Vaccine 40 (2022) 1783–1789

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Towards a predictive model of COVID-19 vaccine hesitancy among American adults

Jack Mewhirter *, Mustafa Sagir, Rebecca Sanders

University of Cincinnati, School of Public and International Affairs, 1203 Crosley Tower, 2600 Clifton Court, Cincinnati, OH 45221, USA

ARTICLE INFO

Article history:
Received 20 July 2021
Received in revised form 29 January 2022
Accepted 2 February 2022
Available online 7 February 2022

ABSTRACT

Designing effective public health campaigns to combat COVID-19 vaccine hesitancy requires an understanding of i) who the vaccine hesitant population is, and ii) the determinants of said population's hesitancy. While researchers have identified a number of variables associated with COVID-19 vaccine hesitancy that could inform such campaigns, little is known about the cumulative or relative predictive power of these factors. In this article, we employ a machine learning model to analyze online survey data collected from 3353 respondents. The model incorporates an array of variables that have been shown to impact vaccine hesitancy, allowing us to i) test how well we can predict vaccine hesitancy, and ii) compare the relative predictive impact of each covariate. The model allows us to correctly classify individuals that are vaccine acceptant with 97% accuracy, and those that are vaccine hesitant with 72% accuracy. Trust in and knowledge about vaccines is, by far, the strongest predictor of vaccination choice. While our results demonstrate that public health campaigns designed to increase vaccination rates must find a way to increase public trust in COVID-19 vaccines, our results cannot speak to the malleability of such beliefs, nor how to enhance trust.

© 2022 Elsevier Ltd. All rights reserved.

1. Introduction

Public health experts agree that mass vaccination is essential for mitigating and eventually ending the destructive COVID-19 pandemic. While the majority of US residents have been vaccinated (as of 1/23/2022, 76.0% have received at least one dose), vaccine supply currently outpaces demand, leaving a substantial portion of the population *voluntarily* unprotected [1]. Such trends bolster long held concerns from the public health community that delay in the acceptance and/or outright refusal of vaccinations—what is commonly referred to as "vaccine hesitancy" [2]—could prove deleterious for the US' COVID-19 mitigation efforts, making herd immunity difficult to achieve [3,4].

Research demonstrates that public health campaigns that are able to i) target the vaccine hesitant population, and ii) effectively address the root causes of hesitancy could provide a powerful tool for US COVID-19 mitigation efforts [5]. Whereas researchers have identified a wide array of demographic, political, psychological, and health-based variables associated with vaccine hesitancy that could be used to inform such campaigns [6–11], to date, no study has (to our knowledge) analyzed the relative importance of such factors. Research identifying (and rank ordering) the factors most predictive of hesitancy could help inform the development of public health campaigns by demonstrating which areas *might be* fruitful to target [5]. That said, it is important to note that some features that predict hesitancy may not be receptive to public health interventions. Those developing public health campaigns should carefully consider both i) how important specific features are at predicting hesitancy, and ii) the malleability of those features.

In this study, we use survey data collected from 3353 US adults to create a predictive model of COVID-19 vaccine hesitancy. We use a machine learning based approach, gradient boosting, to predict whether an individual is vaccine hesitant or acceptant, and in doing so, identify i) how accurately the model can predict hesitancy, ii) the relative importance of the included variables in predicting hesitancy, and iii) the direction of such effects. While the model allows us to correctly predict whether an individual is vaccine hesitant or acceptant with 91% accuracy, it exhibits stronger performance in classifying vaccine acceptant (relative to vaccine hesitant) individuals. Here, we correctly classify individuals that are vaccine acceptant with 97% accuracy, and those that are vaccine hesitant with 72% accuracy. Results demonstrate that trust in and knowledge about vaccines is, by far, the most important

* Corresponding author.
E-mail address: jack.mewhirter@uc.edu (J. Mewhirter).

https://doi.org/10.1016/j.vaccine.2022.02.011
0264-410X/© 2022 Elsevier Ltd. All rights reserved.

predictor of vaccine choice. An array of additional factors improve prediction accuracy: notably, age, and the belief that COVID-19 poses a significant risk to one's health, and the health of others in one's network.

2. Material and methods

2.1. Data collection

The goal of this study is to build a model capable of predicting whether an individual is "vaccine hesitant" or "vaccine acceptant", and in turn, identify the variables most predictive of vaccine choice. To do so, we utilize data collected from an online survey of American adults. The web-based survey panel provider, Qualtrics, was commissioned to i) recruit the online panel, ii) disseminate the survey, and iii) screen out low-quality responses. Individuals in the Qualtrics database are recruited by the firm, and then choose to opt in or out of the survey: information is only collected from (and known for) consenting participants.

The use of online samples recruited from survey firms is advantageous in that it allows researchers to quickly identify and recruit panel participants: a necessary condition for the study of emerging and rapidly changing phenomena. That said, such samples—relative to probability based samples drawn from the broader population—can potentially result in selection bias if those who are recruited for and/or opt into the survey are not representative of the general population. Recent research, however, demonstrates that on average, panels from survey firms in the US (such as Qualtrics) i) are *largely* representative of the US population (at least on a number of studied features—12), and ii) respond to stimuli (e.g., experimental treatment) in a manner consistent with 2019 samples drawn from nationally representative, probability samples [13]. While some concerns have been raised about the age, sex, and race based disparities between Qualtrics respondents and the US population, such errors can often be corrected through survey weights and/or established quotas [12].

