied. This date refers to the first day on which the county met the policy environment criteria as outlined in Section 4.3.

4.4 MULTICOLLINEARITY

Before attempting to make any prediction based on the collected data, we first had to understand how the predictor variables were related to one another. Many statistical and machine learning techniques perform best when the inputted variables are independent of one another. However, many of the relevant variables that we identified were correlated with one another. Figure 5 shows the Pearson correlation coefficients between each of the variables that we studied within the dataset. Several of these correlations were observed across all counties in the U.S. as well (Figure 6).

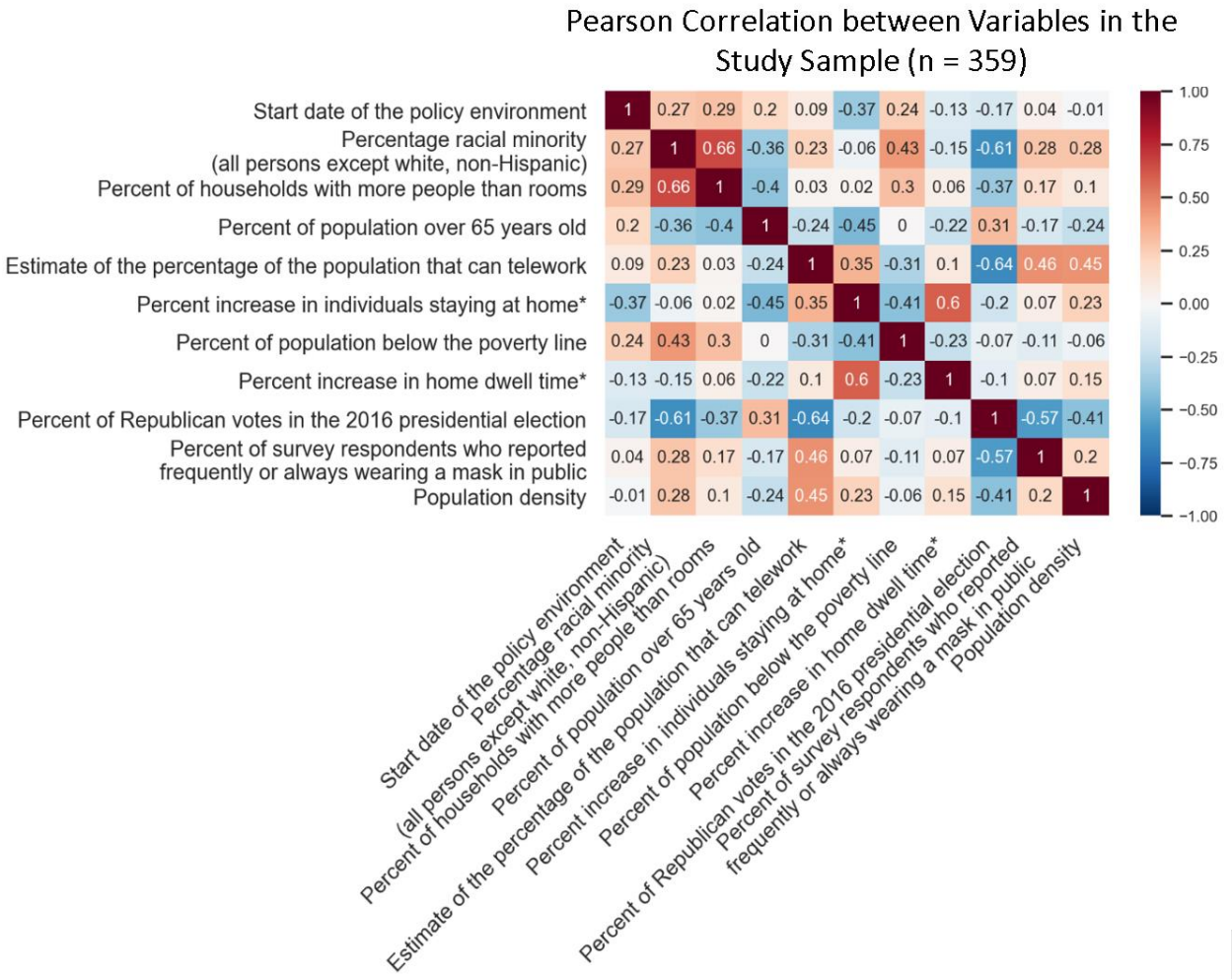

Figure 5 The Pearson correlation coefficient (rounded to 2 decimal places) for each of the variables in the dataset (n = 359). 1.00 indicates a perfect correlation (darker reds) and 1.0 indicates a perfect inverse correlation (darker blues).

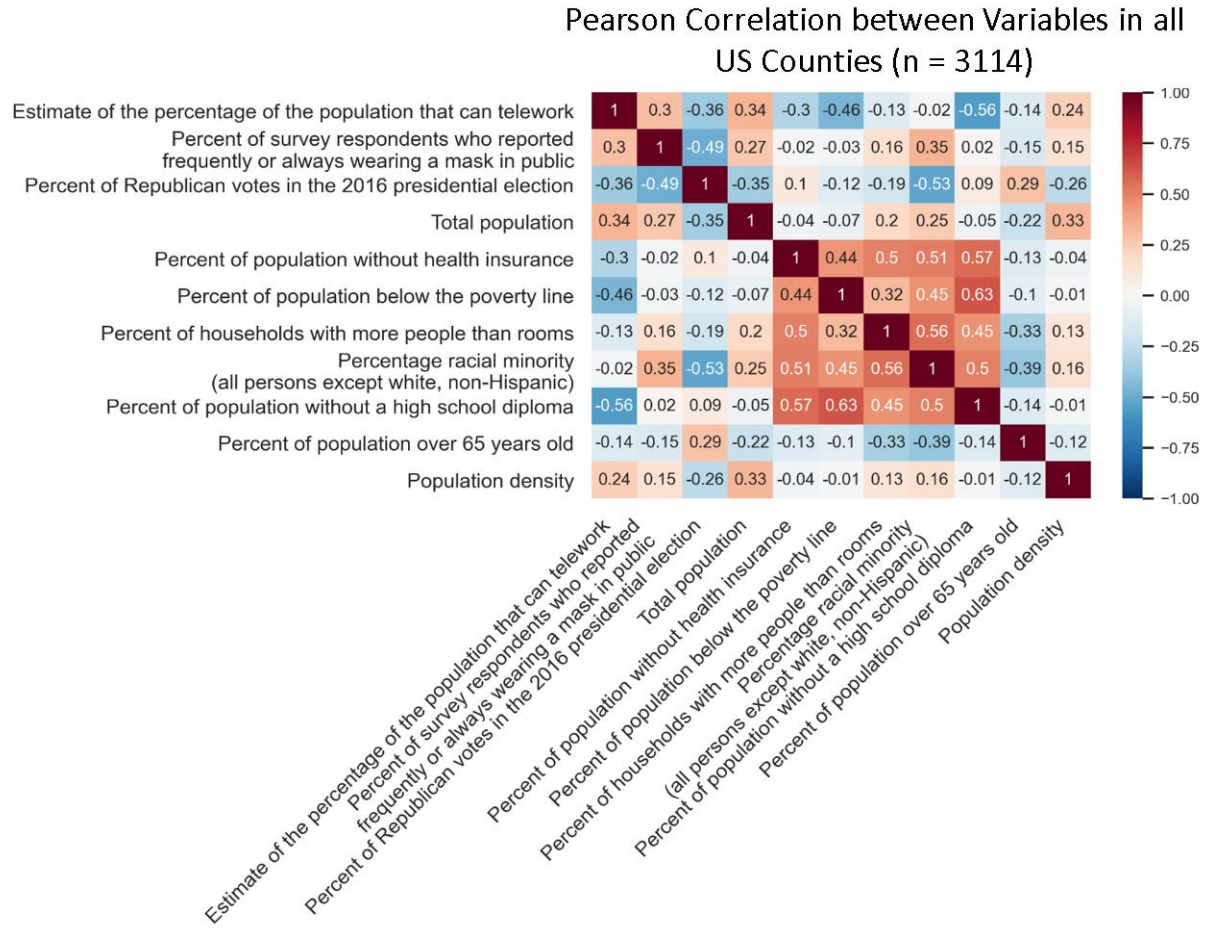

Figure 6. The Pearson correlation coefficient (rounded to 2 decimal places) for each of the counties in the US with available information (n = 3114). 1.00 indicates a perfect correlation (darker reds) and 1.0 indicates a perfect inverse correlation (darker blues).

Rather than only including variables that were independent of another in our analysis, we opted to use a model that would be robust to multicollinearity.[21] To clarify the impact of any one variable, we also computed a simple, unadjusted Pearson correlation coefficient for each of the predictor variables and our outcome variable.

4.5 INTRODUCTION TO RANDOM FORESTS

Random forests are an ensemble machine learning algorithm first introduced by Breiman in 2001.[63] This algorithm can be used on classification and regression tasks. A random forest fits an ensemble of decision trees to various random sub-samples of the training dataset and generates a prediction by committee that is often more accurate than the prediction of any one tree in the forest. In the classification task, a majority voting system is used where the final prediction is the most common class prediction of the decision trees. In the regression task, the prediction of each tree is averaged to produce the final prediction (see Figure 7).

