Increasing Uptake of Social Distancing During COVID-19: How Machine Learning Strategies Can Lead to Targeted Interventions

Grace Charles¹ Mokshada Jain¹ Yael Caplan¹ Hannah Kemp¹
Aysha Keisler² Vincent Huang¹ Sema K. Sgaier^{1,3,4}

¹Surgo Ventures, District of Columbia, United States of America,

²Keisler Social and Behavioral Research, District of Columbia, United States of America,

³Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America,

⁴Department of Global Health, University of Washington, Seattle, Washington, United States of America

Published on: Jan 06, 2022

DOI: <https://doi.org/10.1162/99608f92.17a4e925>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

Social distancing emerged as one of the early critical nonpharmaceutical interventions to fight the spread of COVID-19. However, in the United States, this behavior was not universally adopted. Understanding COVID-relevant behaviors is crucial to helping policymakers develop targeted, actionable interventions that meet the urgent needs of a global pandemic in a precision public health approach—that is, getting the right intervention to the right person, at the right time and place. In this article, we demonstrate how using a toolbox that includes a comprehensive data collection design framework, machine learning tools to do causal discovery (i.e., structural learning of a causal Bayesian network), and clustering analysis can help disentangle the intertwined hierarchy of drivers of social distancing to design targeted, actionable interventions in a precision public health application. We integrated several machine learning techniques to generate insights from a nationally representative social distancing survey of 2,500 U.S. respondents, conducted in March 2020. This approach goes beyond measuring the correlates of social distancing intentions and behavior and narrows in on the potential causal drivers of social distancing. Our approach identifies the factors that social distancing directly and conditionally depends upon: higher financial security, higher information-seeking, and higher worry about the coronavirus, as well as other upstream factors. We also identify four population segments to help target interventions: “Worriers,” “Rule-followers,” “Financially constrained,” and “Skeptics.” Two subgroups in particular, the “Skeptics” and the “Financially constrained,” had low uptake of social distancing, but would require different targeted messages to increase social distancing behavior. Taken together, these results demonstrate how machine learning techniques can help prioritize messages most effective for matched population targets, increasing desirable outcomes while potentially saving resources.

Keywords: social distancing, causal Bayesian network, segmentation, COVID-19, coronavirus, precision public health

Media Summary

Social distancing remains a crucial measure to slow the spread of COVID-19, but it has never been universally adopted in the United States. One reason is that public health campaigns often do not take into account the range of psychobehavioral and structural factors affecting people’s decision to social distance—factors that cut across simplistic demographic groupings.

This article presents a data-driven approach that integrates multiple methods to disentangle the complex drivers of this behavior—the who and why behind social distancing. By integrating multiple approaches, we are able to offer a holistic picture of the drivers behind social distancing and provide actionable and targeted solutions.

Through causal analysis, we found that social distancing intentions and behaviors are influenced most strongly by information-seeking, worries about infection, and financial insecurity. If a public health campaign could

move people from having no intention to social distance to intending always to do so, their odds of social distancing would be 12.8 times higher; and if people were persuaded to seek out more information on the coronavirus, they would be 1.9 times more likely to social distance than people who did not seek information. If made concerned about the pandemic or financial security, people's odds of social distancing would be 1.4 times and 1.2 times higher, respectively.

The approach that led to these insights involved layering several analytical methods. A nationally representative survey investigated the beliefs and perceptions around social distancing of 2,500 people across the United States. Traditional regression methods applied to this data revealed the characteristics of individuals who social distance, and a broad set of factors correlated with social distancing. These results fed into a causal Bayesian network analysis, which mapped the factors with direct and indirect influence upon the decision to social distance, and pointed to the kinds of interventions most likely to increase the odds of people social distancing. Surprisingly, the causal model revealed that factors such as political affiliation and race, which were significant *predictors* of social distancing, were not significant, direct *causal* drivers of social distancing. Last, a cluster analysis identified population segments differentiated by their risk perception, financial stability, beliefs about community social norms, self-efficacy, and information-seeking about social distancing. This analysis could help target messaging more precisely to population segments with differing barriers.

This comprehensive approach shows how integrating machine learning techniques can help go from data mining to prioritizing messaging for matched population targets, increasing desirable outcomes while potentially saving resources. It is an approach that has applications not just in the present pandemic, but in other public health contexts where *changing* people's behaviors is required.

1. Introduction

The practice of social distancing is key to stopping the spread of pandemics, including both the current COVID-19 pandemic as well as others that may strike in future years (Glass et al., 2006). The U.S. Centers for Disease Control and Prevention (CDC) defines social distancing as keeping space between yourself and other people outside of your home by staying at least six feet from other people, refraining from gathering in groups, and staying out of crowded places and avoiding mass gatherings (CDC, 2020). The CDC suggests that social distancing and wearing masks are the best non-pharmaceutical ways to limit the spread of COVID-19, and experts advocate that social distancing in some form will be needed for the immediate future. Despite the importance of social distancing and policies designed to enforce adherence, uptake has been slow in certain areas and at times inadequate (Abouk & Heydari, 2020). Analysis of mobile phone data between February and September 2020 showed that social distancing in the United States has waxed and waned, peaking in mid-April 2020, and then declining throughout that summer (Huang et al., 2020; Unacast, 2020). Furthermore, although the CDC issued recommendations against celebrating the winter 2020 holidays with large groups of friends and family, nearly 33% of Americans said they would travel (*Holiday Travel Survey 2020—Thanksgiving and*

Christmas, 2020), leading to a winter surge in COVID-19 (Mehta et al., 2020). Understanding the factors that drive social distancing uptake and adherence and how these drivers vary for different individuals is important not only to contextualize the spread of COVID-19, but also to respond to future pandemics. Understanding the cascading drivers of and barriers to preventive behaviors such as social distancing can help focus current and future public health outreach and hone communication strategies.

While changing behaviors typically requires a holistic understanding of the context and reasons behind the behavior (Engl & Sgaier, 2020), most studies focus only on certain subsets of potential behavioral drivers. Research from past epidemics (e.g., 2002 SARS and 2009 H1N1) suggests that social distancing and protective behaviors in general are related to a number of demographic and attitudinal factors, perceptions of community norms, and structural factors such as ability to work from home (Bish & Michie, 2010; Kleczkowski et al., 2015; Lin et al., 2018; Rashid et al., 2015; Teasdale et al., 2014; Tracy et al., 2009; Zhang et al., 2019). Context, including everything from existing infrastructure to community cohesion, can also play a large role in the uptake of protective health behaviors, and health officials have emphasized the need for better data to inform more effective interventions (MacDonald et al., 2018). In the face of a global health crisis and increasingly limited budgets, designing more effective and efficient interventions necessitates the design of intentional, targeted, data-driven solutions (Chowkwanyun et al., 2018; Khoury et al., 2016). On one extreme, public health campaigns that take a one-size-fits-all approach may not sufficiently acknowledge the heterogeneous barriers and beliefs of different groups in the population; on the other extreme, completely personalized message campaigns may not be feasible from an implementation standpoint.

Here we demonstrate an approach integrating multiple methods to uncover integrated insights and home in on behavioral causes to inform interventions. To capture potential factors of social distancing in the early phase of the pandemic, we used a holistic behavioral framework called CUBES (to Change behavior, Understand Barriers, Enablers, and Stages of change; Engl & Sgaier, 2020) to design and deploy a survey in late March 2020 that collected data on demographics, beliefs, risk perceptions, social norms, self-efficacy, awareness, and personal experiences on social distancing in the United States. We then learned a Bayesian network (BN) to identify causal relationships between these factors and identify direct causes (and thus, intervention points) of social distancing behavior; we used a predictive model to select a broad set of factors correlated with social distancing as input. Lastly, to facilitate targeted policy recommendations, k-means clustering was used to identify population segments differing in social distancing beliefs and behavior. Taken together, this multifaceted approach can identify *which* factors predict social distancing, *how* we may increase social distancing behavior, and *who* to target to increase social distancing uptake. These findings can be used to identify the intervention or interventions with the highest potential to effect change.

2. Materials and Methods

2.1. Study Design and Participants

In order to capture as many potential causal drivers of social distancing as possible, we used the CUBES framework to design a survey consisting of 24 elective questions (Appendix A) about respondents' actual and self-perceived knowledge, perceptions and structural barriers around social distancing, access to and trust in various sources for coronavirus information and guidelines, and their current health status and social distancing intention and action rates (Engl & Sgaier, 2020). We defined social distancing as a response of 'always' to 1) avoiding crowds, 2) staying home as much as possible, and 3) maintaining six feet of physical distance from others. These survey questions were supplemented with previously collected panel data on demographics, health conditions, religious beliefs, political beliefs, and location (Appendix B).

Our survey was fielded in both English and Spanish by Ipsos, a market research company. The survey was conducted March 27–31, 2020, using a nationally representative probability sample of 2,500 people aged 18 years or older, drawn from the KnowledgePanel, a well-established online probability-based panel representative of the adult U.S. population (Ipsos, 2020). Panel members are recruited based on random sampling of all available household addresses in the United States. Ipsos provides selected households that do not already have internet access with a tablet and internet connection at no cost. The data were weighted to adjust for gender, age, race, education, census region, metropolitan status, and household income using benchmarks from the 2019 March supplement of the U.S. Census Bureau's Current Population Survey (CPS). The margin of sampling error is ± 2.4 percentage points at the 95% confidence level, for results based on the entire sample of adults. The response rate was 51%. The margin of sampling error takes into account the design effect, which was 1.24. All reported descriptive statistics are weighted based on survey weights. Predictive models (described below) were also weighted according to survey weights.