Our data was collected in two waves. The original sample consisted of responses collected from 12,037 US adults between 8/7–9/7/2020 (well before the vaccine rollout), and utilized a quota sampling procedure, with quotas assigned by race, age, gender, and census statistical division (see 15 for more). While the first wave respondents are largely representative of the US population, it is possible that there are key, unmeasured differences. For instance, it is possible that the topic of the survey encouraged responses from a subset of the population, and deterred others. Given that we do not have information on non-respondents, we cannot test or correct for this. However, given that our goal is to predict hesitancy, not assess the average amount of hesitancy in the US population, selection issues are only a concern if the factors that impact hesitancy among those in our sample are fundamentally different from those that predict hesitancy in the general population [14]. While possible, past research in other contexts has shown that this is not often the case [13].

All respondents were resampled between 3/2–3/19/2022, with 3353 responding. Questions pertaining to vaccine hesitancy were only included in wave 2, and thus, only those responses are used in this study. Table P3-a in the Appendix contains descriptive statistics on all variables used across waves 1 and 2. Note that (as detailed in Part 3 of the Appendix) non-random attrition between waves occurred. In the *Alternative Specification* subsection (proceeding our main analysis), we demonstrate that such attrition is likely not meaningfully impacting our results.

This study was reviewed by the University of Cincinnati Institutional Review Board, receiving an exempt determination (IRB #000003152). The survey instrument, data files, and R code used in this analysis can be found in the Harvard Dataverse Repository [https://doi.org/10.7910/DVN/GJVWYF].

2.2. Outcome measure

Following the SAGE Working Group on Vaccine Hesitancy, we define vaccine hesitancy as "delay in acceptance or refusal of vaccination despite availability of vaccination services" (2, p. 4163). To operationalize this, respondents were asked i) if they had already received at least one dose of any COVID-19 vaccine, and if not ii) whether they planned on getting it in the future. Our dependent variable—*Vaccine Hesitant*—takes a value of 0 when a respondent had either i) already received one dose of the vaccine *or* ii) indicated that they planned on getting it "as soon as possible". *Vaccine Hesitant* takes a value of 1 if the respondent i) had not yet received a vaccine, *and* ii) indicated that they were either a) unsure about getting the vaccine, b) planning to wait to get the vaccine, or iii) refusing to get the vaccine. Fig. 1 presents the distribution of vaccination status and hesitancy among the unvaccinated. As shown, roughly 44.4% of our sample had received at least one dose, which was ~6.7% higher than the national average for US adults at the time [1]. Of the 1873 adults in our sample that had not yet received a vaccine, 56.1 % indicated that they were vaccine acceptant, whereas 43.9% indicated that they were vaccine hesitant.

2.3. Predictor variables: trust in & knowledge about vaccines

Trust in and knowledge about the safety and efficacy of vaccines are significant predictors of vaccine acceptance [6,16–18]. To account for this, we rely on a *Vaccine Trust Index.* To create the index, respondents were asked the extent to which they agree that "the vaccine is safe", "the vaccine is effective" and "the public has been provided with enough information about the coronavirus vaccine": values range from 0 to 10, where 0 indicates complete disagreement, and 10 indicates complete agreement. The final index takes the mean value of the three responses. The questions exhibited a high degree of internal consistency (alpha = 0.946), thus justifying the use of an index. We also include the variable *Natural Science Literacy* which captures how scientifically literate an individual is, as we expect this to impact how they process and interpret vaccine information. To construct the variable, we asked respondents to correctly answer a number of natural science based trivia questions[1]: values range from 0 to 1, where 0 = zero percent correct and 1 = 100 percent correct.

2.4. Predictor variables: trust in science & science communication

Recent findings suggest that trust in science—in particular, one's level of trust that scientific research and communication is not tainted by politics—impacts COVID-19 related beliefs [15,19–21]. The variables *Trust Science Community* and *Trust Science Apolitical* capture the extent to which i) respondents have confidence in the scientific community (broadly), and ii) are confident political biases do not impact scientific research. The variables *Trust Science Politicians* and *Trust Science Media* capture the extent to which a respondent believes that politicians and the media can be trusted to appropriately use and accurately communicate scientific findings. Values for all aforementioned variables range from 0 to 10,

[1] The following "true or false" questions (which are primarily drawn from the General Science Survey) were asked: "all radioactivity is man-made" (F); "the sun revolves around the earth" (F); "the continents on which we live have been moving their locations for millions of years and will continue to move in the future" (T); "the center of the earth is very hot" (T); "antibiotics kill bacteria and viruses" (F); "vaccines help develop immunity to disease (T)".