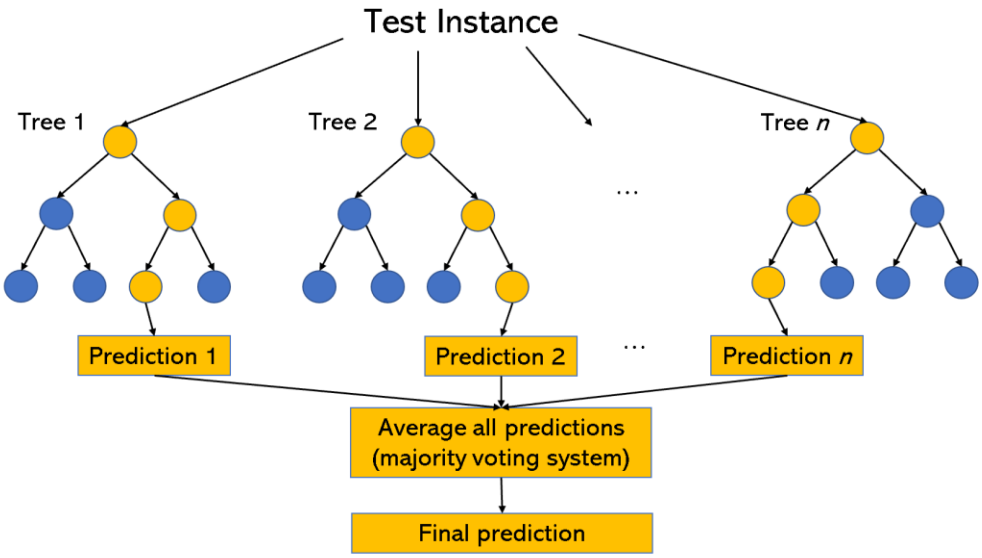

Figure 7. Diagram of a random forest model with *n* estimators completing a regression task.

A simple decision tree model learns to predict a target value by continuously splitting the data at each node in the tree. At each split, the tree iterates through each of the predictor variables to determine which variable most reduces the split-criterion, in this case, the mean-squared error (MSE) (see Figure 8). It then selects the variable associated with the lowest split-criterion and splits the sample according to that variable. Building the decision tree is a greedy

process since at each split, the algorithm makes the most optimal choice to reduce the error function. Due to this greedy strategy for selecting features, trees are especially robust to multicollinearity.[21]

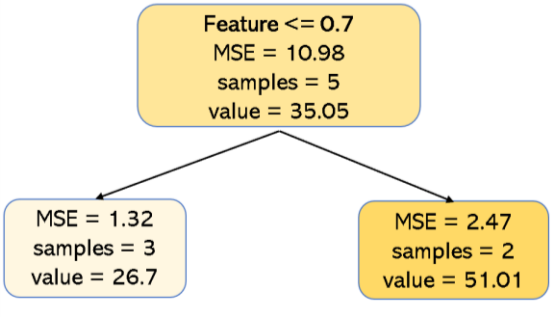

Figure 8. A decision tree with a single node. This decision tree splits the 5 samples into a set of 3 samples and a set of 2 samples based on the value of a feature in each sample. This reduces the MSE from 10.98 in the larger sample to 1.32 and 2.47 in the 2 smaller samples. Instances with a Feature <= 0.7 are associated with lower predicted values than those with a feature above 0.7.

Decision trees can also capture complex structures in data and, if grown sufficiently deep, have a relatively low bias as well, though they are prone to overfitting the training set. Overfitting occurs when the model learns the variance and the noise in the training set to such an extent that the model no longer performs well on data not included in the training set. However, the possibility of the model overfitting the dataset is reduced in a random forest model through a bootstrap process.

In bootstrap aggregation, each model within an ensemble machine learning algorithm is given a random selection (with replacement) of the examples in the training set. The prediction of each component is then averaged to create the final prediction. Since trees are notoriously noisy structures, they benefit greatly from this averaging. Thus, in a random forest, no single tree is built using the entirety of the training data. To ensure that the individual trees are uncorrelated, the features are randomly permutated at each split, so that the best split found may vary, even with the same training data. Breiman originally recommended considering no more than $\left\lfloor \frac{m}{3} \right\rfloor$

features at each split where m is the total number of features in the set. However, we set the model to consider up to m features at each split, which was justified empirically in a more recent paper. That paper found that the error decreases monotonically as the number of features considered increases.[64] The accuracy of the random forest depends both on the strength of each individual tree in the model and on the correlation between any two trees in the dataset.

From the random forest, we can obtain a measure of each feature's "importance" in computing the final prediction as well. At each split in each decision tree, the improvement in the split-criterion – in this case, the reduction in the mean squared error or variance – is the importance measure attributed to the splitting variable. That measure of importance is accumulated over all trees separately for each variable. Because of the random permutation of features selected at each split in the tree, each of the relevant variables has a chance to be the primary split. The ensemble averaging also reduces the contribution of any individual variable. Sci-kit Learn, a popular Python library for machine learning, automatically provides a normalized version of this measure of feature importance.[65] Each variable is assigned an importance measure between 0 and 1 according to their total contribution to reducing the split-criterion. However, random forests can be biased in favor of features with higher cardinalities that may have a one-to-one relationship with the target value. This bias is often an issue when trees are built using both categorical and continuous data, but our model is built using only continuous data. We have also included the cardinality of each of the features in the appendix to ensure that we and the reader can understand the potential for bias in the feature importance computation. Despite this one limitation, we can still use measures of feature importance to understand which of our inputted variables (percent below the poverty line, percent racial

minority, etc.) were most responsible for reducing the mean squared error in the model's COVID-19 caseload prediction.

Additionally, random forests provide a simple way to evaluate how the model will perform on unseen data, known as the out-of-bag (OOB) error estimate. To obtain the OOB error estimate, for each observation z_i we construct a random forest predictor averaging only those trees built from bootstrap samples in which z_i did not appear.[66] Thus, as we build the model, we can simultaneously estimate how well a subset of the component trees performs on unseen observations. Therefore, trees can be fit in one sequence with cross-validation being performed along the way.[67] Sci-kit Learn also automatically provides an OOB error estimate as a part of its RandomForestRegressor class.

We can also estimate the model's performance on unseen data using a slightly more common statistic, the coefficient of determination or R-squared. R-squared measures the proportion of the total variation in the target variable that is captured by the model.[30] To measure how well the model is fitting the training data, we compute the R-squared value on the training set. To measure how well the model will perform on unseen data, we compute the R-squared on the test set. The county-level dataset was split into a train and test set with 70% of the data randomly selected to be part of the training set and 30% selected to be a part of the testing set.

4.6 OUTCOMES STUDIED

For our target value, we computed the average new daily COVID-19 cases observed per 100,000 people in each county studied. A random forest model is only capable of predicting a single value, so needed to condense information about the epidemic trajectory during the policy implementation into a single number. Data on confirmed COVID-19 cases at the county level

was obtained from the publicly available New York Times COVID-19 data repository.[54] We obtained estimates of each county's population from the ACS 2019-2020 5-year estimates. We only considered caseload data from the period in which the policy environment met the criteria outlined in Section 4.3 in our computation. To create this target value, we used the following formula.

$$\text{Average New Cases} = \frac{Cases\ Recorded\ on\ Start\ date\ -\ Cases\ Recorded\ 14\ Days\ after\ End\ Date}{Number\ of\ Days\ the\ Policies\ Were\ in\ Effect\ +\ 14\ days}$$

We took several steps to ensure that this single value would accurately reflect relevant information about the county's epidemic curve. A 14-day lag time was added to the calculation to ensure that all COVID-19 infections that occurred during the period of interest were captured in the data. The incubation period for COVID-19 is thought to extend to 14 days, with a median time of 4-5 days between initial exposure and symptom onset.[68] Increasing the time frame by 14 days also increased the amount of data used to construct the feature and diminished the impact of daily outliers on this calculation. Therefore, despite the limitation of our model only being able to predict a single data point, we constructed a measure that encapsulated a large amount of information about the epidemic trajectory. Figure 9 depicts the average new COVID-19 cases for each of the counties studied.

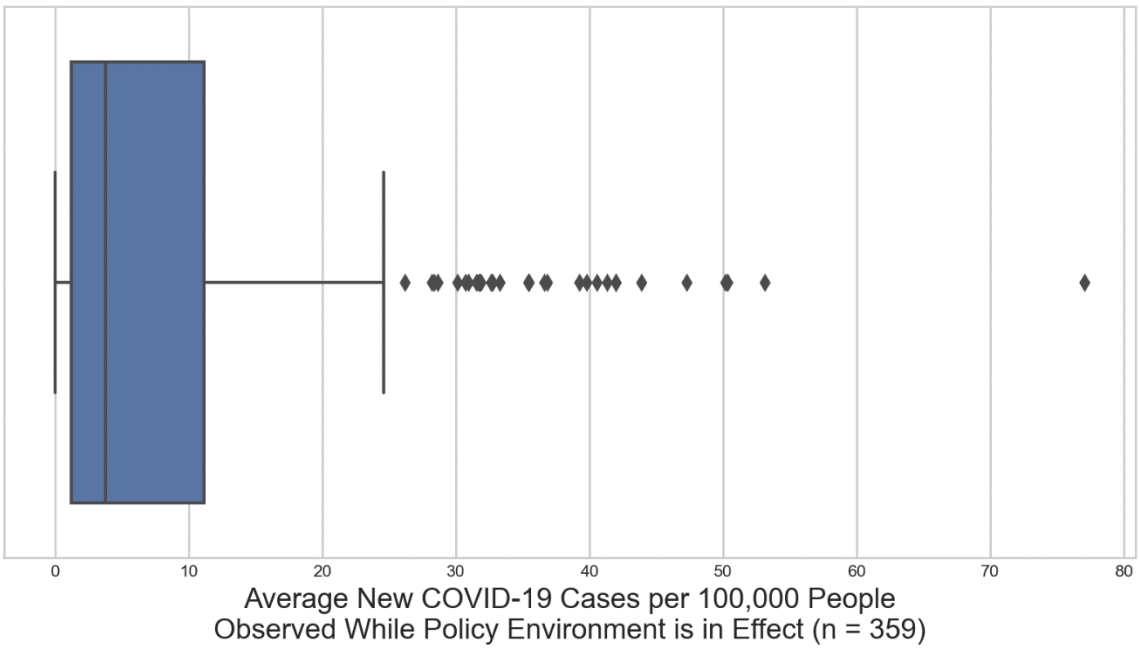

Figure 9. The calculated average new COVID-19 cases per 100,000 people observed during the days that the policy environment of interest in all the counties in the dataset.