2.2. Statistical Analyses

Data were analyzed using R (R Core Team, 2020; Wickham et al., 2021; Kassambara and Mundt, 2020; de Mendiburu, 2021). To conduct descriptive analyses, we used survey weights to calculate weighted mean responses. After conducting descriptive analyses, we used multivariable logistic regression models with sample weights to identify characteristics that were associated with higher likelihood of social distancing. The reason for using this methodology was twofold: (1) it helped to select the variables to be used as input for the subsequent causal modeling without eliminating the conditional dependencies (which may show up as multicollinearity) between variables and (2) it also allowed us to later compare insights gained from a traditional predictive model versus insights gained using a more causal framework. We next identified causal cascades of social distancing behavior using a custom implementation of causal BN structural learning in Python (see Sec. 2.4 for details). Finally, we used unsupervised clustering algorithms to define different population segments with contrasting social distancing perceptions and behaviors (Engl et al., 2019). By using a mix of methods, we gained a much richer understanding of the drivers of social distancing, its key causes and how they interact, and the underlying and contrasting perceptions and barriers to social distancing for different

segments of the population (Table 1). This holistic understanding can help design effective and precise behavioral interventions.

Table 1. Breakdown of methods, our reasoning for using the method, and the outcome of the analysis.

Step	Method	Outcome and relationship to other methods
1	Survey instrument design using the CUBES behavioral framework	Design comprehensive survey that captures as many potential causal drivers of social distancing as possible
2	Predictive modeling	Identify characteristics that were associated with higher likelihood of social distancing; use as a comparison to insights gained with causal modeling
3	Causal BN model	Identify causal cascades of social distancing behavior and proximate causes of the behavior
4	Psychobehavioral segmentation	Define different population segments with contrasting social distancing perceptions and behaviors

2.3. Predictive Model

All predictive models were analyzed in R using the ‘glm’ function in the stats package for general linear modeling. We performed a multiple logistic regression to understand the drivers of social distancing behavior. Our predictors of social distancing behavior (a binomial outcome) included demographic variables (gender, age, race, education, religion, geographic region); other KnowledgePanel characteristics (e.g., political affiliation and partisanship, income level); perceptions of social norms around social distancing; risk perceptions around coronavirus and social distancing; information channels (e.g., how much information respondents are seeking out about coronavirus); and personal coronavirus context (e.g., being in a high-risk group or knowing someone with the disease personally). These variables were chosen for the predictive model to encompass the context and reasons behind social distancing behavior, using the validated CUBES framework as a guide to systematically capture as many potential drivers of social distancing as possible. Full details of the variables included in our logistic regression are given in Appendix B. All continuous variables were standardized so that the odds ratio reflects the difference in shifting one standard deviation on that

variable. The *observational* odds ratio should be interpreted as the difference in *observing a shift* on that variable by one standard deviation.

Variables were tested for multicollinearity, and no variables were found to have a variance inflation factor greater than 2.5.

2.4. Causal Machine Learning

A fundamental requirement for interventions to be effective is that they must be enacted on factors that cause the outcome so one would not be just ‘treating the symptoms, not the disease.’ While predictive models can help us understand the correlates of social distancing behavior, they cannot tell us which variables *cause* this behavior unless paired with experimental data collected with a design like a randomized control trial.

Observational data, such as survey data like ours, are easier to come by and likely the only data available in a crisis. Causal machine learning approaches such as a causal Bayesian network (BN) can be used to infer causal dependencies in observational data. Additionally, BN provides a graphical representation of the complex interdependent web of cause and effect behind a behavior in a compact graphical structure known as a directed acyclic graph (DAG). A DAG represents the variables as nodes, and the directed causal influence between them as directed edges. While the mapping of this complex structure can be subjectively defined by expert panels—an approach called ‘Bayesian Belief Networks’—it is most powerfully learned from data directly via structure-learning machine learning algorithms in a process called ‘causal discovery’ without needing an a priori hypothesis (Appendix F). This feature sets it apart from other statistical causal inference methods such as propensity score matching, which depend on a priori hypothesis to set up and can only test one hypothesis at a time. We decided to use BN to map out the conditional dependencies between variables, and deduce the multiple causal pathways that lead to social distancing behavior.

Figuring out the just potential pathways, however, is not sufficient to help prioritize resources in intervention design; the level of impact of multiple candidate interventions should be compared. With the BN, we could do that by performing interventional queries (i.e., estimating how much the level of social distancing behavior would change if we had done X) to compute the interventional odds ratio for different outcomes of interest (details in Appendix D). The interventional odds ratio estimates how much more likely an outcome is, given a change in an upstream variable (i.e., an intervention) (Pearl, 1995, 2009). This is done by inferring the posterior distribution of a given target variable by setting an interventional variable to a specific value, and in turn, the interventional odds ratio. Unlike the observational odds ratio, the *interventional* odds ratio of an outcome variable should be interpreted as the ratio of the odds of the different outcomes when an intervention is enacted or not—which is very much in the same vein as a randomized control trial design.

2.5. Segmentation

To facilitate targeted policy recommendations, there is a need to identify population segments differing in social distancing beliefs and behavior. While the causal model reveals the causal drivers of social distancing

behavior, it cannot tell us if there are different segments in the population that differ in terms of their sets of social distancing beliefs and drivers. Identifying these segments can help uncover which interventions may be most effective for different parts of the population, and give a sense of how much the population falls into a particular segment.

We used one of the most widely used segmentation algorithms, k-means clustering (MacQueen, 1967), to identify population segments differing in psychobehavioral drivers of social distancing. We chose to use k-means because it is fast, gives straightforward interpretations, and allows control over which variables to include in our segmentation. This method was most appropriate because all our variables were numeric. We aimed to identify segments of individuals that differed on the following six variables: perception of community social norms regarding social distancing, risk perception around social distancing, perceived difficulty of social distancing, degree of information-seeking about coronavirus, level of worry about coronavirus, and perceived self-efficacy in stopping the spread of coronavirus. These variables were selected for segmentation based on their relationship to social distancing behavior observed in the predictive models and their actionability in order to identify population groups and effective interventions. All variables used to define segments were normalized to a 0–1 scale.

After defining the variables to include in our k-means segmentation, we explored segmentation solutions with varying numbers of segments. Below, we describe the data-driven and practical considerations we took into account when choosing our clustering solution.

2.6. Evaluating Competing Segment Solutions

There are numerous considerations in determining the optimal number of segments in a segmentation analysis. The best segmentation solution is ultimately the one that yields the most useful and actionable solutions. Choosing the optimal number of segments is not a trivial task and requires a blend of quantitative analysis and qualitative understanding of which interventions can be practically implemented.

From a data perspective, segment solutions should be those that maximize separation between segments. That is, the ideal solution should produce segments that reveal meaningful differences between groups in the characteristics of their constituent variables. One visual and statistical tool we can use to evaluate separations between segments is the visual ‘elbow’ method, where the overall within-segment sum of squares (i.e., sum of the squared deviations from each observation and the segment centroid) is plotted for increasing numbers of segments, and a ‘bend’ in the elbow plot indicates where the addition of more segments does not help explain meaningfully more amounts of the variance.

From a practitioner and policy perspective, segment solutions must also be practical and actionable and should be chosen such that they can target meaningful drivers and barriers to behavior. For example, while the ‘best’ segmentation solution from a data perspective might include 10 segments, the implementation of this solution could be unwieldy and thus less scalable and relevant to policymakers. On the other hand, while a

segmentation solution with fewer segments may be easier to scale and implement, it still needs to yield specific and actionable insights based on segment characteristics (Engl et al., 2019).

After segments are defined, they are then profiled on a variety of demographic and other panel characteristics. Profiling of these segments allows us to match individuals more easily with policy recommendations.

3. Results

Each analysis method provides a new layer of information, and when considered together with the context, allows us to generate holistic insights and then distill the information to inform specific, targeted policies. Our descriptive results show that while social distancing knowledge was high, intentions and actual behavior were much lower. Our predictive model results selected relevant variables that are associated with social distancing as inputs for causal modeling; these included a wide range of perceptual and structural factors such as risk perception and financial stability. The results of the causal model disentangled the complicated causal relationships between the factors correlated with social distancing. In particular, we quantify that the biggest causal drivers of social distancing are a) intent to distance, b) financial security, c) worry about the coronavirus, and d) degree of information-seeking about the virus. Lastly, our segmentation shows that there are distinct groups in the U.S. population with heterogeneous concerns and barriers to social distancing. We describe the detailed results below.

3.1. Descriptive Statistics

In March 2020, knowledge of social distancing was high, but there was scope to increase it further. Of the population, 87.5% correctly answered at least 11 of the 12 social distancing knowledge questions based on CDC guidance at the time (see Appendix A for more details). Figure 1 shows that survey respondents were accessing most information about coronavirus from state and local government, traditional media sources, and national and global health organizations like the CDC and the World Health Organization (WHO). Importantly, while health care providers were the most trusted source of information about coronavirus, they ranked as one of the least-utilized sources of information.