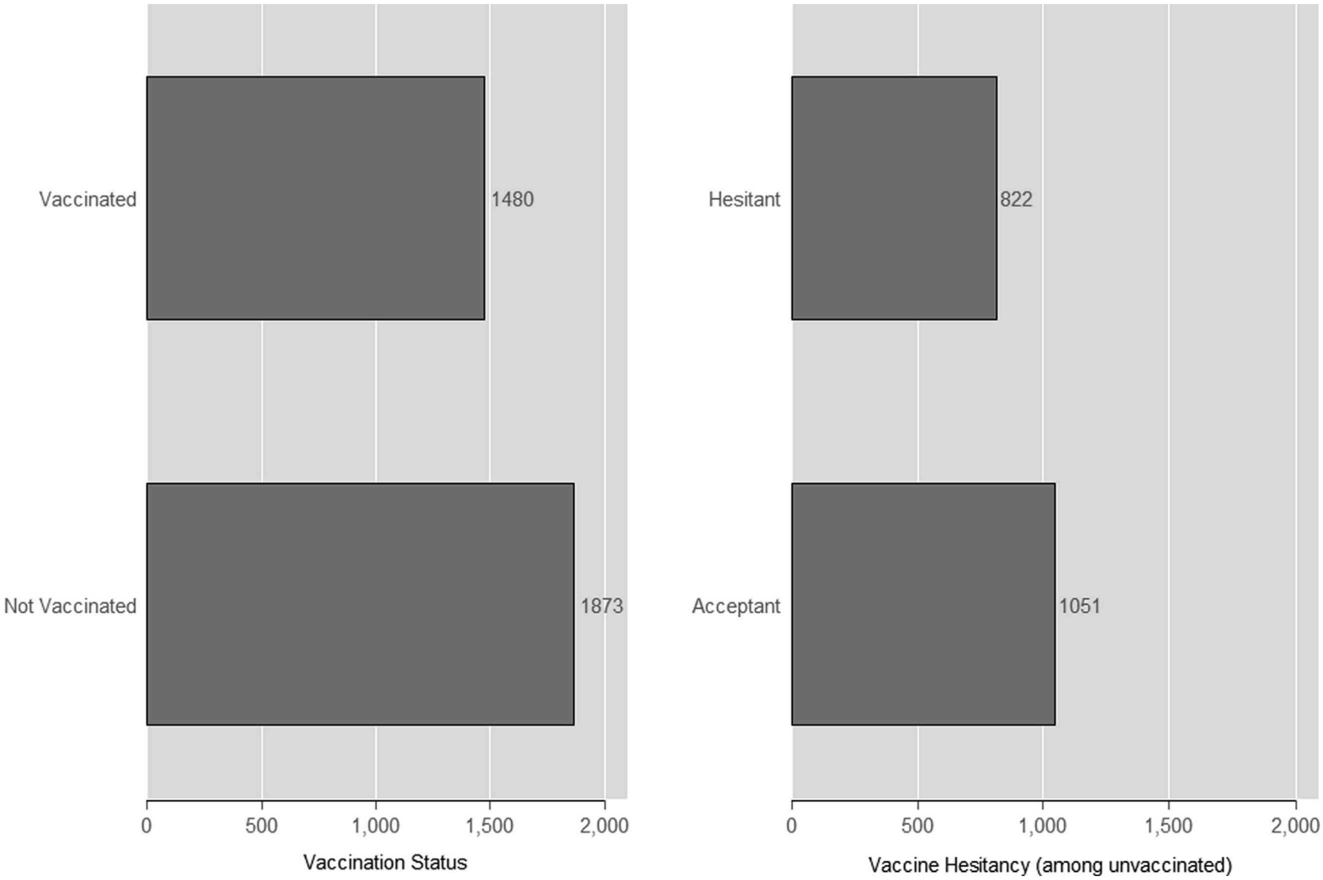

Fig. 1. Vaccination and Hesitancy Status.

with higher values indicating greater trust.[2] Finally, we adjust for general levels of trust in the media through the variable, *Trust Media*, which indicates how often a respondent can trust media to "do what is right": values range from 0 to 10 where 0="Never" and 10="Always".

2.5. Predictor variables: political affiliation & beliefs

Research shows that political discourse in the Unites States has shaped attitudes towards vaccinations, with Republicans—especially those supportive of former President Donald Trump and white Christian evangelicals[3]—exhibiting greater suspicion towards COVID vaccinations [22,23]. The variable *Party ID* captures respondents' political affiliation (1 = Democrat; 2 = Independent; 3 = Libertarian; 4 = Republican; 5 = other) whereas *Biden* and *Trump* capture their support for the current and previous President (0 = No; 1 = Yes). The variables *Trust National, Trust State*, and *Trust Local* capture the extent to which a respondent trusts each of the respective levels of government to do "what is right" (0–10; 0="Never" Trust; 10="Always" Trust). Finally, we include the variable *PS Index* which captures how "politically sophisticated" [24] or knowledgeable an individual is, by gauging their ability to correctly answer several political trivia questions[4]: (0–1; 0 = zero percent correct; 1 = 100 percent correct). The variable *Evangelical* captures whether an individual identifies as a Christian Evangelical (=1) or not (=0).

2.6. Predictor variables: comfort in a healthcare setting

One's level of comfort in a healthcare setting can impact one's willingness to undergo medical procedures including (but not limited to) vaccinations [6,25]. We account for this with the variables *Doctor Comfort* and *Fear Needles*. The variables capture the extent to which a respondent agrees that they "feel comfortable going to the doctor" and "are afraid of needles". Values range from 0 to 10, where 0 = complete disagreement and 10 = complete agreement.

[2] To create these variables, respondents were asked the extent to which they agree that: i) they have "a great deal of confidence in the people running the scientific community"; ii) "a lot of research conducted by scientists is driven by their political motives"; iii) "politicians often skew and misrepresent scientific findings to promote their own interests"; "The news media often skews and misrepresents scientific findings to promote their own interests". Values range from 0 to 10, where 0=complete disagreement, and 10=complete agreement. Responses to statements ii, iii, and iv were then reverse coded (|10-value|) so that each measure (i-iv) indicates greater trust.