5. RESULTS

5.1 PRELIMINARY STATISTICS

Before constructing a model from the data, we computed descriptive statistics for the inputted variables using SciPy's Pearson correlation coefficient. We found that only the following three features had a statistically significant relationship with average new reported COVID-19 cases: the start date of the policy environment (P = .003), the percent of the population that identified as not white, non-Hispanic (P = .007), and the percent of households with more people than rooms (P = .002). Figure 10 depicts the strength and direction of each relationship.

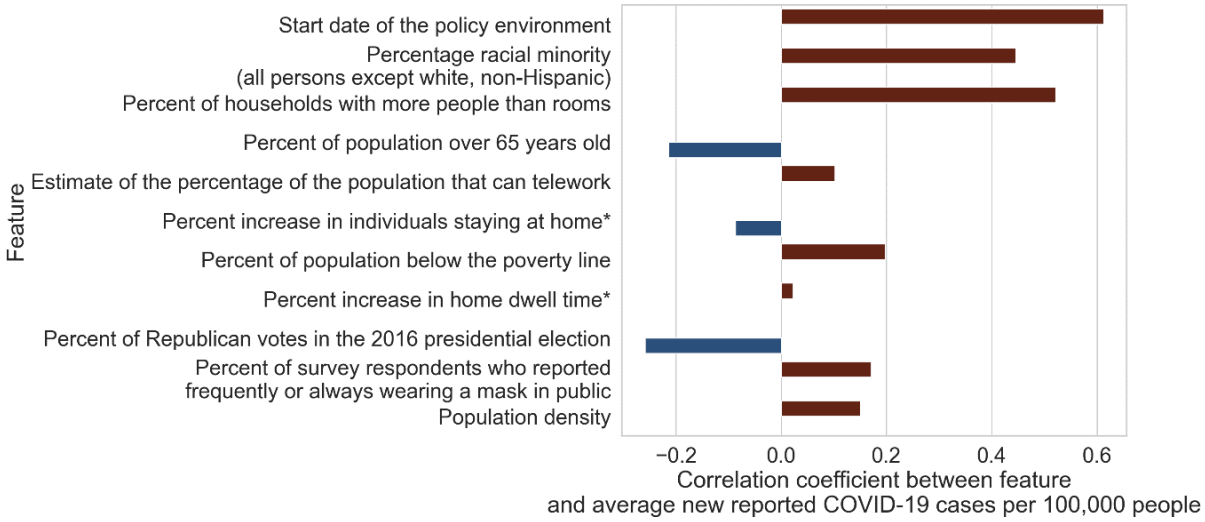

Figure 10. The Pearson correlation coefficient between each of the features in the dataset and the average new COVID-19 cases per 100,000 people. *Percent increase is measured from the first week of January to the first week of April.

5.2 TRAINING THE MODEL ON THE ENTIRE DATASET

Initially, the model was fitted to the entire dataset which included data from the spring, summer, and fall. If a county met the policy criteria during multiple periods between March 2020 and January 2021, only data from the earlier period was included in the model. This was done to prevent any county from being included in the dataset twice and to prevent the model from overfitting. The model could explain approximately 63% of the variance in the average new COVID-19 cases per 100,000 people using those predictor variables. It had an approximate root mean squared error (RMSE) of 6.99 cases per 100,000 people when predicting the average new COVID-19 cases observed in the test data (Table 2).

Table 2. Measures that represent both how much of the variation that the model can account for (Model Evaluation Scores) and the accuracy of the model's prediction on the test dataset (N = 108).

Model Evaluation Scores			**Model Error**		
Training R^2	OOB score	Test R^2	MAE	MSE	RMSE
0.95	0.64	0.60	4.75	48.88	6.99

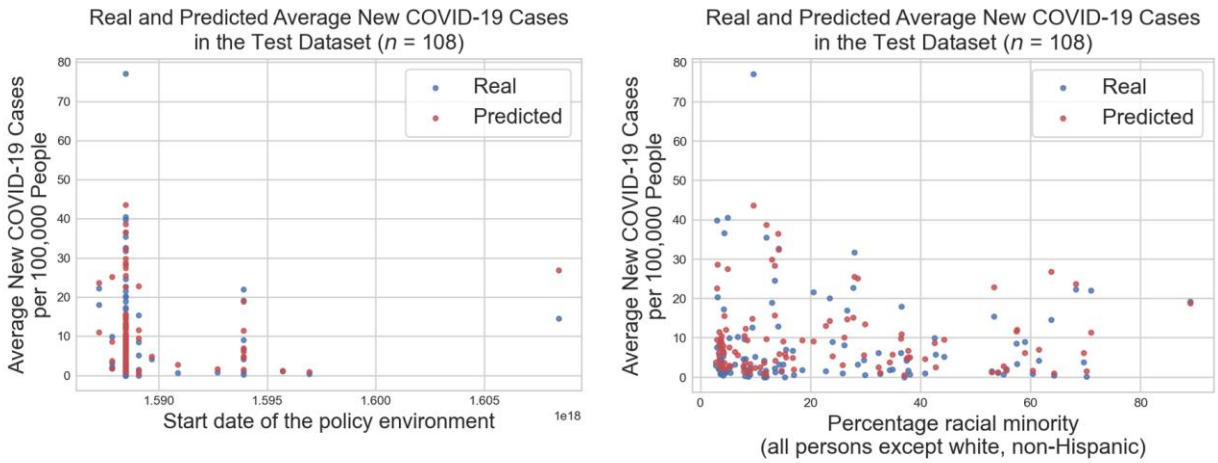

Figure 11. The reported and predicted values of the average new COVID-19 cases per 100,000 people relative to two of the features, the start date of the policy environment and the percent of the population belonging to a racial minority, for each of the counties in the test dataset (n = 108).

In this dataset, the start date of the policy environment was the most important predictor of the average new COVID-19 cases per 100,000 people. The percent of people in the area belonging to a racial minority and the percentage of households with more people than rooms were also important variables. Figure 12 depicts the ranked importance of each of the features inputted into the model.

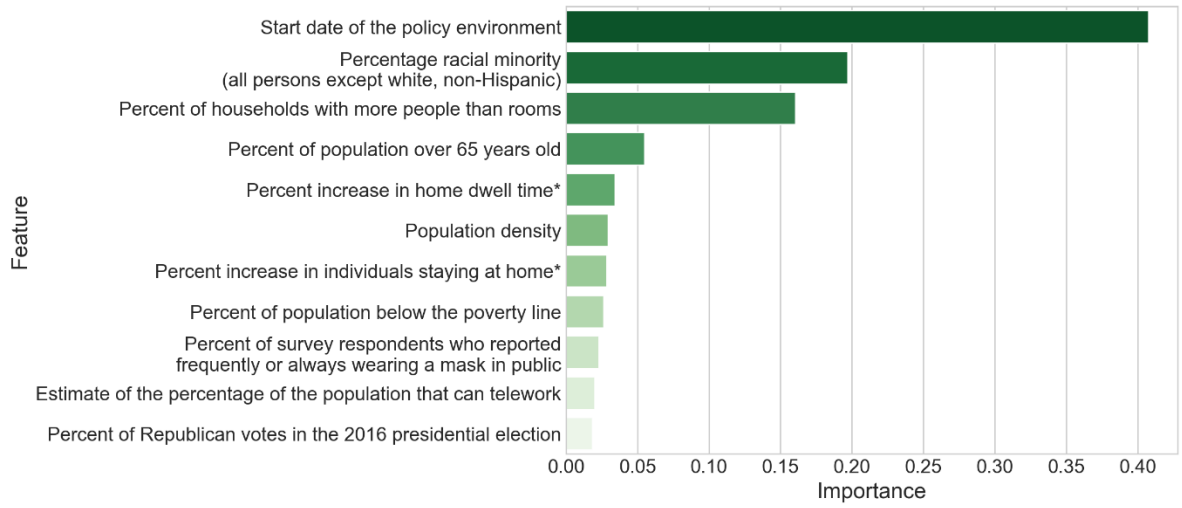

Figure 12. The relative importance of each of the features in explaining the variance in the dataset and reducing the mean squared error. *Percent increase is measured from the first week of January to the first week of April.

5.3 DISCUSSION OF THE SIGNIFICANCE OF START DATE

Around mid-November, there was a substantial increase in new cases per 100,000 people across the U.S. Public health officials attributed this later fall surge to the increase in indoor activities and the large number of people traveling for the holidays observed during the fall and winter.[69] This trend is reflected in many - though not all of - the counties included in this study. Figure 13 depicts the average new COVID-19 cases per 100,000 people for the counties included in this study. Figure 14 shows a representation of one of the estimators within the model. Note how the decision tree has learned that the number of cases observed after August 8th varies significantly from the number of cases observed during earlier phases of the pandemic.