Figure 1. Mean values of how much information, guidance, and opinions respondents reported hearing from different sources, and how much trust respondents reported having in various sources. Values have been normalized to a 0–1 scale. Mean values are representative of the U.S. population.

We found that there was a ‘know-intent-practice’ gap: high levels of knowledge did not correspond to equal levels of intention and action for social distancing. Even though 87.5% of the population had high levels of knowledge about social distancing, only 62.5% intended to always practice social distancing, and an even lower proportion (46.2%) always practiced social distancing.

The gap between intention and action was due to the 21.4% of respondents who intended to always social distance but did not. Counterintuitively, 5% of respondents reported that they were always social distancing, but did not intend to always social distance. There was no significant difference in social distancing behavior between respondents in states with and without stay-at-home orders at the time of the survey (Tukey $p > .70$ for all comparisons; mean social distancing uptake in states with statewide orders was 46.4%; with partial orders 45.4%; and with no orders 46.5%).

3.2. Individual-Level Characteristics of Social Distancers: Predictive Model

We first used multivariable regression to identify variables correlated with social distancing uptake. The predictive model serves three purposes—first, to identify correlations between variables in our data set; second, to serve as a means of variable selection for the causal model; and third, as a point of comparison between correlate identification and input variable selection for the causal model.

For simplicity we assume no interaction terms. Our model’s Area Under the Curve (AUC) was 0.68. Male respondents were less likely to social distance than female respondents (OR = 0.75; 95% CI = 0.62–0.90; $p < .001$; Figure 2), and Catholic respondents were more likely to social distance compared with Evangelical respondents (OR = 1.40; 95% CI = 1.09–1.80; $p < .001$). The effects of belonging to any other religious group were nonsignificant. High-income respondents were less likely to social distance than low-income respondents (for incomes over \$100,000, OR = 0.68; 95% CI = 0.48–0.98; $p < .01$). However, those who perceived themselves as more financially stable were more likely to social distance (OR = 1.05; 95% CI = 1.02–1.07; $p < .01$).

.001). All other respondent characteristics, including race, education, age, geographic region, and political affiliation, were not significant predictors of social distancing uptake. The observation that higher-income individuals are less likely to social distance appears to contradict recent observations reported elsewhere (Chang et al., 2021). However, they may not be contradictory at all. In the model, we included financial security, a variable that captured the degree to which a respondent felt they could a) work from home, b) receive income from their employer even if COVID prevented them from working, and c) pay all their bills for the next two months. We believe this variable would serve as a more direct predictor for social distancing. Second, high income may suggest greater capability of traveling longer distances without being adversely affected by a pandemic lockdown (e.g., not having to rely on public transportation). Conversely, those with no or low income may be more likely to be retired, and may find social distancing easier with fewer professional commitments. Admittedly, these nuances may be challenging to tease out in a predictive model and are perhaps better suited for the causal analysis in Sec. 3.3.

Figure 2. Observational odds ratios of predictors on social distancing with 95% confidence intervals. Variables in red are negatively associated with social distancing; variables in cyan are positively associated with social distancing, at a cut-off of $p < .05$ uncorrected for multiple comparisons. Variables in black have no statistically significant relationship with social distancing behavior. All linear variables have been standardized so that the odds ratio reflects the difference in shifting one standard deviation from the mean on that variable.

In addition to demographic characteristics, some of the strongest predictors of social distancing behavior were related to perceptions and beliefs. Specifically, respondents who were more likely to social distance included those who believed they could help prevent the spread of coronavirus (OR = 1.27; 95% CI = 1.18–1.36; $p < .001$), those with higher risk perceptions around the negative health consequences of not social distancing (OR = 1.07; 95% CI = 1.04–1.11; $p < .001$), those who perceived strong community norms toward social distancing (OR = 1.16; 95% CI = 1.09–1.24; $p < .001$) and those who sought more information about coronavirus (OR = 1.24; 95% CI = 1.14–1.35; $p < .001$). Finally, respondents who perceived certain aspects of social distancing as more difficult—such as feeling isolated, not leaving the house, not attending community gatherings, and not receiving in-person services—were less likely to social distance (OR = 0.85; 95% CI = 0.79–0.92; $p < .001$).

3.3. Identifying Interventions to Increase Social Distancing: Causal Bayesian Network

As we were ultimately interested in intervention recommendations that would improve the level of social distancing intent and behavior, our next step was to employ causal discovery and inference using a causal BN. Because the number of possible structures grows exponentially with the number of variables, the computational problem of identifying the best-fit structure also grows exponentially harder, leading to a tradeoff between algorithm performance (and to the extreme, computational feasibility) and the number of variables that we can include (Chickering et al., 2004). As far as we know, there is no standard way of selecting variables for causal BN analysis. To inform our selection, we first estimated the threshold for the maximum number of variables that would still allow us to achieve an acceptable level of performance with synthetic data sets. For our sample size of 2,500 and assuming the ground truth takes the form of ic-DAG (a graph generation algorithm that samples uniformly from the set of DAGs), this threshold was determined to be 17 variables (Butcher et al., 2021; Ide & Cozman, 2002). We therefore prioritized variables for inclusion in the BN model based on their significance in the predictive model. The finalized data set was then used to learn the causal BN structure.

The learned BN structure revealed a network of relationships between demographics, beliefs, social distancing intentions, and social distancing behaviors, which is represented in a DAG (Figure 3). The arrows depict the directions of conditional dependence, whereas the absence of arrows indicates conditional independence (such as is the case for race, religion, and, interestingly, having health risk). When two variables have one or more common causes but no arrows between themselves, the interpretation is that the two variables are confounded (e.g., financial security and information-seeking are confounded by age group). The learned model showed that information-seeking and social distance intent directly influenced whether people ultimately social distanced. Less direct influences include gender, political party, worries about coronavirus, income, education, age, and financial security. Social distancing behavior did not appear to be conditionally dependent on risk perception, perceived self-efficacy on stopping viral spread, perception of the community norm, and perceived difficulty in social distancing.

Figure 3. Structural graph depicting the causal relationships between respondent demographics, beliefs, and social distancing outcomes. Outcome of interest, social distancing behavior, is highlighted in blue. All significant direct causal drivers of social distancing behavior are outlined in blue, and all significant direct causal drivers of social distancing intentions are outlined in orange.

We were particularly interested in understanding how drivers of social distancing beliefs are related to each other and to social distancing intentions. Our intervention (what-if) analysis revealed that if we were able to persuade people having no intention to social distance to have the intention to always social distance, then their odds of social distancing would be 12.8 times higher (95% CI = 11.1–14.8). We also found that if people were persuaded to seek out information on coronavirus more, they would be 1.9 times (95% CI = 1.7–2.1) more likely to social distance than those who did not. Similarly, if we made people more worried about coronavirus, they would be 1.4 times (95% CI = 1.2–1.5) more likely to social distance. Interventions to increase financial security would result in higher social distancing 1.2 times (95% CI = 1.1–1.4) (Figure 4). Lastly, those belonging to Gen Z or Millennials would have social distanced 1.2 times more (95% CI = 1.0–1.3) had they been Baby Boomers.

Given that intention was the strongest cause of social distancing behavior, we asked which variables increased the odds of always intending to social distance. The trends were similar to causal factors of social distancing behavior, with the exception that income now also became a significant cause (Figure 4).

Figure 4. What-if analysis (reference/intervention) for our causal BN. (A) Interventional odds ratio of always social distancing versus never or sometimes social distancing and (B) interventional odds ratio of intending to always social distance versus intending to never or sometimes social distance, with 95% confidence intervals. For example, respondents are 12.8 times more likely to social distance if one is able to persuade them to always intend to do so. Nonsignificant drivers are in black, while significant drivers are in cyan.

3.4. Breaking Down Social Distancing Beliefs and Behavior by Population Segments: Segmentation

Our segmentation analysis revealed population segments that were differentiated by social distancing beliefs and perceptions. Our ‘elbow’ analysis, where the overall within-segment sum of squares is plotted for increasing numbers of segments, found a ‘bend’ in the elbow plot around three to four segments. We next looked at the composition and characteristics of segmentation solutions with varying numbers of segments. Segment solutions from three to six groups were explored. The three-segment solution identified the following: one segment with high rates of social distancing, and two segments with lower rates of social distancing. While this solution offered some resolution on the reasons why individuals did not social distance, we felt that solutions with more segments might help us understand the reasons why individuals did social distance. The four-segment solution of our segmentation revealed four groups with contrasting characteristics: one segment with high rates of social distancing and high risk perception around COVID; a second segment with lower risk perception, but still high rates of social distancing; a third segment with high risk perception, but more financial barriers to social distancing and lower rates of social distancing; and a fourth segment with low social distancing, low risk perception, low information-seeking, and low rates of social distancing. Meanwhile, the five- and six-segment solutions identified one better-performing social distancing group, one poorly performing group, and three to four groups with intermediate levels of social distancing. However, these solutions did not offer much more differentiation than the four-segment solution for the groups with intermediate rates of social distancing. Based on these findings, we considered the four-segment solution most actionable. Results from this segmentation are reported here.