[3] White Christian evangelicals are a key part of the conservative Republican coalition in the US. We include this group not because of their religious beliefs, but because evangelicals are often characterized as vaccine hesitant in media coverage of vaccine debates.

[4] Respondents were asked the following multiple-choice questions: "Which position does Mike Pence currently hold? (Attorney General; Vice President*; Not Sure); Which party currently has the most members in the US House of Representatives in Washington DC? (Democrats*; Republicans; Not Sure); The system of government where power is divided between a national and regional governments is called what? (Federal system*; Mixed system; Not sure); On which of the following does the US federal government currently spend the least? (Foreign Aid*; Social Security; Not sure); Which levels of government are primarily charged with funding public schools? (National and State governments; State and local governments*; Not sure); Which level of government is primarily responsible for setting zoning policies and building codes? (National government; Local governments*)

2.7. Predictor variables: risk associated with catching & spreading COVID-19

Risk has long been acknowledged as a "core concept in theories of health behavior" [26,p. 136]. As related to vaccinations, findings suggest that those who perceive themselves to be and/or objectively are i) at a greater risk of catching the illness, ii) vulnerable to severe illness, and/or iii) likely to pass on the illness (if infected) to others who are vulnerable to severe illness, are more likely to get vaccinated [21,27]. *Perceived Risk* captures the extent to which an individual agrees that the coronavirus "poses a serious risk to my health". *Perceived Network Risk* captures the extent to which they agree it "poses a serious risk to those with whom [they] regularly interact" (0–10 scale where 0 = complete disagreement and 10 = complete agreement).

We include an array of variables that capture factors associated with severe virus-related illness, including *Age* (in years), as well as eight variables that identify whether an individual has any of the following conditions[5]: pregnancy; asthma; lung disease; diabetes; immune disorder; obesity; heart problems; liver or kidney problems.[6] Finally, we include variables that capture the chances of getting infected by modeling whether they have already been infected (*Infected Personal*: 0 = No; 1 = Yes), whether someone in their immediate network has been infected (*Infected Network*: 0 = No; 1 = Yes), the proportion of people in the county in which they reside that have been infected (*County Cases*), the proportion of people in their county that have been infected in the past two weeks (*County Cases 2wk*), and the population density of their county (*County Density*).[7]

2.8. Predictor variables: mental and financial health based incentives

We anticipate that financial and mental health-based incentives could play a large role in motivating vaccine acceptance/hesitancy. We capture this with three variables: *Pandemic Impact; Pandemic Impact Network; Vaccine Required.* Respondents were asked the extent to which i) their financial health and ii) their mental health was impacted by the pandemic (0–10, where 0 ="major negative impact" and 10="major positive impact"): *Personal Impact* takes the mean value of these responses. *Pandemic Impact Network* is captured in the same manner, but in relation to people they "consider close". *Vaccine Required* captures whether an "employer, school, or other entity" requires that they get the vaccination (1 = Yes; 0 = No or Unsure).

2.9. Predictor variables: demographics

Finally, we account for a number of demographic variables that have been shown to influence COVID-19 related vaccine hesitancy [28,29]: *Race* (1 = white; 2 = Black; 3 = Hispanic; 4 = other); *Male* (0 = No; 1 = Yes); *College Degree* (1 = Yes; 0 = No); *Household Income* (1–12).[8]

2.10. Statistical model

We apply a gradient boosting (GB—also known as boosted trees) model to analyze the collected data and predict respondents' decisions on vaccine choice. GB is a supervised learning algorithm that predicts an outcome of interest based on an ensemble of weak prediction models. Given that our outcome variable is dichotomous, our GB model iteratively estimates a series of classification trees. Each model integrates information obtained from previously estimated trees, searching for an additive model that reduces the loss function [30,31]. Each model is adjoined to the previous model, and the process continues for a number of iterations. As the model seeks to repeatedly minimize the loss function, the overall prediction accuracy of the model increases with each iteration. GB uses individual trees as a base learner and the final prediction is based on an ensemble of models. As opposed to other, more commonly employed ensemble learning methods (namely, random forests)—where i) all trees are created independently by only including a random subset of predictor variables, ii) each tree is created to have maximum depth, and iii) each tree contributes equally to the final model—the trees in GB are contingent on past trees, have minimum depth, and contribute unequally to the final model.

In supervised machine learning, the dataset is split into two sets: a "training set" and a "test set". The training set (which contains the bulk of the data) allows for the algorithm to learn which variables are important and how. Using this information, the algorithm is then used to "predict" the expected value of the dependent variable (in our case, 0 or 1) in the "test set", conditional on the observed values of the predictor variables. The observations in the test set are also known as out-of-sample/validation set observations. In our data, 789 of the 3353 observations compose the test set: the remaining observations compose the training set.

The analysis is conducted in R using "Tidymodels" and "xgboost" packages for modeling and tuning purposes and "vip" and "pdp" packages for model interpretability.[9] Table 1 reports the confusion matrix and performance metrics of our model for the validation/test set observations. Overall, our model predicts whether a respondent is vaccine hesitant *or* vaccine acceptant with 91% accuracy. That said, the model better predicts vaccine acceptant individuals: as demonstrated by the sensitivity and specificity metrics, our model accurately predicts vaccine acceptant individuals with 97 % accuracy, and vaccine hesitant individuals with 72% accuracy. Accordingly, the area under the receiver operating characteristic curve (ROC-AUC) is 93%.[10]

[5] Variable names are: Condition Pregnancy; Condition Asthma; Condition Lung; Condition Diabetes; Condition Immune; Condition Obesity; Condition Heart; Condition Organ.