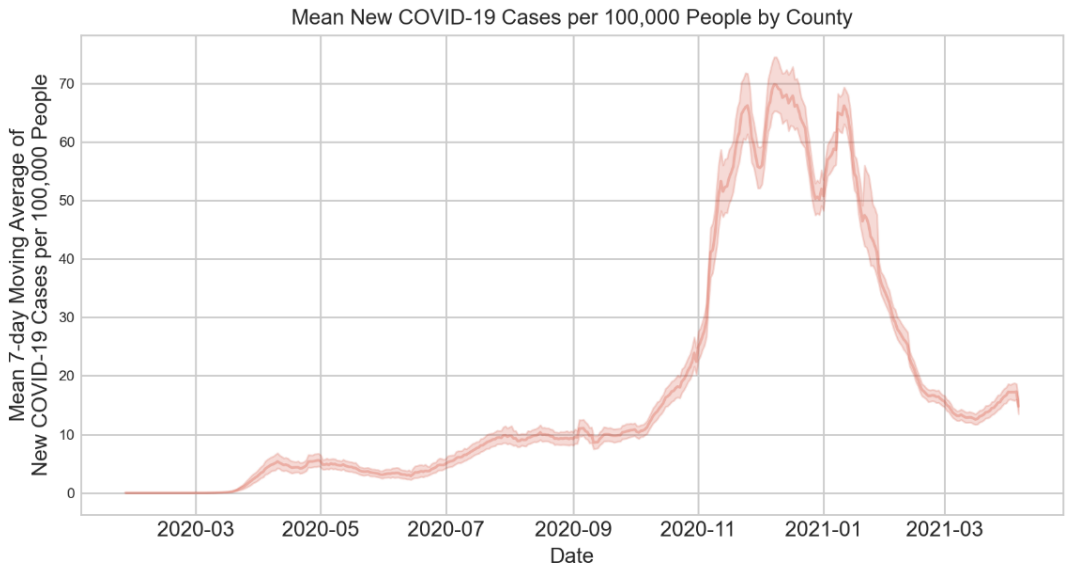

Figure 13. The mean of the 7-day moving average of new COVID-19 cases per 100,000 people for each of the counties included in the study (n = 359). The shaded region depicts the 95% confidence intervals.

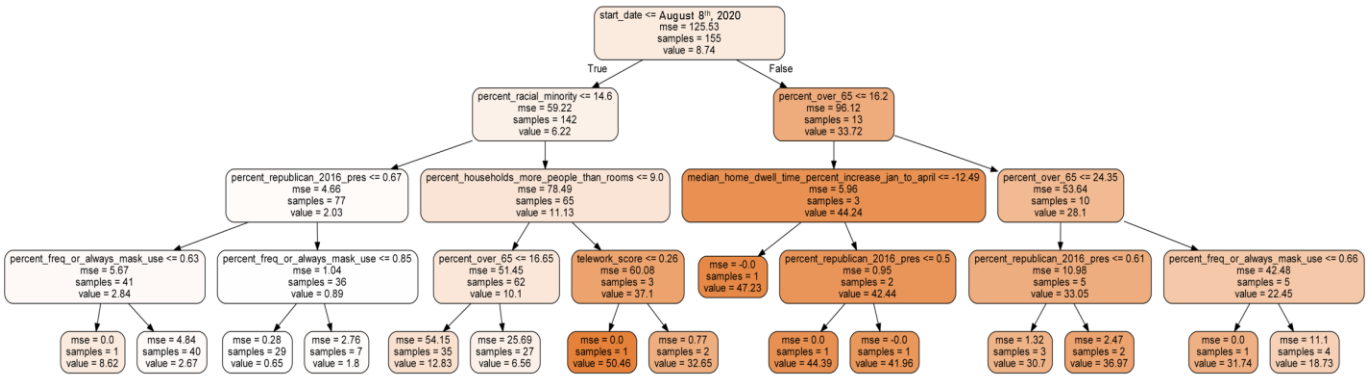

Figure 14. A sample estimator with maximum depth restricted to 5. Note that in the actual model, maximum depth was restricted to 70 and 1000 estimators were used. Additionally, the start date was converted from an integer representation to an approximate day to improve the interpretability of this diagram.

5.4 SPLITTING THE DATASET

To verify our assumption that the start date was only important for initially splitting the dataset according to the phase of the pandemic, we manually split the dataset and trained separate models on each subset of the data. We divided the counties into a dataset with start dates before August 8th, 2020 and another dataset with start dates after August 8th, 2020. Since some counties met the policy criteria during multiple periods, those counties were now included in both datasets. The spring and summer sample included 327 counties. The fall dataset included only 115 counties.

5.5 TRAINING THE MODEL ON THE SPRING AND SUMMER DATA

Once again, we conducted a preliminary correlation analysis before constructing the random forest model to better understand the relationships between the inputted variables and the number of new COVID-19 cases (Figure 15). Notably, there was no longer a statistically significant relationship between the average number of new COVID-19 cases and the start date

of the policy environment (P = .068). However, the positive relationship between the percent of the population belonging to a racial minority and the average number of daily new COVID-19 cases was even more significant in this new dataset (P < .001) as was the positive relationship between the percent of households with more people than rooms and the number of new COVID-19 cases (P < .001). In the spring dataset, we also found two other statistically significant relationships. The percent of people over 65 and the percent of Republican votes in the 2016 election had a significant inverse correlation with the average new COVID-19 cases in the spring and summer dataset (r = -.40, P = .007, r = -.36, P = .01 respectively). We also know that these two features also had a positive correlation with one another from our previous analysis (r = .31, P = .02).

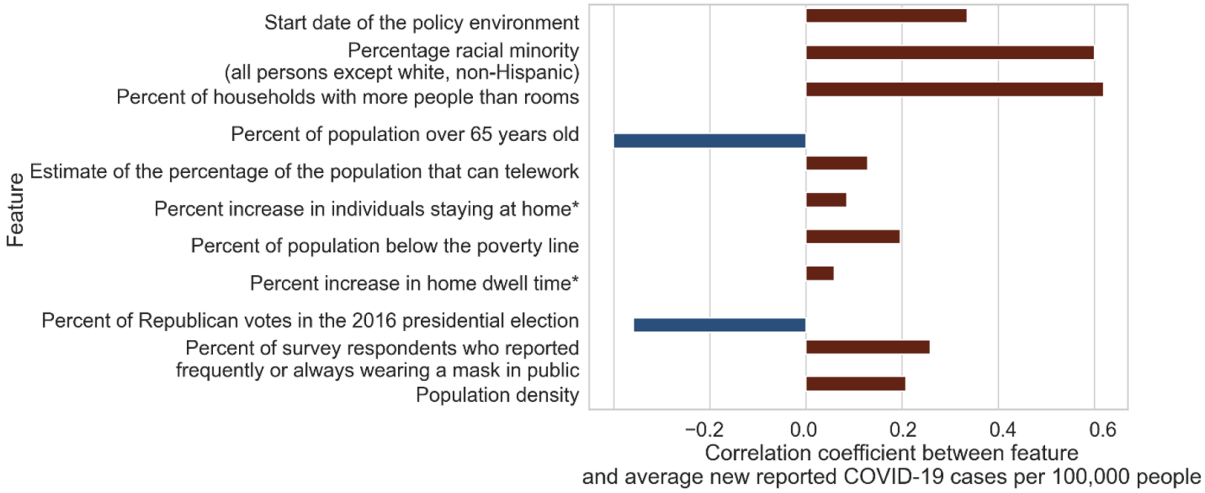

Figure 15. The Pearson correlation coefficient between each of the features in the dataset and the average new COVID-19 cases per 100,000 people during the spring and summer. *Percent increase is measured from the first week of January to the first week of April.

When we constructed the random forest model using only data from the spring and summer, the model was able to explain less of the variance in the data but had a lower mean-squared error (MSE) on the test set. A reduction in the variance explained was expected as we had previously learned that the start date was accounting for a significant amount of the variance

within the data. However, the random forest model was still able to explain 41% of the variance in COVID-19 cases observed (Table 3).

Table 3. Measures that represent both how much of the variation that the model can account for (Model Evaluation Scores) and the accuracy of the model's prediction on the test dataset (Model Error) (n = 99). Note that all values are rounded to 2 decimal places.

Model Evaluation Scores			**Model Error**		
Training R^2	OOB score	Test R^2	MAE	MSE	RMSE
.92	.48	.41	3.78	35.63	5.97

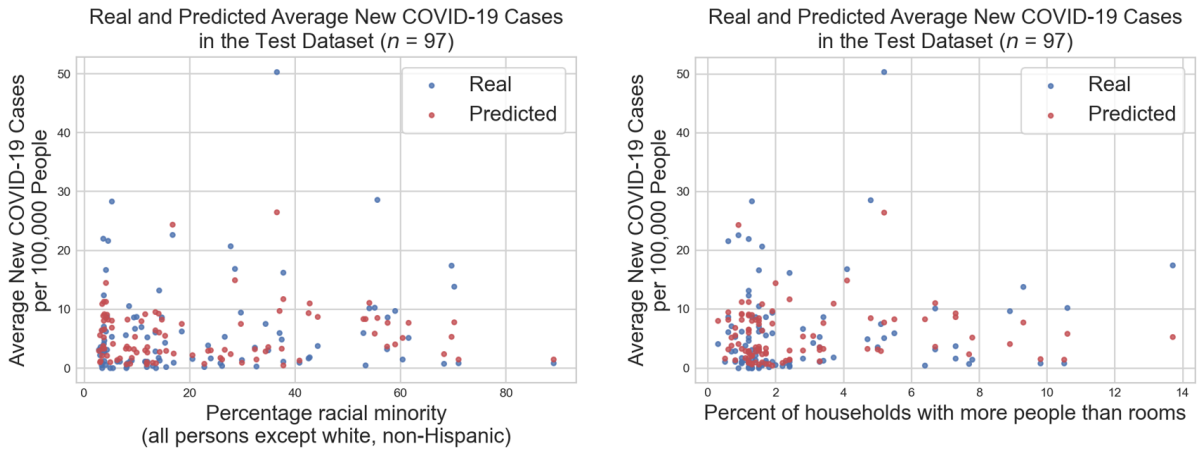

Figure 16. The reported and predicted values of the average new COVID-19 cases per 100,000 people relative to two of the features, the percent of the population belonging to a racial minority and the percent of households with more people than rooms, for each of the counties in the test dataset (n = 99).