In the four-segment solution (Table 2), segments were distinguished by their social distancing rates in a pattern consistent with the relation between the segmenting variables and the predictive model results. Critically, these

segments were differentiated by variables that our BN model had found to be causal drivers of social distancing behavior. Profiling revealed that these segments also differed in their social distancing behaviors and intentions.

Segment 1, the "Worriers" segment, had high rates of information-seeking and were on average more worried than other segments about coronavirus. Profiling revealed that respondents from this group social distanced at a rate of 55.3%. Segment 2, the "Rule followers" segment, were seeking out information about coronavirus at high rates, but were less worried about the virus than Segment 1 (post hoc Tukey test $p < .05$). Profiling revealed that this segment had higher financial stability than the other segments (Tukey $p < .05$) and social distanced at a rate of 53.8%, a rate comparable to Segment 1 (Tukey $p = .94$). Segment 3, the "Financially constrained" segment, also had a high degree of information-seeking and a middling amount of worry about coronavirus. However, profiling revealed that this segment only social distanced at a rate of 44.1%. While this segment was similar to Segment 2 in terms of beliefs, profiling revealed that Segment 3 had lower financial stability scores than Segment 2 (Tukey $p < .05$). Finally, Segment 4, the "Skeptics" segment, had significantly lower rates of information-seeking and were less worried about coronavirus than Segments 1 or 3. Profiling showed that this segment was only social distancing at a rate of 25.9%.

Segment profile characteristics differed significantly on other demographic characteristics (Table 2 and Appendix C). Republican respondents and men were overrepresented in Segments 2 and 4 with low worry about coronavirus. The "Skeptics" of Segment 4 had the highest proportion of men (57.5%).

Table 2. Characteristics of the four population segments identified and profiling variable means.

	Segment 1	Segment 2	Segment 3	Segment 4
Description	Worriers	Rule followers	Financially constrained	Skeptics
Segment size (% of survey)	28.5	27.3	20.8	23.4
Segment variables (variables used to calculate segments)				
Degree of information-seeking	0.88 (ab)	0.89 (a)	0.86 (b)	0.69 (c)
Mean worry about coronavirus	0.73 (a)	0.29 (c)	0.55 (b)	0.29 (c)

Perception of strong community norms of social distancing	0.84 (a)	0.85 (a)	0.38 (c)	0.64 (b)
Risk perception of social distancing impact on health	0.68 (a)	0.37 (c)	0.62 (b)	0.33 (d)
Perceived difficulty of social distancing	0.53 (a)	0.23 (d)	0.36 (b)	0.30 (c)
Perceived self-efficacy in preventing spread of coronavirus	0.77 (b)	0.85 (a)	0.80 (b)	0.47 (c)
Social distancing profile (mean values for each segment)				
Social distancing uptake (%)	55.3 (a)	53.8 (a)	44.1 (b)	25.9 (c)
Social distancing intent (%)	71.1 (a)	67.8 (a)	73.1 (a)	36.6 (b)
Financial stability score (0–1)	0.63 (b)	0.7 (a)	0.6 (b)	0.61 (b)

Note. All values of characteristics were normalized to a 0–1 scale, with 0 being the lowest possible value and 1 being the highest possible value. Segment variables represent variables used in k-means calculations of segments. Profiling variables represent mean value or percentage for all individuals in each segment. Tukey post hoc pairwise tests show significant differences between segments where $a > b > c > d$, with $p < .05$ for each step. Variables shown in bold font were causal drivers of social distancing in our learned BN.

3.5. Combining Results From Segmentation and Causal Bayesian Network

Our segment analysis revealed two segments of the population with troublingly low rates of social distancing, but different sets of beliefs and perceptions. Individuals in the “Financially constrained” segment were social distancing 44% of the time. They had high intention to social distance, and high amounts of information-seeking, but were only moderately worried about coronavirus. This segment also had lower average financial stability scores than the “Rule followers” segment, a segment with a similar belief profile, but higher rates of social distancing. The “Skeptics” segment was social distancing at an even lower rate of 25.9%. Members of this segment had low intent to social distance and the lowest degree of information-seeking of all four segments. Profiling revealed that members of this segment were more likely to be male and a member of the Republican Party.

This indicates that the two segments should be targeted with different types of messaging. Our results from the causal BN suggest that certain interventions can be prioritized (Figure 4).

For the “Financially constrained” segment, even though their profile indicates weaknesses in risk perception as well as in financial security, causal analysis suggests that increasing financial security through means such as stimulus checks or enhanced unemployment benefits could be the more critical step to boosting social distancing behavior. Individuals in this segment appear to want to social distance but cannot do so, at least in part because of financial barriers.

For the “Skeptics” segment, the identifying characteristics are weaknesses in risk perception and the fact that they did not seek out information. Our causal analysis suggests that seeking information is a critical gateway to social distancing behavior. Thus, messages about social distancing focused on risk perception and social distancing intentions should be delivered through less traditional channels (e.g., trusted health care providers rather than traditional media), given that members of this group sought information at lower rates. Messaging that focuses on increasing intentions, such as pledges to social distance, could also help increase social distancing behavior in this segment.

4. Discussion

Influencing human behaviors such as social distancing is a complex task. With many factors at play, it can be overwhelming and difficult to arrive at a precise, targeted approach. The machine learning methodologies used here range from frequently used and well-understood algorithms (i.e., regression and k-means clustering) to novel and emerging approaches (i.e., causal machine learning). The three approaches help reveal insights from three different angles:

1. What are the characteristics of individuals who social distance? (regression)
2. What interventions can we employ to increase the odds of people social distancing? (causal BN)
3. Can we hone public health interventions by identifying population segments with different barriers to and beliefs around social distancing? (k-means clustering)

Our predictive models revealed that individuals’ perceptions and beliefs (e.g., risk perception, social norms, self-efficacy) are significant predictors of social distancing behavior. While many polls and studies of social distancing have focused primarily on how demographic characteristics relate to behavior (Murad, 2020; Pew Research, 2020), our predictive model included key perceptual drivers of and structural barriers to social distancing. This may explain why factors such as political partisanship and age that have been found to be important differentiators of social distancing in various polls and univariate analyses were not significant predictors of behavior in our study (Allcott et al., 2020; Murad, 2020). For example, unlike other surveys of social distancing (Pew Research, 2020), we did not find a significant correlation between social distancing behavior and political affiliation. However, we did find that Republicans were overrepresented in low-social-distancing population segments.

We used the causal BN model to gain a more system-level view of the causal cascade to social distancing in the United States. While our predictive model revealed a set of 10 variables correlated with social distancing behavior, our BN model suggested which variables were the most direct causal drivers of this behavior. We found that social distancing intentions and behavior were conditionally dependent on seeking out information about coronavirus, being more worried about coronavirus, and having higher financial security. Interestingly, while our predictive model found significant correlations between social distancing and individuals' risk perception, social norms, and self-efficacy, our causal model showed that these variables are downstream correlates of social distancing behaviors and intent, rather than their causes.

Like our predictive model, our BN showed that demographics (other than age group) had weak, if any, causal effects on social distancing outcomes. While gender and political affiliation were causally linked to degree of worry about social distancing, neither of these variables was a direct driver of social distancing behavior or intent. Race and religion were also not direct causal drivers of social distancing behavior or intent. Messages focusing on these aspects of social distancing may not be as effective. It is worth noting that the media coverage of the political and demographic landscape of the pandemic has been very dynamic since March 2020, which could affect the beliefs and behaviors of individuals, and in turn, the applicability of these conclusions to the current environment. The causal model also suggests that lack of financial security was an important barrier to social distancing. Accordingly, policies that work to reduce financial barriers, such as increasing unemployment benefits, could be important tools to increase social distancing. Conversely, delays in financial support may have discouraged social distancing.

The k-means clustering analysis helped to identify four population segments best differentiated by their risk perception, financial stability, beliefs about community social norms, self-efficacy, and information-seeking about social distancing. This analysis could help hone messaging by identifying population segments with different barriers to social distancing.

The combination of the insights generated from the three approaches allow us to recommend a series of public health interventions targeting low-social-distancing segments of the population with the right messages through the right messengers. Public health officials may choose to implement only one or a subset of results, but they can be empowered to know the causal drivers behind social distancing and the segments of the population they are likely to reach with different interventions. The strategic content suggested by our analysis should be delivered with best practices: use nontechnical, clear, and consistent messaging from trusted information sources that provides actionable messages and tangible examples (Lunn et al., 2020; World Health Organization, 2017).

5. Next Steps and Limitations

The development of tailored interventions needs to be paired with effective campaigns to translate insights into action. We believe there are a number of pathways to implement targeted interventions, whether at the

individual or community scale. First, segmentation results can be further analyzed to develop ‘typing tools’—short decision trees, with the same questions used in segmentation—that can quickly and accurately classify individuals into different social distancing segments and then engage them in real time with tailored messages and interventions (Sgaier et al., 2018). Typing tools may be carried out on tech platforms (e.g., a short online survey) or in person.