[6] Note that our reliance on self-reported measures capture whether an individual believes that they have a condition, which may contradict the opinion of a medical provider. For instance, while only 10% of our round 2 sample indicates being obese, it is likely much higher. Here, obese individuals may i) have a different conceptualization of what constitutes obesity, or ii) given the social stigma surrounding it, be less likely to self-report as obese. As such, our analysis will reveal how one's perceptions of their own health conditions impact their vaccine choice (as opposed to the effect of the actual condition).

[7] Respondents were asked to provide their zip code from which county and state identifiers were extracted. County case data comes from the New York Times coronavirus case database (https://github.com/nytimes/covid-19-data). Data on population density comes from 2018 Census American Community Survey (5 year) estimates.

[8] Values are: <$10,000 (=1); $10,000-$19,999 (=2); $20,000-$29,999 (=3); $30,000-$39,999 (=4); $40,000-$49,999 (=5); $50,000-$59,999 (=6); $60,000-$69,999 (=7); $70,000-$79,999 (=8); $80,000-$89,999 (=9); $90,000-$99,999 (=10); $100,000-$149 ,999 (=11); >$150,000 (=12)

[9] After splitting the dataset into training and testing sets (75% training set – 25% test set), the model is built by using "tidymodels" and "xgboost" packages and using "xgboost" engine on the training set. Performance metrics on the out of bag sample is calculated, then, 10-fold cross validation is applied, and prediction accuracy metrics (accuracy, ROC-AUC) are re-assessed. To optimize the predictive performance, the model is tuned to find the optimal "tree_depth", "learn_rate", and "sample_size". After tuning, the optimum model parameters are applied to the model and performance metrics (accuracy, sensitivity, specificity, and ROC-AUC) are calculated in the test set. Please see pages 2–4 in the Appendix for full details regarding how our analysis was performed. Also note that all data and code needed to replicate our analysis can be found on corresponding Harvard Dataverse site [https://doi.org/10.7910/DVN/GJVWYF].

[10] The receiver operator characteristic (ROC) curve is a performance metric for binary classification models. It is a probability curve that plots the true positive rate against the false positive rate at various threshold values. The area under the curve (AUC) is the measure of the ability of a classifier to make a distinction between outcome values and is used as a summary of the ROC curve. Essentially, higher AUCs signify a better prediction between the positive and negative classes.

Table 1
Confusion Matrix and Performance Metrics of Boosted Tree Model.

	Real Output	
Prediction	Vaccine Acceptant (0)	Vaccine Hesitant (1)
Acceptant (0)	580	53
Hesitant (1)	20	136
Accuracy	0.91	
Sensitivity	0.97	
Specificity	0.72	
ROC-AUC	0.93	

2.11. Variable importance and direction

GB allows us to determine the extent to which each included covariate contributes to predicting the outcome variable. Fig. 2 presents a variable importance plot (VIP), showing the accuracy loss associated with excluding each variable from the GB model. The gain built-in feature of xgboost package calculates the variable importance based on the relative impact of the specific feature to the model, calculated by taking each feature's contribution for each tree in the GB model. For result interpretability (considering the number of covariates included in the model) only the top 10 most important predictors are included. In Figure P1-b in the Appendix, we include VIP that considers the top 30 predictors. To better illustrate the importance of the features, we rescale importance scores so that the largest value is 100.

As shown, the *Vaccine Trust Index* is, *by far*, the most important predictor (importance score = 100), followed by *Age* (=4.38), *Perceived Network Risk* (=2.45), *Perceived Risk* (2.00), *Trust State* (=1.64), *County Cases 2wk* (=1.62), *County Density* (=1.28), *Trust Local* (=0.84), *Household Income* (=0.91), *Race-Black (0.68)* and *County Cases* (=0.63). Collectively the top four variables in the model—those that maintain importance scores > 2—explain the vast majority of variation in vaccine choice. To illustrate the predictive power of the key variables, we estimate two new GB models: one that solely includes the *Vaccine Trust* variable, and another that includes the top 4 variables identified through the original model. Relative to the full model, the single predictor model and four predictor model exhibit only slightly lower levels of performance, as demonstrated by the indicated by the accuracy (0.891; 0.894), sensitivity (0.965; 0.953), specificity (0.656; 0.704), and ROC-AUC (0.916; 0.922) metrics. Given that the goal of this project is to identify the most predictive determinants of vaccine hesitancy, we focus on these top 4 variables for the remainder of the article.

To understand and visualize the relationship between each of the four most important predictor variables and vaccine choice, we utilize four partial dependence plots (PDPs). PDPs present the complex estimated nonparametric prediction function as a low-dimensional graph, presenting the marginal effect of each predictor variable on the class probability of the outcome variable [32]. Fig. 3 contains the PDPs for the vaccine-hesitant category.

The PDPs show that a greater degree of trust in/knowledge of vaccines—the strongest predictor of vaccine hesitancy—is negatively associated with hesitancy (but only for values >~2.5). Age is also negatively associated with hesitancy up to a point: though hesitancy appears to increase as age surpasses 80, this should be interpreted with caution as very few respondents (~1.5%) were 80 years of age or older. Greater concern regarding the impact that COVID-19 would have on individuals in one's network has a negative, relatively linear impact on hesitancy as values exceed ~5. *Perceived Risk* has a negative relatively-linear impact on hesitancy as values exceed ~6.5.