Now that the dataset only contained information on counties that implemented policies before August 8th, the start date became the least important feature in the dataset. Of course, the start date may have also been recorded as the least important feature because random forests are biased towards high cardinality features. Start-date has the fewest unique values of any of the features in the dataset (see Appendix).

However, the percent of people identifying as a racial minority and the percent of households with more people than rooms remained the most important features within the

dataset. The percent of the population over 65 years old and the percent increase in individuals staying at home were also important to reducing variance in the model.

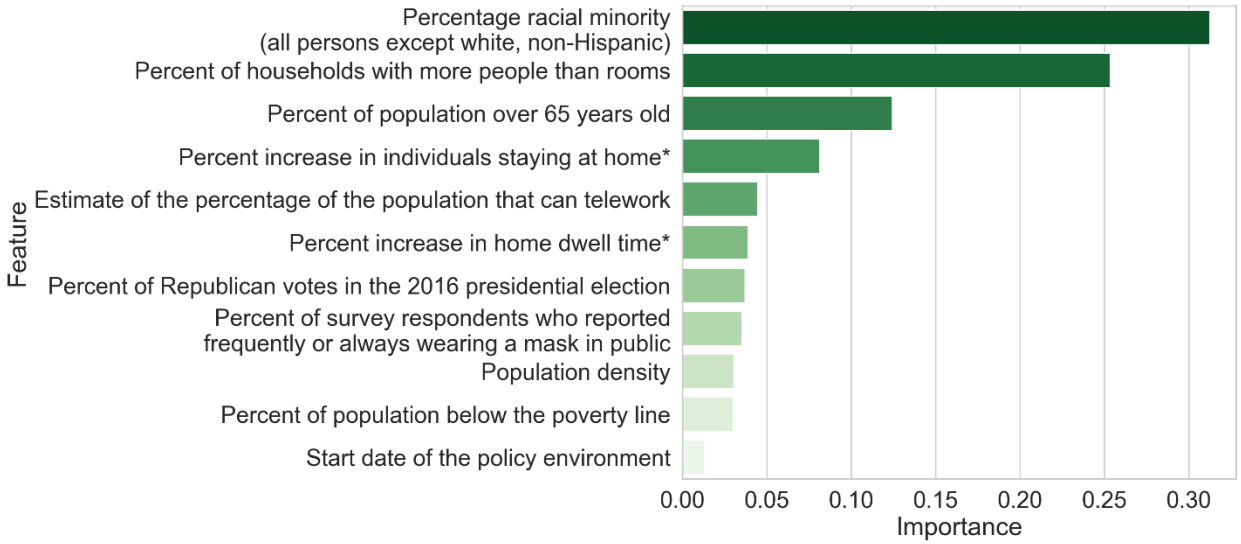

Figure 17. The relative importance of each of the features in explaining the variance in the average new COVID-19 cases in the spring and summer dataset and reducing the mean squared error. *Percent increase is measured from the first week of January to the first week of April.

5.6 DISCUSSION OF THE RESULTS

We found that the percent of the population identifying as a racial minority and the percent of households in the area with more people than rooms were the two most important variables in determining the model's prediction of the average number of COVID-19 cases per day observed during the spring and summer when social distancing policy was accounted for. This finding aligns with previous evidence that more urban, diverse areas bore the COVID-19 caseload burden at earlier stages of the pandemic.[70] The results also confirm previous analysis suggesting that race and household composition are significant determinants of an individual's likelihood to be infected by COVID-19.[48]

Percent racial minority being the most important predictive feature aligns with our understanding of race and social determinants of health. Social determinants of health are

defined as “factors apart from medical care that can be influenced by social policies and shape health in powerful ways.”[40] Socioeconomic factors such as an individual’s income, education, and wealth can be powerful determinants of that individual's health. During the COVID-19 pandemic, people with lower incomes and lower educational attainment were more likely to work in frontline positions where they were at increased risk of being exposed to COVID-19.[9] Black and Hispanic individuals are disproportionately represented in these low-income positions. There are also significantly more likely to live in crowded, multi-generational households.[71] Our own analysis found a strong correlation between the percent of people belonging to a racial minority and households with more people than rooms.[71] When members of these households become infected at work, they bring the virus home to vulnerable family members that they cannot effectively isolate from. These results suggest that when social distancing policies are in effect, the virus continues to spread amongst people who cannot effectively obey social distancing orders because they work in frontline industries and/or live in crowded households.

Some people may erroneously suggest that perhaps race is such an important predictive factor because certain racial groups are less willing to socially distance than others. USA Today reported that “The message of social distancing doesn’t seem to be hitting home, with people still playing basketball, having card parties and hosting sleepovers, say black mayors.” Yet, such messaging ignores the reality of racial discrimination and segregation in this country. Black and Hispanic Americans are not willingly putting themselves into situations in which they are exposed to COVID-19, they have been forced into those situations by the lingering impacts of legal segregation and modern discrimination. Black and Latino Americans are overrepresented in industries that cannot be conducted from home and are more vulnerable to being laid off due to a persistent racial gap in both wealth and income. Thus, we’ve found that in the spring and

summer, the inability of people of color to effectively social distance due to structural inequalities shaped the epidemic trajectory in areas that combatted the virus with social distancing policies.

5.7 TRAINING THE MODEL ON THE FALL DATASET

We then ran the model on the fall dataset to see if these same relationships would hold. The OOB error estimate suggests that the model would be unable to account for the variance in an unseen dataset. Despite the relatively high R-squared value ($R^2 = .28$) on the fall test set ($n = 34$), this OOB error indicates that the model might have been overfitted to this smaller subsample of counties (see Table 4).

Table 4. Measures that represent both how much of the variation that the model can account for (Model Evaluation Scores) and the accuracy of the model's prediction on the test dataset (*n* = 36) (Model Error). Note that all values are rounded to 2 decimal places.

Model Evaluation Scores			**Model Error**		
Training R^2	OOB score	Test R^2	MAE	MSE	RMSE
.86	-0.04	0.28	17.28	441.586	21.01

Additionally, we observed that the ranking of feature importance measures in the fall model differed substantially from our spring and summer model. Now behavioral factors, such as the percent increase in home dwell time, partisan preference, and likelihood of wearing a mask in public, were identified as the most important features in the model (Figure 18).

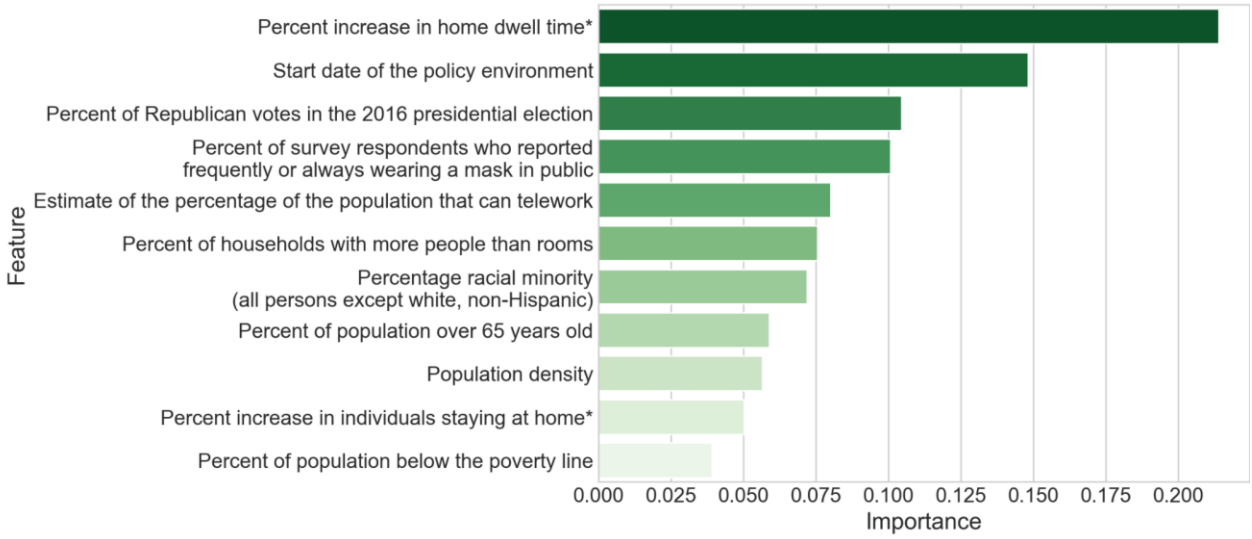

Figure 18. The relative importance of each of the features in explaining the variance in the average new COVID-19 cases in the fall and winter dataset. *Percent increase is measured from the first week of January to the first week of April.

The above results are provided to suggest avenues for future analysis. These three features would have been the most out-of-date at the time we were measuring COVID-19 cases. The mask survey information was obtained in July. The percent increase in metrics of mobility was measured from January to April. Additionally, we used data from the 2016 presidential election because the more recent voting data was not available at the start of this study. Perhaps the accuracy of the model was limited by the outdated nature of its most important features, or perhaps we just needed a larger fall dataset to draw meaningful conclusions about the relationship between the features and the confirmed COVID-19 cases.