Perhaps counterintuitively, mass media campaigns are another pathway to target different population segments. Tailored mass media messages may be developed based on our understanding of the beliefs and barriers of different population segments. These messages may be further personalized by using credible messengers that align with the profiles of each segment. Having a broad array of targeted messages can help target particular audiences without reaching out to individuals. As noted above, mass media campaigns may also choose to focus only on the causal drivers that are most likely to make a difference for the largest segments of the population.

Our research has several limitations. First, this is a cross-sectional study reporting on beliefs, perceptions, and behaviors around social distancing in late March 2020, but the context of the novel coronavirus is rapidly evolving. Since our survey was conducted, there have been significant developments in the disease trajectory of COVID-19, as well as local, state, national, and international interventions. Second, all social distancing behavior is self-reported and subject to social desirability bias, which may mean that we have overestimated rates of social distancing uptake. Third, while we have identified key potential causal drivers of social distancing, as well as population segments to target in order to increase social distancing, applying and testing the efficacy of behavioral interventions remains a challenge. Applying specific actions to reach population segments is often highly dependent on local or even hyperlocal contexts and requires buy-in from community leaders. Fourth, while our survey is nationally representative, we did not have enough statistical power to estimate social distancing rates at subregional levels such as state or zip code.

Finally, all the models come with assumptions; for BN, there are several. First, the data requirements for causal BN models are more stringent for causal models than for regression-type predictive models, including the need for a much bigger sample size to produce reliable estimates of the dependencies between variables from their distributions. This may be difficult to achieve in certain situations. Second, while BN excels in modeling observed confounding, in some circumstances, the directions of the dependencies in the model may be misleading if the training data does not include variables that are causal to both the outcome of interest and variables upstream. Therefore, for this approach it is crucial that researchers ensure that relevant variables are included in the training data as much as possible. This is why we used a validated framework, CUBES, in order to systematically capture as many potential causal drivers of social distancing as possible. Third, many traditional statistical models such as regression estimate easily interpretable goodness-of-fit measures such as R^2 . Unlike prediction models, where we can assess the error rate of the predicted values of the outcome variable against the observed, with a causal model we are interested in accuracy of recovering the latent

relationships between variables in the data. However, because the true cause and effect relationships between different factors *are* latent (i.e., the ‘ground truth’ is not known), our causal machine learning performance can only be inferred by estimating the performance of algorithms on recovering known causal structure of synthetic data sets with characteristics that are similar to the data set that is being modeled (see details in Appendix E). The performance inference work is detailed in Butcher et al., 2021.

6. Conclusion

Social distancing behavior has been and will continue to remain a salient non-pharmaceutical infection prevention and control intervention. Behavioral drivers and perceptions are key to understanding uptake of social distancing, as awareness alone is inadequate to stir action. Further, social distancing behavior is heterogeneous; distinct segments exist in the population that vary in their behavioral drivers and uptake of social distancing. Causal machine learning can help policymakers design precision public health interventions that identify which groups to target and with what messages. More broadly, we believe that public health interventions and communication campaigns that incorporate data-driven insights, are more tailored to target different priority segments in the population, and leverage trusted sources like local governments and health care providers, could yield considerable gains in the uptake of desirable protective behaviors.

Acknowledgments

The authors thank James Baer, Rahul Joseph, Danielle Schmutz, Dr. Peter Smittenaar, and Dr. Neela Saldanha for their insightful comments on this manuscript.

Disclosure Statement

Grace Charles, Mokshada Jain, Yael Caplan, Hannah Kemp, Aysha Keisler, Vincent Huang, and Sema K. Sgaier have no financial or non-financial disclosures to share for this article.

Data Repository/Code

The source data frame and the R analysis script are available on Harvard Dataverse:

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/RBZONT>. The structure-learning algorithm is a proprietary implementation of OrderMCMC and qNML; access can be granted from the authors on a per-request basis.

References

Abouk, R., & Heydari, B. (2020). The immediate effect of COVID-19 policies on social distancing behavior in the United States. *SSRN*. <https://doi.org/10.2139/ssrn.3571421>

Allcott, H., Boxell, L., Conway, J., Ferguson, B., Gentzkow, M., & Goldman, B. (2020). *What explains temporal and geographic variation in the early US Coronavirus pandemic?* (Working Paper No. w27965). National Bureau of Economic Research. <https://doi.org/10.3386/w27965>

Bish, A., & Michie, S. (2010). Demographic and attitudinal determinants of protective behaviours during a pandemic: A review. *British Journal of Health Psychology*, 15(Part 4), 797–824. <https://doi.org/10.1348/135910710x485826>

Butcher, B., Huang, V., Robinson, C., Reffin, J., Sgaier, S., Charles, G., & Quadrianto, N. (2021). Causal datasheet for datasets: An evaluation guide for real-world data analysis and data collection design using Bayesian Networks. *Frontiers in Artificial Intelligence*, 4, Article 612551. <https://doi.org/10.3389/frai.2021.612551>

Centers for Disease Control and Prevention. (2020, February 11). *COVID-19 and your health*. <https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html>

Chickering, D. M., Heckerman, D., & Meek, C. (2004). Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5, 1287–1330. <https://www.jmlr.org/papers/volume5/chickering04a/chickering04a.pdf>

Chowkwanyun, M., Bayer, R., & Galea, S. (2018). “Precision” public health—Between novelty and hype. *New England Journal of Medicine*, 379(15), 1398–1400. <https://doi.org/10.1056/NEJMp1806634>

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., & Leskovec, J. (2021). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840), 82–87. <https://doi.org/10.1038/s41586-020-2923-3>

de Mendiburu, F. (2021). agricolae: Statistical Procedures for Agricultural Research. R package version 1.3-5. <https://CRAN.R-project.org/package=agricolae>

Engl, E., & Sgaier, S. K. (2020). CUBES: A practical toolkit to measure enablers and barriers to behavior for effective intervention design. *Gates Open Research*, 3(886). <https://doi.org/10.12688/gatesopenres.12923.2>

Engl, E., Smittenaar, P., & Sgaier, S. (2019). Identifying population segments for effective intervention design and targeting using unsupervised machine learning: An end-to-end guide. *Gates Open Research*, 3(1503). <https://doi.org/10.12688/gatesopenres.13029.2>

Glass, R. J., Glass, L. M., Beyeler, W. E., & Min, H. J. (2006). Targeted social distancing designs for pandemic influenza. *Emerging Infectious Disease Journal*, 12(11), Article 1671. <https://doi.org/10.3201/eid1211.060255>

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10(524), Article 524. <https://doi.org/10.3389/fgene.2019.00524>

Holiday Travel Survey 2020—Thanksgiving and Christmas. (2020, November 17). The Vacationer. <https://thevacationer.com/holiday-travel-survey-2020/>

Huang, V. S., Sutermeister, S., Caplan, Y., Kemp, H., Schmutz, D., & Sgaier, S. K. (2020). Social distancing across vulnerability, race, politics, and employment: How different Americans changed behaviors before and after major COVID-19 policy announcements. *medRxiv*. <https://doi.org/10.1101/2020.06.04.20119131>

Ide, J. S., & Cozman, F. G. (2002). Random generation of Bayesian networks. In G. Bittencourt, & G. L. Ramalho (Eds.), *Advances in artificial intelligence* (pp. 366–376). Springer. https://doi.org/10.1007/3-540-36127-8_35

Ipsos. (2020). *Knowledge panel information* [Web Page]. <https://www.ipsos.com/en-us/solutions/public-affairs/knowledgepanel>

Kassambara, A., & Mundt, F. (2020). factoextra: Extract and visualize the results of multivariate data analyses. R package version 1.0.7. <https://CRAN.R-project.org/package=factoextra>

Khoury, M. J., Iademarco, M. F., & Riley, W. T. (2016). Precision public health for the era of precision medicine. *American Journal of Preventive Medicine*, 50(3), 398–401. <https://doi.org/10.1016/j.amepre.2015.08.031>

Kleczkowski, A., Maharaj, S., Rasmussen, S., Williams, L., & Cairns, N. (2015). Spontaneous social distancing in response to a simulated epidemic: A virtual experiment. *BMC Public Health*, 15, Article 973. <https://doi.org/10.1186/s12889-015-2336-7>

Lin, L., McCloud, R. F., Jung, M., & Viswanath, K. (2018). Facing a health threat in a complex information environment: A national representative survey examining American adults' behavioral responses to the 2009/2010 a(H1N1) pandemic. *Health Education & Behavior*, 45(1), 77–89. <https://doi.org/10.1177/1090198117708011>

Lunn, P., Belton, C., Lavin, C., McGowan, F., Timmons, Shane, & Robertson, D. (2020). *Using behavioural science to help fight the coronavirus* [Report]. Behavioural Research Unit, Environmental Systems Research Institute (Esri).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (pp. 281–297). University of California Press

Mehta, S. H., Clipman, S. J., Wesolowski, A., & Solomon, S. S. (2020). “There’s no place like home for the holidays:” Travel and SARS-CoV-2 test positivity following Thanksgiving weekend. *medRxiv*.