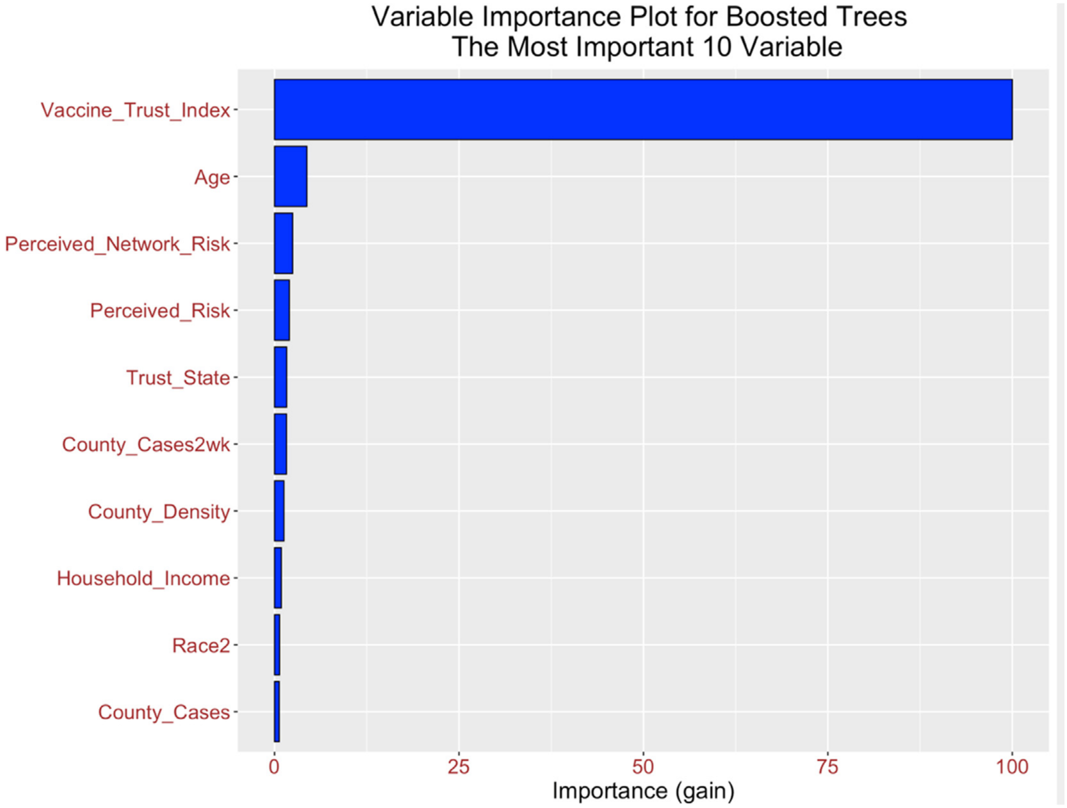

Fig. 2. Variable Importance Plot.

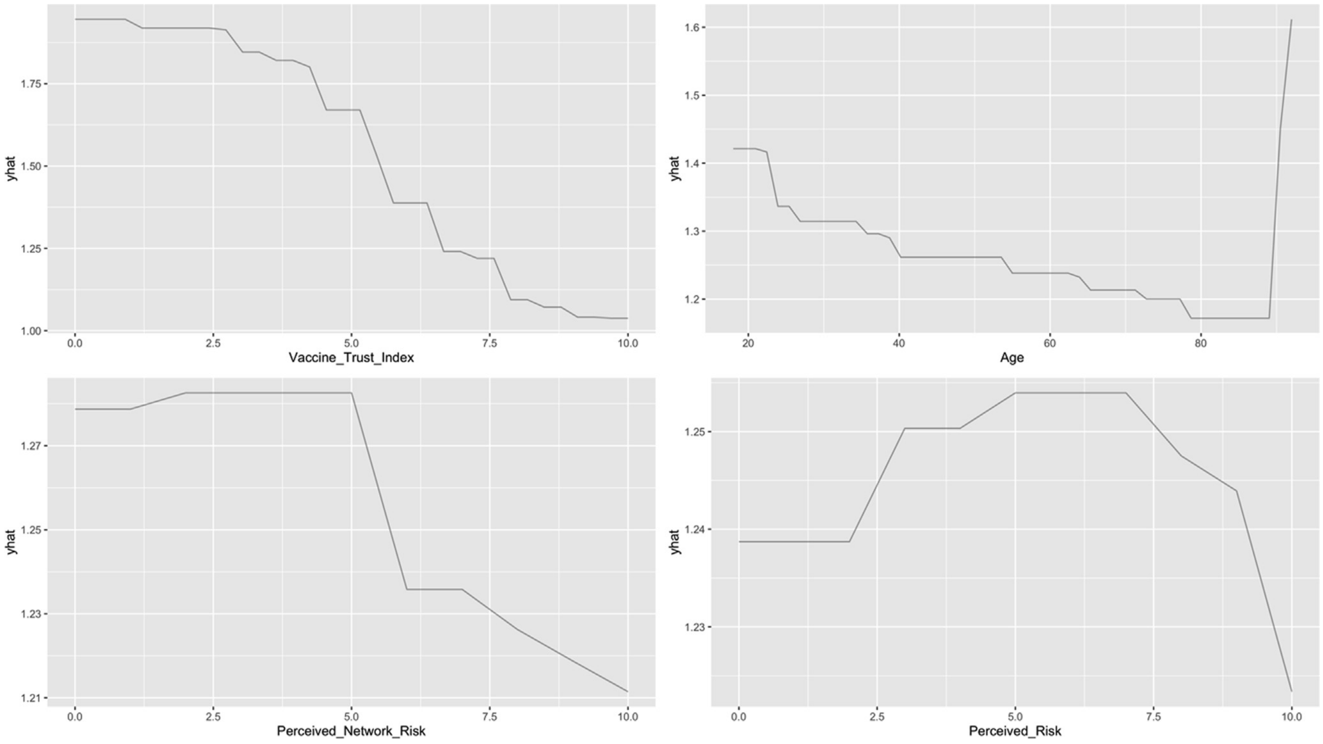

Fig. 3. Dependency Plots for '*Vaccine Hesitant*'