5.8 LIMITATIONS

Other limitations should be noted as well. Controlling for an area's COVID-19 mitigation policy is both difficult and inexact. We did not attempt to control for any policies recorded in COVID AMP that were unrelated to social distancing or mask mandates. These other kinds of policies, like "Travel restrictions," "Enabling and relief measures," or "Contact tracing," may

have a significant impact on COVID-19 restrictions but restricting the policy criteria further would have caused a smaller sample size. With a smaller training size, the model may not have been able to successfully identify patterns in the dataset. Another limitation of this work is that using a random forest model necessitates reducing complex time-series data to a single feature. We had to transform every measure of mobility into a single data point, but future analysis may want to consider multiple features related to mobility measured at multiple months during the pandemic. Also, while we did not remove variables that were correlated, we did try to avoid using mutually exclusive variables. For instance, rather than including multiple features representing different racial or ethnic groups, we elected to use a single feature that represented the proportion of the population that is not white, non-Hispanic. The US Census Bureau records racial groups as mutually exclusive categories, so knowing how many people belong to one racial or ethnic group also lets a researcher know how many people do not belong in the other racial categories. That might cause certain features to be treated as more important because they contain information about another more relevant feature. However, as evidence emerges that COVID-19 has disproportionately impacted Black and Hispanic Americans more than other groups, a future analysis may want to consider analyzing just those two categories rather than using the broader variable used in this study. Future analysis should consider these limitations and build on the study design presented in this paper.

6. CONCLUSION

When social distancing policy is held as a control, the spread of COVID-19 in the spring and summer at the county level is most significantly impacted by the racial composition of the county as well as the crowdedness of households in the area. We have thus substantiated

previous evidence suggesting a strong relationship between both race and household crowdedness and the spread of COVID-19 during the first waves of the pandemic. We also observed that in the counties studied, those variables may have been more important predictors of COVID-19 caseload burden than other socioeconomic and social factors such as income, partisan preference, etc. Those same relationships were not found in our analysis of the fall, suggesting that our dataset was perhaps insufficient to train a representative model or that such relationships were not present. Despite the inconclusiveness of our fall analysis, we have still identified important factors that impact the spread of the novel coronavirus.

Our study suggests that when social distancing policies are implemented, the virus primarily continues to spread through the households of marginalized individuals. In the spring, when non-essential businesses closed and the general population was told to stay inside, COVID-19 continued to circulate amongst people who could not effectively social distance due to the size of their household and/or the nature of their job. The pandemic and our policy response to the pandemic thus exposed and heightened existing structural inequalities.

The most explanatory variables in our spring model were the crowdedness of households and the percentage of people in the area that identify as a racial minority. The first feature speaks to an issue that policymakers have understood but perhaps not responded to effectively. Lt. Gov. Garlin Gilchrist II, the head of the Michigan Coronavirus Task Force on Racial Disparities, argued "You can't isolate if you live in a home with one room or one bedroom or one bathroom."[72] Dr. Mary Bassett, the Director of the FXB Center for Health and Human Rights at Harvard University, commented that the focus on comorbidities in discussing the COVID-19 racial gap "makes me angry, because this is really is about who still has to leave their home to work, who has to leave a crowded apartment, get on crowded transport, and go to a crowded

workplace, and we just haven't acknowledged that those of us who have the privilege of continuing to work from our homes aren't facing those risks."[8] The results of our study align with these comments by prominent public health officials. In areas where government officials had multiple mitigation policies in place, race and household crowdedness explained why some areas still observed more COVID-19 cases than others. While more privileged individuals were protected by policy, marginalized people were still at risk.

In future pandemics, we cannot repeat these same mistakes and enact policies that fail to protect people who are already marginalized. Truly tackling these issues may require major reforms, but several groups have already proposed policy solutions that could protect vulnerable individuals in the short term. The Brookings Institute called for "enrolling all uninsured frontline essential workers and their families in a new "Medicare COVID" program that would cover all testing, treatment, and vaccinations related to COVID-19."[73] Zhang et al. have even confirmed that such a policy may be effective. Zhang et al. found that "the COVID-19 daily growth rate declined in states with a higher percentage of essential workers enrolled in Medicaid (std. coeff. = −0.04, $p < 0.05$) and a higher percentage of low-income essential workers enrolled in Medicaid (std. coeff. = −0.06, $p < 0.01$)." [3] Other relief measures, such as hazard pay, may help protect vulnerable populations as well. Currently, unpublished analysis of the COVID AMP database found that states that simultaneously enacted enabling and relief measures and social distancing policies saw fewer COVID-19 cases than those states that only enacted social distancing policies. Additionally, the United Nations High Commission on Human Rights has identified several promising practices worldwide that address the disproportionate impact of COVID-19 on minority populations. Some of the identified practices include "allocating aid in Greece, urgent measures for food solidarity in Italy, access to public services in Portugal, actions on social

services in Spain, [and] providing protection measures in the UK."[3] In conclusion, while social distancing is effective, our policy response to pandemic influenza in the United States cannot only involve the enactment of social distancing measures. We must also consider how to best ensure that historically marginalized groups are not disproportionately exposed to and subsequently infected by the virus.

REFERENCES

1. Spiegel MI, Tookes H. Business Restrictions and COVID Fatalities. *SSRN Electron J*. Published online November 11, 2020. doi:10.2139/ssrn.3725015
2. Pak A, McBryde E, Adegboye OA. Does High Public Trust Amplify Compliance with Stringent COVID-19 Government Health Guidelines? A Multi-country Analysis Using Data from 102,627 Individuals. *Risk Manag Healthc Policy*. 2021;Volume 14:293-302. doi:10.2147/RMHP.S278774
3. Zhang X, Warner ME. Covid-19 policy differences across us states: Shutdowns, reopening, and mask mandates. *Int J Environ Res Public Health*. 2020;17(24):1-17. doi:10.3390/ijerph17249520
4. Killeen GF, Kiware SS. Why lockdown? Why national unity? Why global solidarity? Simplified arithmetic tools for decision-makers, health professionals, journalists and the general public to explore containment options for the 2019 novel coronavirus. *Infect Dis Model*. 2020;5:442-458. doi:10.1016/j.idm.2020.06.006
5. Ferguson NM, Laydon D, Nedjati-Gilani G, et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. *ImperialAcUk*. 2020;(March):3-20. doi:10.25561/77482
6. Kaur S, Bherwani H, Gulia S, Vijay R, Kumar R. Understanding COVID-19 transmission, health impacts and mitigation: timely social distancing is the key Change point. doi:10.1007/s10668-020-00884-x
7. Thakkar N, Zimmermann M, Burstein R, Wenger E, Famulare M. *Comparing COVID-19 Dynamics in King and Yakima Counties What Do We Already Know?*
8. Richard OJ, Gebeloff R, Lai KKR, Wright W, Smith M. The Fullest Look Yet at the

Racial Inequity of Coronavirus - The New York Times. The New York Times. Published 2020. Accessed February 26, 2021. https://www.nytimes.com/interactive/2020/07/05/us/coronavirus-latinos-african-americans-cdc-data.html?action=click&module=RelatedLinks&pgtype=Article

9. Blau FD, Koebe J, Meyerhofer PA. Essential and Frontline Workers in the COVID-19 Crisis | Econofact. ECONOFACT. Published 2020. Accessed March 14, 2021. https://econofact.org/essential-and-frontline-workers-in-the-covid-19-crisis
10. Harper CA, Satchell LP, Fido D, Latzman RD. Functional Fear Predicts Public Health Compliance in the COVID-19 Pandemic. *Int J Ment Health Addict*. Published online April 27, 2020:1-14. doi:10.1007/s11469-020-00281-5
11. Coroiu A, Moran C, Campbell T, Geller AC. Barriers and facilitators of adherence to social distancing recommendations during COVID-19 among a large international sample of adults. Capraro V, ed. *PLoS One*. 2020;15(10):e0239795. doi:10.1371/journal.pone.0239795
12. Markel H, Lipman HB, Navarro JA, et al. Nonpharmaceutical interventions implemented by US cities during the 1918-1919 influenza pandemic. *J Am Med Assoc*. 2007;298(6):644-654. doi:10.1001/jama.298.6.644
13. Lipton E, Steinhauer J. Social Distancing for Coronavirus Has a History. The New York Times. Published 2020. Accessed December 11, 2020. https://www.nytimes.com/2020/04/22/us/politics/social-distancing-coronavirus.html
14. Glass RJ, Glass LM, Beyeler WE, Min HJ. Targeted social distancing design for pandemic influenza. *Emerg Infect Dis*. 2006;12(11):1671-1681. doi:10.3201/eid1211.060255

15. Rothstein MA, Alcalde MG, Nanette MPH, et al. *QUARANTINE AND ISOLATION: LESSONS LEARNED FROM SARS A Report to the Centers for Disease Control and Prevention*.; 2003.