<https://doi.org/10.1101/2020.12.22.20248719>

Murad, Y. (2020, March 20). *Most U.S. adults practice some degree of social distancing amid coronavirus spread*. Morning Consult. <https://morningconsult.com/2020/03/20/coronavirus-social-distancing-poll/>

Pearl, J. (1995). From Bayesian networks to causal networks. In G. Coletti, D. Dubois, & R. Scozzafava (Eds.), *Mathematical models for handling partial knowledge in artificial intelligence* (pp. 157–182). Springer US.

https://doi.org/10.1007/978-1-4899-1424-8_9

Pearl, J. (2009). *Causality: Models, reasoning and inference*. Cambridge University Press.

<https://doi.org/10.1017/CBO9780511803161>

Pew Research, C. (2020). *Republicans, Democrats move even further apart in Coronavirus concerns* [Report].

<https://tinyurl.com/pew-research-poll>

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Rashid, H., Ridda, I., King, C., Begun, M., Tekin, H., Wood, J. G., & Booy, R. (2015). Evidence compendium and advice on social distancing and other related measures for response to an influenza pandemic. *Paediatric Respiratory Reviews*, 16(2), 119–126. <https://doi.org/10.1016/j.prrv.2014.01.003>

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

<https://doi.org/10.1214/aos/1176344136>

Sgaier, S. K., Engl, E., & Kretschmer, S. (2018). Time to scale psycho-behavioral segmentation in global development. *Stanford Social Innovation Review*, Fall. <https://doi.org/10.48558/ghdj-w903>

Teasdale, E., Santer, M., Geraghty, A. W., Little, P., & Yardley, L. (2014). Public perceptions of non-pharmaceutical interventions for reducing transmission of respiratory infection: Systematic review and synthesis of qualitative studies. *BMC Public Health*, 14, Article 589. <https://doi.org/10.1186/1471-2458-14-589>

Tracy, C. S., Rea, E., & Upshur, R. E. (2009). Public perceptions of quarantine: Community-based telephone survey following an infectious disease outbreak. *BMC Public Health*, 9, Article 470.

<https://doi.org/10.1186/1471-2458-9-470>

Unacast. (2020). *Social distancing scoreboard*. <https://www.unacast.com/covid19/social-distancing-scoreboard>

Wickham, H., & François, R., Henry, L., & Müller K. (2021). dplyr: A Grammar of data manipulation. R package version 1.0.6. <https://CRAN.R-project.org/package=dplyr>

World Health Organization. (2017). *Communicating risk in public health emergencies: A WHO guideline for emergency risk communication (ERC) policy and practice*.

Zhang, X., Wang, F., Zhu, C., & Wang, Z. (2019). Willingness to self-isolate when facing a pandemic risk: Model, empirical test, and policy recommendations. *International Journal of Environmental Research and Public Health*, 17(1), Article 197. <https://doi.org/10.3390/ijerph17010197>

Appendix

Appendix A

Text of Social Distancing Survey

Social Distancing Adherence: Developing a Model of Behavior Change

The purpose of this survey is to understand opinions and actions related to the coronavirus disease (COVID-19), the disease that is spreading in the US right now. The survey is being conducted by the [blinded] The survey will take about 15 minutes.

During the survey we will ask you questions about your behaviors and opinions regarding the coronavirus. There are no right or wrong answers to any of the questions, we simply want to understand your opinion. Your responses will be kept secure, confidential and anonymous. Your responses will be used for research purposes only, and only researchers associated with the study will have access to your responses.

You may quit the survey at any time. If you have any questions about this study, please reach out to the contact below.

[blinded]

If you have any questions about the coronavirus, please refer to the CDC

<https://www.cdc.gov/coronavirus/2019-ncov/index.html>

Do you wish to continue on to the survey?

No [THANK AND TERMINATE]

Yes [CONTINUE]

1. Thinking about the current coronavirus situation, which of these actions do **YOU** think are OK for **YOU** to do right now? Please check all that apply.

Going to your regular trivia night at the bar

Taking public transportation, such as subways, buses, etc.
Going for a walk outside by yourself
Standing close to people in a grocery line
Going to the grocery store on a limited basis to get food
Socializing with a small group of healthy friends outdoors
Going to an indoor event of 500 people
Going to a dinner party of 15 people if you feel healthy
Going to a dinner party of 15 people only if you wash your hands before you go and immediately after you get home.
Going to a party while you have a cough so long as you wash your hands often
Hanging out with friends indoors that do not have coronavirus symptoms
Going into the office as usual if you are under 30 years of age

You’ve probably heard people talk about “social distancing.” **Social distancing means to remain away from both small and large groups and stay at least six feet away from others (not in your household) when possible to slow or stop the spread of a disease. For many people, this means staying at home as much as possible.**

2. How often are you CURRENTLY practicing the following social distancing behaviors?

1	2	3	4	5	6	7
Never			Sometimes			Always

Avoiding groups of people outside of your home?

Staying home as much as possible, with the exception of solo outdoor activities?

Maintaining physical distance from people, with the goal of staying at least 6 feet from others?

3. People may *want to* take some of the social distancing precautions we listed, but they are unable to due to work and other factors. How often would you *ideally like* to take the following precautions related to the coronavirus?

1	2	3	4	5	6	7
Never			Sometimes			Always

Avoiding groups of people outside your home?

Staying home as much as possible, with the exception of solo outdoor activities?

Maintaining physical distance from people, with the goal of staying at least 6 feet from others?

4. Please rate your agreement with the following statements.

1	2	3	4	5	6	7
Strongly Disagree			Neutral			Strongly Agree

I understand what I should do to prevent MYSELF AND MY FAMILY from getting sick from coronavirus.

I understand what I should do to prevent MY COMMUNITY from getting sick from coronavirus.

I understand what people that are particularly vulnerable (e.g. older adults, those with certain underlying medical conditions) should do to prevent getting sick from coronavirus.

I understand why health officials tell us to remain **away from groups and stay at least six feet away from others when possible** during the coronavirus pandemic.

I understand why health officials tell us to stay home as much as possible during the coronavirus pandemic.

5. Please rate your agreement with the following statements.

1	2	3	4	5	6	7
Never			Sometimes			Always

My friends and family are doing everything they can to remain away from groups and stay at least six feet away from others when possible due to coronavirus.

People in my community are doing everything they can to remain away from groups and stay at least six feet away from others when possible to coronavirus.

6. Please rate your agreement with the following statements about your work and life right now.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

Strongly Disagree			Neutral			Strongly Agree
----------------------	--	--	---------	--	--	----------------

I am able to work from home (telework) right now.

I receive income from my employer even if the coronavirus prevents me from working right now.

I am able to pay all my bills for the next two months.

I have access to a home where I can stay for the next two months.

I have control over who comes and goes from the home that I am in.

7. Now tell us how **likely** you think it is the coronavirus will impact the following.

1	2	3	4	5	6	7
Not at all likely						Very Likely

Your physical health

Your mental health

Your close friends' and family's health

Your financial stability

The American economy

8. Please tell us how **severely** you think coronavirus could impact the following.

1	2	3	4	5	6	7
No impact			Some impact			Severe Impact

Your physical health

Your mental health

Your close friends' and family's health

Your financial stability

The American economy

9. What do you think the impact is of **YOU going out and/or interacting with others outside your home** right now? Specifically, what is the impact on...

1	2	3	4	5	6	7	8	9
---	---	---	---	---	---	---	---	---

Very Negative Impact		Somewhat Negative impact		No impact		Somewhat Positive Impact		Very Positive Impact
----------------------------	--	--------------------------------	--	-----------	--	--------------------------------	--	----------------------------

Your physical health?

Your mental health?

Your close friends' and family's health?

Your financial stability?

Your community's economy?

10. How **likely** is it that **YOU going out and/or interacting with others outside your home** will have an impact on....

1	2	3	4	5	6	7
Not at all likely						Very Likely

Your health?

Your mental health?

Your close friends' and family's health?

Your financial stability?

Your community's economy?

11. How difficult are the following aspects of remaining **away from groups and staying at least six feet away from others when possible**?

1	2	3	4	5	6	7	N/A
Not Difficult						Very Difficult	

Feeling socially isolated and/or lonely at home?

Not physically leaving the house often?

Not attending any religious or community gatherings in person?

Managing essential day-to-day tasks, such as getting and making food, going to the pharmacy for medication?

Not receiving in-person services such as hair and nail salon services, laundry service, bank visits, etc.?

Postponing in person healthcare visits?

Managing childcare?

Making decisions about how I will protect my and my family's health?

12. How much are you keeping up with information about the coronavirus?

1	2	3	4	5	6	7
Not at all			Some			A lot

13. How much information, guidance, and opinions about the coronavirus **do you** hear from the following sources?

1	2	3	4	5	N/A
None		Some		All	

Traditional media sources, such as newspapers (print and online)

Social media

Friends and family

President Trump

US Congress

Your state and local government

Your religious leaders (pastor, priest, rabbi, etc.)

Your employer

Your healthcare providers

National and global health organizations (WHO, CDC, NIH, etc.)

14. Now we'll show you those sources of information again. Tell us how much you **TRUST** each source of information about coronavirus.

1	2	3	4	5	6	7	N/A
No trust			Some trust			A lot of trust	

Traditional media sources, such as newspapers (print and online)

Social media

Friends and family

President Trump

US Congress

Your state and local government

Your religious leaders (pastor, priest, rabbi, etc.)