2.12. Alternative specifications

In this section, we assess whether and to what extent our results are i) model dependent, and ii) impacted by non-random attrition between survey waves.

We replicate our analyses using four alternative approaches—logistic regression, decision trees, bagging, and random forests—to assess the robustness of our findings. In Part 2 of the Appendix (pages 5–17), we describe estimation procedures, detail model performance, and present variable importance plots. In Tables P2-a and P2-b (page 5 in the Appendix) we assess variation in i) performance and ii) variable importance across models. As shown, the GB approach outperforms each alternate model. *Vaccine Trust* is the most important predictor across models. While the ordered importance of the remaining 10 predictors varies, those considered in Fig. 3 rank highly in each.

In Table P3-a in the Appendix, we present summary statistics for all variables across waves, demonstrating that a number of wave 1 variables impact the likelihood of participating in wave 2 [indicating non-random attrition). To adjust for this, we leverage a propensity score weighting approach to evaluate—using wave 2 data—how the variables used in the primary model impact vaccine hesitancy, once attrition is taken into account [33].[11] Specifically, the model seeks to examine *what would have been observed had attrition occurred randomly*. The procedure and results are described in detail in pages 18–20 of the Appendix. The findings, presented in Tables P3-b and P3-c, demonstrate that while non-random attrition occurred, failing to correct for it does not significantly alter model performance, nor does it dramatically change to what extent the covariates impact vaccine choice. As such, we do not expect that non-random attrition has a meaningful impact on the primary results presented in this manuscript.

3. Discussion

Our findings have implications for public health interventions designed to overcome COVID-19 vaccine hesitancy. Notably, while a number of variables—including age, and the perception that COVID-19 poses a risk to oneself and those in one's network—are associated with hesitancy, our results strongly suggest that addressing the public's lack of trust in the COVID-19 vaccine is of paramount importance. While certain populations such as Republicans, African Americans, and Christian evangelicals evince higher levels of COVID-19 vaccine hesitancy, these identities per se may not be driving forces of hesitancy. Rather, it may be that features that are strongly associated with hesitancy—such as lack of trust in vaccines—are clustered in certain constituencies, thus shaping demographic patterns. If true, public health campaigns should not simply rely on politically, racially, or religiously targeted messaging, but instead consider the deeper sources of mistrust and discomfort.

While our study provides important insights relevant to scholars and practitioners alike, there are a number of limitations that should be taken into account in future research. First, our model better classifies vaccine acceptant individuals (97%) relative to vaccine hesitant individuals (72%). While we, in an effort to maximize predictive capacity, included a wide array of covariates, such inaccuracies demonstrate that there (likely) exists a number of unmodeled features. Second, our reliance on a quota-based online convenience sample and the presence of non-random attrition present complications for the generalizability of our findings. That said, we argue that given the strength of the relationship between vaccine trust and vaccine choice (both in the unweighted and weighted models), it is unlikely that this particular finding is specific to our sample. Finally, while the model demonstrates how strongly the included variables contribute to the predictive capacity of the model, we cannot assume that this relationship is causal. Given these limitations, we consider this article an important "first step" for the use of machine learning models in predicting COVID-

[11] This procedure is commonly used to correct for selection bias arising from multiple sources, including non-response to/attrition from surveys [34,35].

19 vaccine hesitancy: one that we hope is extended and improved upon in future research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Science Foundation (SES #2034367).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.vaccine.2022.02.011.