16. Talus Analytics, Georgetown University Center for Global Health Science and Security. COVID AMP. Published 2020. Accessed March 10, 2021. https://covidamp.org/about/doc

17. Thu TPB, Ngoc PNH, Hai NM, Tuan LA. Effect of the social distancing measures on the spread of COVID-19 in 10 highly infected countries. *Sci Total Environ*. 2020;742:140430. doi:10.1016/j.scitotenv.2020.140430

18. Subramanian V, Kattan MW. Why Is Modeling Coronavirus Disease 2019 So Difficult? *Chest*. 2020;158(5):1829-1830. doi:10.1016/j.chest.2020.06.014

19. Hill A. Modeling COVID-19 Spread vs Healthcare Capacity. Published 2020. Accessed April 7, 2021. https://alhill.shinyapps.io/COVID19seir/

20. Eikenberry SE, Mancuso M, Iboi E, et al. To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infect Dis Model*. 2020;5:293-308. doi:10.1016/j.idm.2020.04.001

21. Delen D, Eryarsoy E, Davazdahemami B. No Place Like Home: Cross-National Data Analysis of the Efficacy of Social Distancing During the COVID-19 Pandemic. *JMIR public Heal Surveill*. 2020;6(2):e19862. doi:10.2196/19862

22. Gao S, Rao J, Kang Y, et al. Association of Mobile Phone Location Data Indications of Travel and Stay-at-Home Mandates With COVID-19 Infection Rates in the US. *JAMA Netw open*. 2020;3(9):e2020485. doi:10.1001/jamanetworkopen.2020.20485

23. Banerjee T, Nayak A. U.S. county level analysis to determine if social distancing slowed the spread of COVID-19. *Rev Panam Salud Publica/Pan Am J Public Heal*. 2020;44.

doi:10.26633/RPSP.2020.90

24. Badr HS, Du H, Marshall M, Dong E, Squire MM, Gardner LM. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *Lancet Infect Dis*. Published online 2020. doi:10.1016/S1473-3099(20)30553-3

25. Chang S, Pierson E, Koh PW, et al. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*. 2021;589(7840):82-87. doi:10.1038/s41586-020-2923-3

26. Engle S, Stromme J, Zhou A. Staying at Home: Mobility Effects of COVID-19. *SSRN Electron J*. Published online April 16, 2020. doi:10.2139/ssrn.3565703

27. Gupta S, Nguyen TD, Lozano Rojas F, et al. Tracking Public and Private Responses to the COVID-19 Epidemic: Evidence from State and Local Government Actions. Published online 2020. Accessed March 14, 2021. http://www.nber.org/papers/w27027

28. Lee MI, Zhao J, Sun QI, et al. Human mobility trends during the early stage of the COVID-19 pandemic in the United States. Published online 2020. doi:10.1371/journal.pone.0241468

29. SteelFisher GK, Blendon RJ, Bekheit MM, Lubell K. The public's response to the 2009 H1N1 influenza pandemic. *N Engl J Med*. 2010;362(22). doi:10.1056/NEJMP1005102

30. Lee M, Zhao J, Sun Q, et al. Human mobility trends during the early stage of the COVID-19 pandemic in the United States. *PLoS One*. 2020;15(11 November). doi:10.1371/journal.pone.0241468

31. Monod M, Blenkinsop A, Xi X, et al. Age groups that sustain resurging COVID-19 epidemics in the United States. *Science (80-)*. 2021;371(6536):eabe8372. doi:10.1126/science.abe8372

32. Clinton J, Cohen J, Lapinski J, Trussler M. Partisan pandemic: How partisanship and public health concerns affect individuals' social mobility during COVID-19. *Sci Adv*. 2021;7(2):eabd7204. doi:10.1126/sciadv.abd7204

33. Baum MA. Red state, blue state, flu state: Media Self-Selection and Partisan Gaps in Swine Flu Vaccinations. *J Health Polit Policy Law*. 2011;36(6):1021-1059. doi:10.1215/03616878-1460569

34. Parker C, Mejia J, Pestilli F. The Spread of COVID-19 Increases With Individual Mobility and Depends on Political Leaning. *Res Sq*. Published online February 4, 2021. doi:10.21203/rs.3.rs-147801/v1

35. Gollwitzer A, Martel C, Brady WJ, et al. Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nat Hum Behav*. 2020;4(11):1186-1197. doi:10.1038/s41562-020-00977-7

36. Hart Research Associates/Public Opinion Strategies. (No Title). NBC News/Wall Street Journal Survey Datasets. Published March 2020. Accessed March 14, 2021. https://www.documentcloud.org/documents/6810602-200149-NBCWSJ-March-Poll-Final-3-14-20-Release.html

37. Kushner Gadarian S, Goodman SW, Pepinsky TB. Partisanship, Health Behavior, and Policy Attitudes in the Early Stages of the COVID-19 Pandemic. *SSRN Electron J*. Published online March 31, 2020. doi:10.2139/ssrn.3562796

38. National Conference of State Legislatures. COVID-19: Essential Workers in the States. Accessed March 14, 2021. https://www.ncsl.org/research/labor-and-employment/covid-19-essential-workers-in-the-states.aspx

39. Zhang X, Warner ME. Covid-19 policy differences across us states: Shutdowns,

reopening, and mask mandates. *Int J Environ Res Public Health*. 2020;17(24):1-17. doi:10.3390/ijerph17249520

40. Braveman P, Gottlieb L. The social determinants of health: It's time to consider the causes of the causes. *Public Health Rep*. 2014;129(SUPPL. 2):19-31. doi:10.1177/00333549141291s206

41. McCormack G, Avery C, Spitzer AKL, Chandra A. Economic Vulnerability of Households with Essential Workers. *JAMA - J Am Med Assoc*. 2020;324(4):388-390. doi:10.1001/jama.2020.11366

42. Thompson HA, Mousa A, Dighe A, et al. Report 38: SARS-CoV-2 setting-specific transmission rates: a systematic review and meta-analysis. doi:10.25561/84270

43. *Summary Report on Outbreaks and Exposure Settings for COVID-19 Cases in King County, WA.*; 2020.

44. Azar KMJ, Shen Z, Romanelli RJ, et al. Disparities in outcomes among COVID-19 patients in a large health care system in California. *Health Aff*. 2020;39(7):1253-1262. doi:10.1377/hlthaff.2020.00598

45. Centers for Disease Control and Prevention (CDC). *COVIDView: Key Updates for Week 21, Ending May 23, 2020.*; 2020.

46. Millett GA, Jones AT, Benkeser D, et al. Assessing differential impacts of COVID-19 on black communities. *Ann Epidemiol*. 2020;47:37-44. doi:10.1016/j.annepidem.2020.05.003

47. Williams DR, Cooper LA. COVID-19 and Health Equity - A New Kind of "herd Immunity." *JAMA - J Am Med Assoc*. 2020;323(24):2478-2480. doi:10.1001/jama.2020.8051

48. Selden TM, Berdahl TA. COVID-19 and racial/ethnic disparities in health risk,

employment, and household composition. *Health Aff*. 2020;39(9):1624-1632. doi:10.1377/hlthaff.2020.00897

49. Selden TM, Berdahl TA. COVID-19 and racial/ethnic disparities in health risk, employment, and household composition. *Health Aff*. 2020;39(9):1624-1632. doi:10.1377/hlthaff.2020.00897
50. Social Distancing Metrics. Accessed March 10, 2021. https://docs.safegraph.com/docs/social-distancing-metrics
51. United States Census Bureau. American Community Survey.
52. MIT Election Data and Science Lab. County Presidential Election Returns. Published online 2018. https://doi.org/10.7910/DVN/VOQCHQ
53. The New York Times and Dynata. Mask-Wearing Survey Data: Estimates from The New York Times, based on roughly 250,000 interviews conducted by Dynata from July 2 to July 14. Published 2020. Accessed March 11, 2021. https://github.com/nytimes/covid-19-data/tree/master/mask-use
54. The New York Times. Coronavirus (COVID-19) Data in the United States. Published 2020. Accessed March 11, 2021. https://github.com/nytimes/covid-19-data
55. Reitsma MB, Salomon JA, Goldhaber-Fiebert JD. Mapping Inequality in SARS-CoV-2 Household Exposure and Transmission Risk in the USA. *J Gen Intern Med*. Published online February 18, 2021:1. doi:10.1007/s11606-021-06603-0
56. Mongey S, Pilossoph L, Weinberg A, Griffin KC. *Which Workers Bear the Burden of Social Distancing Policies?*; 2020. Accessed December 13, 2020. http://www.nber.org/papers/w27085
57. Raine S, Liu A, Mintz J, Wahood W, Huntley K, Haffizulla F. Racial and ethnic

disparities in covid-19 outcomes: Social determination of health. *Int J Environ Res Public Health*. 2020;17(21):1-16. doi:10.3390/ijerph17218115

58. Ballotpedia. Status of lockdown and stay-at-home orders in response to the coronavirus (COVID-19) pandemic, 2020. Published 2020. Accessed April 2, 2021. https://ballotpedia.org/Status_of_lockdown_and_stay-at-home_orders_in_response_to_the_coronavirus_(COVID-19)_pandemic,_2020
59. Safegraph. Social Distancing Metrics. Published 2020. Accessed March 10, 2021. https://docs.safegraph.com/docs/social-distancing-metrics
60. US Centers for Disease Control and Prevention (CDC). CDC SVI 2018 Documentation. Published January 31, 2020. Accessed March 19, 2021. https://svi.cdc.gov/Documents/Data/2018_SVI_Data/SVI2018Documentation.pdf
61. Dingel JI, Neiman B. *How Many Jobs Can Be Done at Home?*; 2020. Accessed December 13, 2020. http://www.nber.org/papers/w26948
62. DingelNeiman-workathome/occupations_workathome.csv at master · jdingel/DingelNeiman-workathome. Accessed March 3, 2021. https://github.com/jdingel/DingelNeiman-workathome/blob/master/occ_onet_scores/output/occupations_workathome.csv
63. Breiman L. *Random Forests*. Vol 45.; 2001.
64. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;63(1):3-42. doi:10.1007/s10994-006-6226-1
65. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.24.1 documentation. Accessed March 9, 2021. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

66. Cobb JS, Seale MA. Examining the effect of social distancing on the compound growth rate of COVID-19 at the county level (United States) using statistical analyses and a random forest machine learning model. *Public Health*. 2020;185:27-29. doi:10.1016/j.puhe.2020.04.016

67. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer; 2009.