Your employer

Your healthcare providers

National and global health organizations (WHO, CDC, NIH, etc.)

15. Overall, is the information you're getting from these sources the same or different?

1	2	3	4	5	6	7
Very different information from sources			Somewhat different information from sources			Very Similar information from sources

16. Please rate your level of agreement with the following statements.

1	2	3	4	5	6	7
Strongly Disagree			Neutral			Strongly Agree

I trust there are safety nets that are in place now to protect me and my family if we stay home to ensure we have food and basic services

If I experience financial hardship because of social distancing, I trust that the government will help

17. How much uncertainty are you feeling right now about...

1	2	3	4	5	6	7	N/A
A lot of uncertainty						No uncertainty	

Your own and your family's future health?

Your community's future health?

Your own and your family's future financial stability?

America's future financial stability?

18. Overall, how confident are you in the government's response to the coronavirus?

1	2	3	4	5	6	7
Not at all confident			Somewhat confident			Very Confident

19. Please rate your health.

	1	2	3	4	5
	Poor	Fair	Good	Very Good	Excellent
Physical Health					
Mental Health					

20. Please rate how anxious or worried you've been because of coronavirus over the past three days.

1	2	3	4	5	6	7
Not anxious at all			Somewhat anxious			Very anxious

21. Please rate your agreement with the following statements.

1	2	3	4	5	6	7
Strongly disagree			Neutral			Strongly agree

I feel like I can help reduce the spread of the coronavirus in my community.

My actions have an impact on the spread of the coronavirus in my community.

I am someone who takes risks in life.

I am someone who “plays it safe” in life.

22. How important is your own well-being compared to the well-being of others around you?

1	2	3	4	5	6	7
My well-being is much more important			My well-being and others' wellbeing is equally important			Others' well-being is much more important

23. How confident are you filling out medical forms by yourself?

1	2	3	4	5
Not at all confident	Somewhat confident	Neutral	Very confident	Extremely confident

24. Select all that apply:

1. You think you might currently have or had the coronavirus
2. You have been diagnosed with the coronavirus
3. You know someone personally that has or had the coronavirus
4. You have heard of cases of the coronavirus in your community

Thank you for sharing your opinion with us! Your participation is very important as we study the health and wellbeing of Americans.

If you have any questions about the survey, please contact

[blinded]

If you have any questions about the coronavirus, please refer to the CDC

<https://www.cdc.gov/coronavirus/2019-ncov/index.html>

Appendix B

Technical Appendix.

Predictive Model specification variable descriptions and weights

Number of observations: Original observations = 2,500, 201 observations removed due to missing predictors, final sample size = 2,299.

Model specification: Generalized linear model, response = social distancing behavior, predictors = all listed

below.

Model type = binomial, weighted by survey Weights.

Variable description in model	Question(s) from survey	Question response range or categories	Variable calculation, if any
Social distancing behavior (response variable)	How often are you CURRENTLY practicing the following social distancing behaviors? a. Avoiding groups of people outside of your home? b. Staying home as much as possible, with the exception of solo outdoor activities? c. Maintaining physical distance from people, with the goal of staying at least 6 feet from others?	1 (never) – 7 (always) for each sub-question	Respondent <i>always</i> following social distancing behavior if answered 7 (always) to all three questions
Self-efficacy in stopping spread of virus	Please rate your agreement with the following statements a. I feel like I can help reduce the spread of the coronavirus in my community. b. My actions have an impact on the spread of the coronavirus in my community.	1 (strongly disagree) – 7 (strongly agree) for each sub-question	Mean of responses to all sub-questions. All continuous variables were standardized so that the odds ratio reflects the difference in shifting one standard deviation on that variable

Social distancing risk perception	<p>How likely is it that YOU going out and/or interacting with others outside your home will have an impact on...</p> <p>a. Your health</p> <p>b. Your close friends' and family's health</p>	1 (not at all likely) to 7 (very likely) for each sub question	Mean of responses to all sub-questions. All continuous variables were standardized so that the odds ratio reflects the difference in shifting one standard deviation on that variable
Degree of information seeking	How much are you keeping up with information about the coronavirus?	1 (not at all) – 7 (a lot)	All continuous variables were standardized so that the odds ratio reflects the difference in shifting one standard deviation on that variable
Degree of financial stability	<p>Please rate your agreement with the following statements about your work and life right now</p> <p>a. I am able to work from home (telework) right now.</p> <p>b. I receive income from my employer even if the coronavirus prevents me from working right now.</p> <p>c. I am able to pay all my bills for the next two months</p>	1 (strongly disagree) – 7 (strongly agree)	Mean of responses to all sub-questions. All continuous variables were standardized so that the odds ratio reflects the difference in shifting one standard deviation on that variable
Perception: is community following social distancing?	<p>Please rate your agreement with the following statements:</p> <p>People in my community are doing everything they can to remain away from groups and stay at least six feet away from others when possible to coronavirus.</p>	1 (never) – 7 (always)	All continuous variables were standardized so that the odds ratio reflects the difference in shifting one standard deviation on that variable

Degree of worry about Coronavirus	Please rate how anxious or worried you've been because of coronavirus over the past three days.	1 (not anxious at all) – 7 (very anxious)	All continuous variables were standardized so that the odds ratio reflects the difference in shifting one standard deviation on that variable
Degree of trust in information sources	<p>Tell us how much you TRUST each source of information about coronavirus</p> <p>a. Traditional media sources, such as newspapers (print and online)</p> <p>b. Social media</p> <p>c. Friends and family</p> <p>d. President Trump</p> <p>e. US Congress</p> <p>f. Your state and local government</p> <p>g. Your religious leaders (pastor, priest, rabbi, etc.)</p> <p>h. Your employer</p> <p>i. Your healthcare providers</p> <p>j. National and global health organizations (WHO, CDC, NIH, etc.)</p>	1 (no trust) – 7 (a lot of trust)	Mean of responses to all sub-questions. All continuous variables were standardized so that the odds ratio reflects the difference in shifting one standard deviation on that variable

High risk of COVID-19	Calculated based on responses to demographic and health questions	High risk / normal risk	Respondent was categorized as high-risk if any of the following criteria were met: a. Age 65+ b. Diabetic c. BMI > 40 d. Heart disease or heart condition e. Asthma, bronchitis, or COPD
-----------------------	---	-------------------------	---

Perceived difficulty of social distancing	<p>How difficult are the following aspects of remaining away from groups and staying at least six feet away from others when possible?</p> <p>a. Feeling socially isolated and/or lonely at home?</p> <p>b. Not physically leaving the house often?</p> <p>c. Not attending any religious or community gatherings in person?</p> <p>d. Managing essential day-to-day tasks, such as getting and making food, going to the pharmacy for medication?</p> <p>e. Not receiving in-person services such as hair and nail salon services, laundry service, bank visits, etc.?</p> <p>f. Postponing in person healthcare visits?</p> <p>g. Managing childcare?</p> <p>h. Making decisions about how I will protect my and my family's health?</p>	1 (not difficult) – 7 (very difficult)	Mean of responses to all sub-questions. All continuous variables were standardized so that the odds ratio reflects the difference in shifting one standard deviation on that variable
---	---	--	---

Context: aware of someone with COVID-19	Select all that apply: a. You think you might currently have or had the coronavirus b. You have been diagnosed with the coronavirus c. You know someone personally that has or had the coronavirus	Yes/No	If respondent indicated ‘yes’ to any sub question, yes, otherwise no.
Currently working or retired	Current employment status	Working, retired, not working	Currently working or retired grouped together vs. unemployed
Religion	What is your religion	Evangelical/Protestant, Catholic, Jewish, No religion, other religion	
Geographic region	Based on respondent geographic information	Northeast, Midwest, south, west	
Education	What is your highest level of education?	Less than high school, high school, some college, bachelor’s degree or higher	
Race	Self-reported race	Black/African American, White, Asian American, Other	
Age	Self-reported age	Gen Z Millennial Gen X Boomer +	Gen Z: age 18- 23 Millennial: age 24 – 39 Gen X: age 40 – 55 Boomer +: age 56+
Political partisanship	Self-reported political party affiliation	-3 (strong Republican) – 0 (independent) – 3 (strong Democrat)	
Gender	Self-reported gender	Male, Female	

Characteristics and responses of KnowledgePanel members who completed the survey between March 27, 2020 and March 31, 2020. All percentages weighted according to survey means.

Characteristic	Weighted No. (%) ^a
Age	
Generation Z (age 18-23)	175 (7)
Millennial (age 24-39)	759 (30.4)
Generation X (age 40-55)	600 (24)
Boomers + (age 56+)	966 (38.6)
Gender	
Female	1290 (51.6)
Male	1210 (48.4)
Education	
Less than high school	265 (20.6)
High school	708 (28.3)
Some college	695 (27.8)
Bachelor's degree or higher	832 (33.3)
Race/Ethnicity	
White	1580 (63.2)
Black or African American	306 (12.2)
Latino/Hispanic	376 (14.7)
Other race/multiple races	247 (10)

Income category	
<\$30,000	338 (13.5)
\$30,000–\$99,999	1209 (48.4)
\$100,000+	953 (38.1)
Work category	
Not working	713 (28.5)
Working	1787 (71.5)
Religion	
Catholic	600 (24)
Evangelical/Protestant	845 (22.8)
Jewish	43 (1.7)
No religion	577 (23.1)
Other religion	436 (17.4)
Political partisanship (continuous scale, -3 = strong Republican, 3 = strong Democrat)	Mean = 0.24
Geographic region	
Midwest	520 (20.8)
Northeast	437 (17.5)
South	948 (37.9)
West	595 (23.8)

^aNumber and percentage weighted based on survey weights.