References

[1] Centers for Disease Control. COVID-19 Vaccination Trends in the United States, National and Jurisdictional. January 2022. https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2.
[2] MacDonald NE. Vaccine hesitancy: definition, scope and determinants. Vaccine 2015;33(34):4161–4.
[3] Mandavilli A. Reaching 'Herd Immunity' Is Unlikely in the U.S., Experts Now Believe. The New York Times, May 3, 2021. https://www.nytimes.com/2021/05/03/health/covid-herd-immunity-vaccine.html.
[4] Oliu-Barton M, Pradelski BSR, Aghion P, Artus P, Kickbusch I, Lazarus JV, et al. SARS-CoV-2 elimination, not mitigation, creates best outcomes for health, the economy, and civil liberties. Lancet 2021;397(10291):2234–6.
[5] Noar SM. A 10-year retrospective of research in health mass media campaigns: where do we go from here? J Health Commun 2006;11(1):21–42.
[6] Dubé E, Laberge C, Guay M, Bramadat P, Roy R, Bettinger JA. Vaccine hesitancy: an overview. Human Vaccines Immunotherap 2013;9(8):1763–73.
[7] Hornsey MJ, Harris EA, Fielding KS. The psychological roots of anti-vaccination attitudes: a 24-nation investigation. Health Psychol 2018;37(4):307–15.
[8] Kaplan Robert M, Milstein Arnold. Influence of a COVID-19 vaccine's effectiveness and safety profile on vaccination acceptance. Proc Natl Acad Sci 2021;118(10):e2021726118.
[9] Marti M, de Cola M, MacDonald NE, Dumolard L, Duclos P, Borrow R. Assessments of global drivers of vaccine hesitancy in 2014—Looking beyond safety concerns. PLoS ONE 2017;12(3):e0172310. https://doi.org/10.1371/journal.pone.0172310.
[10] Murphy J, Vallières F, Bentall RP, Shevlin M, McBride O, Hartman TK, McKay R, et al. Psychological characteristics associated with COVID-19 vaccine hesitancy and resistance in Ireland and the United Kingdom. Nat Commun 2021;12(1):1–15.
[11] Ruiz JB, Bell RA. Predictors of intention to vaccinate against COVID-19: results of a nationwide survey. Vaccine 2021;39(7):1080–6.
[12] Boas TC, Christenson DP, Glick DM. Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. Polit Sci Res Methods 2020;8(2):232–50.
[13] Coppock A, McClellan OA. Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. Res Polit 2019;6(1). 2053168018822174.
[14] Baker R, Brick JM, Bates NA, Battaglia M, Couper MP, Dever JA, et al. Summary report of the AAPOR task force on non-probability sampling. J Surv Stat Methodol 2013;1(2):90–143.
[15] McLaughlin D, Mewhirter J, Sanders R. The belief that politics drive scientific research and its impact on COVID-19 risk assessment. Plos one 2021;16(4):e0249937.
[16] Jamison AM, Quinn SC, Freimuth VS. "You don't trust a government vaccine": narratives of institutional trust and influenza vaccination among African American and white adults. Soc Sci Med 2019;221:87–94.
[17] Goldenberg MJ. Vaccine hesitancy: public trust, expertise, and the war on science. University of Pittsburgh Press; 2021.
[18] Quinn SC, Jamison AM, An Ji, Hancock GR, Freimuth VS. Measuring vaccine hesitancy, confidence, trust and flu vaccine uptake: results of a national survey of White and African American adults. Vaccine 2019;37(9):1168–73.
[19] Donovan J. Concrete Recommendations for cutting through misinformation during the COVID-19 pandemic. Am J Public Health 2020;110(S3):S286–7.
[20] Dube J, Simonov A, Sacher S, Biswas S. News media and distrust in scientific experts. VoxEU 2020.
[21] McLaughlin D, Mewhirter J, Sanders R. The belief that politics drive scientific research & its impact on COVID-19 risk assessment. PLoS ONE 2021;16(4):e0249937.
[22] Hornsey MJ, Finlayson M, Chatwood G, Begeny CT. Donald Trump and vaccination: the effect of political identity, conspiracist ideation and presidential tweets on vaccine hesitancy. J Exp Soc Psychol 2020;88:103947. https://doi.org/10.1016/j.jesp.2019.103947.
[23] Fridman A, Gershon R, Gneezy A, Capraro V. COVID-19 and vaccine hesitancy: a longitudinal study. PLoS ONE 2021;16(4):e0250123. https://doi.org/10.1371/journal.pone.0250123.
[24] Gomez BT, Matthew Wilson J. Political sophistication and economic voting in the American electorate: a theory of heterogeneous attribution. Am J Polit Sci 2001;45(4):899–914.
[25] Hobson-West P. Understanding vaccination resistance: Moving beyond risk. Health, Risk Soc 2003;5(3):273–83.
[26] Brewer NT, Chapman GB, Gibbons FX, Gerrard M, McCaul KD, Weinstein ND. Meta-analysis of the relationship between risk perception and health behavior: the example of vaccination. Health Psychol 2007;26(2):136–45.
[27] Dryhurst S, Schneider CR, Kerr J, Freeman ALJ, Recchia G, van der Bles AM, et al. Risk perceptions of COVID-19 around the world. J Risk Res 2020;23(7-8):994–1006.
[28] Cordina M, Lauri MA, Lauri J. Attitudes towards COVID-19 vaccination, vaccine hesitancy and intention to take the vaccine. Pharmacy Practice (Granada) 2021;19(1):2317. https://doi.org/10.18549/PharmPract.2021.1.2317.
[29] Khubchandani J, Sharma S, Price JH, Wiblishauser MJ, Sharma M, Webb FJ. COVID-19 vaccination hesitancy in the United States: a rapid national assessment. J Community Health 2021;46(2):270–7.
[30] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29(5):1189–232.
[31] Chen Tianqi, He Tong, Benesty Michael, Khotilovich Vadim, Tang Yuan, Cho Hyunsu. Xgboost: Extreme gradient boosting. R package version 0.4-2 1. 2015; (4): 1–4.
[32] Greenwell BM. pdp: An R package for constructing partial dependence plots. R J 2017;9(1):421–36.
[33] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41–55.
[34] Olmos A, Govindasamy P. Propensity scores: a practical introduction using R. J MultiDisc Evaluat 2015;11(25):68–88.
[35] McLaughlin DM, Mewhirter JM, Wright JE, Feiock R. The perceived effectiveness of collaborative approaches to address domestic violence: the role of representation, 'reverse-representation, 'embeddedness, and resources. Public Manage Rev 2021;23(12):1808–32. https://doi.org/10.1080/14719037.2020.1774200.