68. US Centers for Disease Control and Prevention (CDC). Management of Patients with Confirmed 2019-nCoV | CDC. Accessed March 11, 2021. https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html

69. The New York Times. The U.S. Passes 4 Million Cases in November Alone, Doubling October's Tally . Published November 30, 2020. Accessed April 2, 2021. https://www.nytimes.com/live/2020/11/28/world/covid-19-coronavirus

70. Adhikari S, Pantaleo NP, Feldman JM, Ogedegbe O, Thorpe L, Troxel AB. Assessment of Community-Level Disparities in Coronavirus Disease 2019 (COVID-19) Infections and Deaths in Large US Metropolitan Areas. *JAMA Netw open*. 2020;3(7):e2016938. doi:10.1001/jamanetworkopen.2020.16938

71. Reitsma MB, Salomon JA, Goldhaber-Fiebert JD. Mapping Inequality in SARS-CoV-2 Household Exposure and Transmission Risk in the USA. *J Gen Intern Med*. Published online February 18, 2021:1. doi:10.1007/s11606-021-06603-0

72. Friess S. Lt. Gov. Garlin Gilchrist on the Racial Disparities Revealed by COVID-19. The Hour. Published May 28, 2020. Accessed March 30, 2021. https://www.hourdetroit.com/community/racial-disparities-coronavirus-garlin-gilchrist/

73. Too Many Black Americans Are Dying from COVID-19. Scientific American. Accessed December 22, 2020. https://www.scientificamerican.com/article/too-many-black-americans-are-dying-from-covid-19/

APPENDIX

Table 5. The cardinality of each of the features in our dataset. *Percent increase is measured from the first week of January to the first week of April.

Feature	Number of Unique Values
Percent increase in individuals staying at home*	359
Percent of Republican votes in the 2016 presidential election	359
Population density	359
Percent increase in home dwell time*	359
Percent increase in median non-home dwell time*	359
Estimate of the percentage of the population that can telework	359
Total population	358
Average new reported COVID-19 cases per 100,000 people	351
Percent of survey respondents who reported frequently or always wearing a mask in public	264
Percentage racial minority (all persons except white, non-Hispanic)	246
Percent of population below the poverty line	149
Percent of population over 65 years old	143
Percent of households with more people than rooms	75
Start date of the policy environment	15

Table 6. The Pearson correlation coefficients between the predictor variables across all US counties with available information (n = 3114). Note that measures of mobility are not available for all US counties.

	Estimate of the percentage of the population that can telework	Percent of survey respondents who reported frequently or always wearing a mask in public	Percent of Republican votes in the 2016 presidential election	Total population	Percent of population without health insurance	Percent of population below the poverty line	Percent of households with more people than rooms	Percentage racial minority (all persons except white, non-Hispanic)	Percent of population without a high school diploma	Percent of population over 65 years old	Population density
Estimate of the percentage of the population that can telework	1										
Percent of survey respondents who reported frequently or always wearing a mask in public	0.295327	1									
Percent of Republican votes in the 2016 presidential election	-0.35539	-0.49347	1								
Total population	0.341889	0.27404	-0.34874	1							
Percent of population without health insurance	-0.30388	-0.02421	0.10474	-0.04077	1						
Percent of population below the poverty line	-0.46342	-0.02683	-0.12299	-0.0744	0.439669	1					
Percent of households with more people than rooms	-0.13264	0.155887	-0.1933	0.197268	0.502299	0.323377	1				
Percentage racial minority (all persons except white, non-Hispanic)	-0.01912	0.35411	-0.53057	0.24626	0.51398	0.445648	0.563525	1			
Percent of population without a high school diploma	-0.56164	0.015198	0.090851	-0.05028	0.56868	0.625999	0.451369	0.500475	1		
Percent of population over 65 years old	-0.14485	-0.15008	0.285878	-0.21771	-0.13083	-0.10069	-0.33137	-0.38695	-0.13842	1	
Population density	0.241953	0.153745	-0.26327	0.33175	-0.03985	-0.00875	0.133136	0.157799	-0.01413	-0.1228	1

Table 7. The P-values for the above correlations between the predictor variables across all US counties with available information (n = 3114). Note that measures of mobility are not available for all US counties.

	Estimate of the percentage of the population that can telework	Percent of survey respondents who reported frequently or always wearing a mask in public	Percent of Republican votes in the 2016 presidential election	Total population	Percent of population without health insurance	Percent of population below the poverty line	Percent of households with more people than rooms	Percentage racial minority (all persons except white, non-Hispanic)	Percent of population without a high school diploma	Percent of population over 65 years old	Population density
Estimate of the percentage of the population that can telework	0										
Percent of survey respondents who reported frequently or always wearing a mask in public	1.08E-63	0									
Percent of Republican votes in the 2016 presidential election	2.46E-93	1.25E-190	0								
Total population	4.45E-86	8.68E-55	5.50E-90	0							
Percent of population without health insurance	1.66E-67	0.176160058	4.39E-09	0.0229511	0						
Percent of population below the poverty line	1.46E-165	0.134429971	5.75E-12	3.25E-05	2.48E-147	0					
Percent of households with more people than rooms	1.08E-13	2.11E-18	1.15E-27	1.11E-28	1.06E-198	1.07E-76	0				
Percentage racial minority (all persons except white, non-Hispanic)	0.286192733	1.86E-92	1.84E-224	1.91E-44	3.27E-210	8.75E-152	1.22E-261	0			
Percent of population without a high school diploma	2.47E-258	0.397090442	3.67E-07	0.005015253	3.06E-266	0	3.93E-156	9.58E-198	0		
Percent of population over 65 years old	4.60E-16	3.84E-17	1.03E-59	1.02E-34	2.33E-13	1.80E-08	1.13E-80	4.45E-112	8.68E-15	0	
Population density	1.04E-42	6.15E-18	1.15E-50	7.39E-81	0.026219446	0.625688796	8.77E-14	6.69E-19	0.430595076	6.20E-12	0

Table 8. The Pearson correlation coefficients for all the predictor variables inputted in our model (n = 359). *Percent increase is measured from the first week of January to the first week of April.

	Start date of the policy environment	Percentage racial minority (all persons except white, non-Hispanic)	Percent of households with more people than rooms	Percent of population over 65 years old	Estimate of the percentage of the population that can telework	Percent increase in individuals staying at home*	Percent of population below the poverty line	Percent increase in home dwell time*	Percent of Republican votes in the 2016 presidential election	Percent of survey respondents who reported frequently or always wearing a mask in public	Population density
Start date of the policy environment	1										
Percentage racial minority (all persons except white, non-Hispanic)	0.27198	1									
Percent of households with more people than rooms	0.293186	0.664361	1								
Percent of population over 65 years old	0.196404	-0.36098	-0.39946	1							
Estimate of the percentage of the population that can telework	0.091019	0.23336	0.027943	-0.24174	1						
Percent increase in individuals staying at home*	-0.37491	-0.06088	0.024387	-0.44548	0.354973	1					
Percent of population below the poverty line	0.244966	0.428882	0.297164	0.001515	-0.31029	-0.40603	1				
Percent increase in home dwell time*	-0.13346	-0.14749	0.06356	-0.22262	0.100812	0.59742	-0.23163	1			
Percent of Republican votes in the 2016 presidential election	-0.17315	-0.60992	-0.36868	0.313816	-0.63604	-0.20395	-0.07357	-0.10322	1		
Percent of survey respondents who reported frequently or always wearing a mask in public	0.041786	0.279445	0.172917	-0.17042	0.458495	0.067194	-0.11462	0.071502	-0.56607	1	
Population density	-0.00718	0.275921	0.10211	-0.24386	0.452185	0.233549	-0.06026	0.154839	-0.41238	0.202806	1

Table 9. The P-values for the above correlations between the predictor variables across all the counties studied (n = 359). *Percent increase is measured from the first week of January to the first week of April.

	Start date of the policy environment	Percentage racial minority (all persons except white, non-Hispanic)	Percent of households with more people than rooms	Percent of population over 65 years old	Estimate of the percentage of the population that can telework	Percent increase in individuals staying at home*	Percent of population below the poverty line	Percent increase in home dwell time*	Percent of Republican votes in the 2016 presidential election	Percent of survey respondents who reported frequently or always wearing a mask in public	Population density
Start date of the policy environment	0										
Percentage racial minority (all persons except white, non-Hispanic)	1.66E-07	0									
Percent of households with more people than rooms	1.51E-08	4.60E-47	0								
Percent of population over 65 years old	0.00018	1.72E-12	3.48E-15	0							
Estimate of the percentage of the population that can telework	0.085046	7.90E-06	0.597709	3.60E-06	0						
Percent increase in individuals staying at home*	2.00E-13	0.249925	0.645139	6.66E-19	4.23E-12	0					
Percent of population below the poverty line	2.64E-06	1.70E-17	9.41E-09	0.977179	1.88E-09	1.11E-15	0				
Percent increase in home dwell time*	0.011365	0.005108	0.229636	2.07E-05	0.056351	4.20E-36	9.26E-06	0			
Percent of Republican votes in the 2016 presidential election	0.000987	5.94E-38	5.32E-13	1.21E-09	4.27E-42	9.95E-05	0.164257	0.050688	0		
Percent of survey respondents who reported frequently or always wearing a mask in public	0.429924	7.29E-08	0.001003	0.001189	4.62E-20	0.204031	0.029912	0.17645	8.39E-32	0	
Population density	0.892155	1.08E-07	0.053234	2.94E-06	1.71E-19	7.76E-06	0.254806	0.003268	3.58E-16	0.000109	0