Appendix C

Social Distancing Segment Profiles and Differences

Characteristics of the four population segments and profiling variable means were identified. All values of characteristics were normalized to a 0–1 scale, with 0 being the lowest possible value and 1 being the highest possible value. Segment variables represent variables used in k-means calculations of segments. Profiling variables represent mean value or percentage for all individuals in each segment. Tukey post hoc pairwise tests show significant differences between segments where $a > b > c > d$, with $p < .05$ for each step. Segments with the same letter are not significantly different

	Segment 1	Segment 2	Segment 3	Segment 4
Segment size (% of survey)	697 (28.5%)	666 (27.3%)	509 (20.8%)	572 (23.4%)
Segment variables (variables used to calculate segments)				
Perception of high community norms of social distancing	0.84 (a)	0.85 (a)	0.38 (c)	0.64 (b)
Risk perception of social distancing impact on health	0.68 (a)	0.37 (c)	0.62 (b)	0.33 (d)
Perceived difficulty of social distancing	0.53 (a)	0.23 (d)	0.36 (b)	0.30 (c)
Degree of information seeking	0.88 (ab)	0.89 (a)	0.86 (b)	0.69 (c)
Mean worry about coronavirus	0.73 (a)	0.29 (c)	0.55 (b)	0.29(c)
Perceived self-efficacy in preventing spread of coronavirus	0.77 (b)	0.85 (a)	0.80 (b)	0.47(c)
Social distancing profiling variables (mean values for each segment)				

Social distancing uptake (% , survey mean = 46.2%)	55.3 (a)	53.8 (a)	44.1 (b)	25.9 (c)
Social distancing intent (%)	71.1 (a)	67.8 (a)	73.1 (a)	36.6 (b)
Social distancing knowledge (%)	90.8 (a)	89.9 (a)	91.0 (a)	77.5 (b)
Demographic profiling variables (mean values for each segment)				
Average age	47.4 (b)	52.0 (a)	44.2 (c)	47 (b)
Gender: Male (%)	39.4 (c)	51.8 (b)	47.1 (bc)	57.5 (a)
Political party: Republican (%)	31.5 (b)	46.3 (a)	37.8 (b)	49.7 (a)
Political party: Democrat (%)	64.4 (a)	46.6 (b)	56.6 (a)	44.1 (b)
Religion: Evangelical (%)	28.1 (b)	36 (a)	32.6 (ab)	38.4 (a)
Religion: Catholic	31.4 (a)	23.7 (b)	19.3 (b)	21.7 (b)
Religion: No religion (%)	19.4 (b)	22.7 (ab)	26.7 (a)	25.1 (ab)
Income: \$100,000 + (%)	35.9 (b)	43.4 (a)	37.7 (ab)	37.1 (ab)
Income: \$30,000 - \$99,999 (%)	49.0 (a)	46.8 (a)	46.7(a)	50.1(a)
Income: <\$30,000	15.2 (a)	9.9 (b)	15.6 (a)	12.8 (ab)
Region: Northeast	19.5 (a)	19.7 (a)	13.6 (c)	15.4 (bc)
Region: Midwest	19.9 (a)	20.4 (a)	22.1 (a)	21.0 (a)
Region: South	36.0 (a)	35.1 (a)	41.2 (a)	40.0 (a)

Region: West	24.6 (a)	24.7 (a)	23.0 (a)	23.6 (a)
---------------------	----------	----------	----------	----------

Appendix D

Interventional Performance Calculation for Causal Bayesian Network

One of the key uses of a causal Bayesian Network model is that, for a given outcome variable of interest, one can test hypothetical interventions on each variable. One can then compute the interventional odds ratio (OR) of how the outcome may change based on the intervention.

$$OR = \frac{P(A = a_1 | do(B = b_1))}{P(A = a_1 | do(B = b_2))} \bigg/ \frac{P(A = a_2 | do(B = b_1))}{P(A = a_2 | do(B = b_2))} = \frac{a}{c} \bigg/ \frac{b}{d}$$

The results of this intervention encompass both the causal structure learned and the parameters (estimated by Maximum Likelihood) of the conditional probability tables at each variable. We calculate the standard error for the odds ratios by:

$$SE = \frac{1}{a'} + \frac{1}{b'} + \frac{1}{c'} + \frac{1}{d'}$$

where a' is the frequency obtained from the training data for term a ; $P(A = a_1 | do(B = b_1))$. 95% Confidence intervals are then obtained by $\log(OR) \pm \sqrt{1.96} \cdot SE$

Appendix E

Interventional Performance Calculation for Causal Bayesian Network

To estimate the confidence in the causal BN results, we ran the same BN structure learning algorithms on 50 synthetic data sets with similar characteristics, including the sample size ($n = 2,500$), number of variables ($n = 17$), and similar average number of variable levels ($n = 3$). The key difference between the synthetic data sets and our real survey data set is that we know the ‘ground-truth’ (i.e., the actual causal pathways between the variables) of the synthetic data sets. This means we can infer how well we expect our structure learning algorithms will perform for data sets with those characteristics.

Our evaluation of synthetic data found that we could expect to recover most of the ‘skeleton’ of our DAG; meaning that, for this set of data set characteristics, a BN does a good job of recovering all the relationships between variables, but does not say anything about the direction (i.e., causality) of relationships between variables. We found that mean skeleton recall was quite high (0.98 on a linear scale from 0 to 1), meaning that

we could expect to recover, on average, 98% of correct associative relationships (e.g., undirected edges) between variables. We found that skeleton precision was only 0.69, meaning that, on average, around 69% of all undirected edges would be correct (this metric accounts for both false positives and false negatives). When we consider the directed, causal relationships between variables, we found that DAG recall was around 0.82 and DAG precision was around 0.58. This means that we could expect to find around 82% of directed relationships between variables. In addition, after accounting for false positives and false negatives, we could expect that around 58% of all directed relationships we predicted would be correct.

This performance evaluation has a few implications for our work. First, we can be relatively confident that the relationships between variables are accurate. However, because of our low expected DAG precision, we should be careful about our assertions around causality. In order to improve expected model performance, we could expand our sample size. For example, increasing the sample size to 5,000 would increase DAG recall to 0.86 and DAG precision to 0.82.

Appendix F

Causal Discovery and Causal Inference With a Causal Bayesian Network

It has been shown that conditional dependencies between variables can be represented by a directed acyclic graph (DAG) in an approach called the causal Bayesian network (BN) (Pearl, 1995, 2009). Specifically, Bayesian networks are probabilistic graphical models that leverage these conditional dependencies to model causation. For the purpose of this article, by causal we mean conditional dependencies as defined by Pearl et al. The structure of a DAG is denoted by a set of nodes (i.e., representing the variables) and edges (i.e., representing the causal connections between the nodes). The absence of an edge between a given pair of nodes suggests a conditional independence between the variables the nodes represent. In a causal BN, a node A is deemed directly causal of a node C if changing the value of A (i.e., intervention) without changing values of other nodes affects the distribution of C :

$$P(do(A = a_1), do(B = b_1)) \neq P(do(A = a_2), do(B = b_1))$$

where $P(do(A=a_1))$ is the interventional distribution. The $do()$ is known as the do-operator; $do(A=a_1)$ means setting the value of variable A to a_1 .

Accompanying each node is a set of parameters that specify the probability of that node taking on a value (e.g., social distancing behavior = TRUE). For discrete or categorical variables, the parameters are represented as conditional probability tables (CPTs). In other words, the underlying ‘causal’ ordering of factors is identified through the structural output of the DAG. The DAG shows which variables are directly or indirectly causal of the outcome of interest, which are causal through upstream pathways, and which are outside the causal chain.

Here we used a data-driven approach by using structural learning algorithms to construct the DAG from the observational data set. To generate the causal graph and perform causal inference, a variation of the Markov Chain Monte Carlo method in the topological order space with the quotient normalized maximum likelihood (qNML) score function was used in a custom implementation here (Butcher et al., 2021). The qNML score is a recent development that is like the Bayesian information criterion (BIC) (Schwarz, 1978). Both are based on the minimum description length principle such that the less complex model is preferred given the same information provided. Both are score equivalent and hyperparameter-free. We chose qNML because in our empirical analyses, we found qNML to give superior results when recovering causal structures in synthetic data sets that had statistical properties similar to our survey data (Butcher et al., 2021). It should be noted, however, that a variety of algorithms (e.g., Peter-Clark, Greedy Equivalence Search, etc.) are also available for other types of data sets with different assumptions of underlying ground truth processes that produce the real-life observations and data (Glymour et al., 2019).

©2022 Grace Charles, Mokshada Jain, Yael Caplan, Hannah Kemp, Aysha Keisler, Vincent Huang, and Sema K. Sgaier. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.