Using Machine Learning to Identify Important Predictors of COVID-19 Infection Prevention Behaviors During the Early Phase of the Pandemic

Caspar J. van Lissa, Wolfgang Stroebe, Michelle R. vanDellen, N. Pontus Leander, Maximilian Agostini, Tim Draws, Andrii Grygoryshyn, Ben Gützgow, Jannis Kreienkamp, Clara S. Vetter, Georgios Abakoumkin, Jamilah Hanum Abdul Khaiyom, Vjolica Ahmedi, Handan Akkas, Carlos A. Almenara, Mohsin Atta, Sabahat Cigdem Bagci, Sima Basel, Edona Berisha Kida, Allan B.I. Bernardo, Nicholas R. Buttrick, Phatthanakit Chobthamkit, Hoon-Seok Choi, Mioara Cristea, Sára Csaba, Kaja Damnjanović, Ivan Danyliuk, Arobindu Dash, Daniela Di Santo, Karen M. Douglas, Violeta Enea, Daiane Gracieli Faller, Gavan J. Fitzsimons, Alexandra Gheorghiu, Ángel Gómez, Ali Hamaidia, Qing Han, Mai Helmy, Joevarian Hudiyana, Bertus F. Jeronimus, Ding-Yu Jiang, Veljko Jovanović, Željka Kamenov, Anna Kende, Shian-Ling Keng, Tra Thi Thanh Kieu, Yasin Koc, Kamila Kovyazina, Inna Kozytska, Joshua Krause, Arie W. Kruglanksi, Anton Kurapov, Maja Kutlaca, Nóra Anna Lantos, Edward P. Lemay, Jr., Cokorda Bagus Jaya Lesmana, Winnifred R. Louis, Adrian Lueders, Najma Iqbal Malik, Anton P. Martinez, Kira O. McCabe, Jasmina Mehulić, Mirra Noor Milla, Idris Mohammed, Erica Molinario, Manuel Moyano, Hayat Muhammad, Silvana Mula, Hamdi Muluk, Solomiia Myroniuk, Reza Najafi, Claudia F. Nisa, Boglárka Nyúl, Paul A. O'Keefe, Jose Javier Olivas Osuna, Evgeny N. Osin, Joonha Park, Gennaro Pica, Antonio Pierro, Jonas H. Rees, Anne Margit Reitsema, Elena Resta, Marika Rullo, Michelle K. Ryan, Adil Samekin, Pekka Santtila, Edyta M. Sasin, Birga M. Schumpe, Heyla A. Selim, Michael Vicente Stanton, Samiah Sultana, Robbie M. Sutton, Eleftheria Tseliou, Akira Utsugi, Jolien Anne van Breen, Kees Van Veen, Alexandra Vázquez, Robin Wollast, Victoria Wai-Lan Yeung, Somayeh Zand, Iris Lav Žeželj, Bang Zheng, Andreas Zick, Claudia Zúñiga, Jocelyn J. Bélanger

PII: S2666-3899(22)00067-8

DOI: https://doi.org/10.1016/j.patter.2022.100482

Reference: PATTER 100482

To appear in: *Patterns*

Received Date: 7 December 2021

Revised Date: 23 December 2021

Accepted Date: 4 March 2022

Please cite this article as: van Lissa CJ, Stroebe W, vanDellen MR, Leander NP, Agostini M, Draws T, Grygoryshyn A, Gützgow B, Kreienkamp J, Vetter CS, Abakoumkin G, Abdul Khaiyom JH, Ahmedi V, Akkas H, Almenara CA, Atta M, Bagci SC, Basel S, Kida EB, Bernardo ABI, Buttrick NR, Chobthamkit P, Choi H-S, Cristea M, Csaba S, Damnjanović K, Danyliuk I, Dash A, Di Santo D, Douglas KM, Enea V, Faller DG, Fitzsimons GJ, Gheorghiu A, Gómez Á, Hamaidia A, Han Q, Helmy M, Hudiyana J, Jeronimus BF, Jiang D-Y, Jovanović V, Kamenov Ž, Kende A, Keng S-L, Thanh Kieu TT, Koc Y, Kovyazina K, Kozytska I, Krause J, Kruglanksi AW, Kurapov A, Kutlaca M, Lantos NA, Lemay Jr. EP, Jaya Lesmana CB, Louis WR, Lueders A, Malik NI, Martinez AP, McCabe KO, Mehulić J, Milla MN, Mohammed I, Molinario E, Moyano M, Muhammad H, Mula S, Muluk H, Myroniuk S, Najafi R, Nisa CF, Nyúl B, O'Keefe PA, Olivas Osuna JJ, Osin EN, Park J, Pica G, Pierro A, Rees JH, Reitsema AM, Resta E, Rullo M, Ryan MK, Samekin A, Santtila P, Sasin EM, Schumpe BM, Selim HA, Stanton MV, Sultana S, Sutton RM, Tseliou E, Utsugi A, Anne van Breen J, Van Veen K, Vázquez A, Wollast R, Wai-Lan Yeung V, Zand S, Žeželj IL, Zheng B, Zick A, Zúñiga C, Bélanger JJ, Using Machine Learning to Identify Important Predictors of COVID-19 Infection Prevention Behaviors During the Early Phase of the Pandemic, *Patterns* (2022), doi: https://doi.org/10.1016/j.patter.2022.100482.

Using Machine Learning to Identify Important Predictors of COVID-19 Infection Prevention Behaviors During the Early Phase of the Pandemic

Caspar J. van Lissa[1]

Wolfgang Stroebe[2]

Michelle R. vanDellen[3]

N. Pontus Leander[2, 69]

Maximilian Agostini[2]

Tim Draws[4]

Andrii Grygoryshyn[5]

Ben Gützgow[2]

Jannis Kreienkamp[2]

Clara S. Vetter[5]

Georgios Abakoumkin[6]

Jamilah Hanum Abdul Khaiyom[7]

Vjolica Ahmedi[8]

Handan Akkas[9]

Carlos A. Almenara[10]

Mohsin Atta[11]

Sabahat Cigdem Bagci[12]

Sima Basel[13]

Edona Berisha Kida[8]

Allan B. I. Bernardo[14]

Nicholas R. Buttrick[15]

Phatthanakit Chobthamkit[16]

Hoon-Seok Choi[17]

Mioara Cristea[18]

Sára Csaba[19]

Kaja Damnjanović[20]

Ivan Danyliuk[21]

Arobindu Dash[22]

Daniela Di Santo[23]

Karen M. Douglas[24]

Violeta Enea[25]

Daiane Gracieli Faller[13]

Gavan J. Fitzsimons[26]

Alexandra Gheorghiu[25]

Ángel Gómez[27]

Ali Hamaidia[28]

Qing Han[29]

Mai Helmy[30,31]

Joevarian Hudiyana[32]

Bertus F. Jeronimus[2]

Ding-Yu Jiang[33]

Veljko Jovanović[34]

Željka Kamenov[35]

Anna Kende[36]

Shian-Ling Keng[37]

Tra Thi Thanh Kieu[38]

Yasin Koc[2]

Kamila Kovyazina[39]

Inna Kozytska[21]

Joshua Krause[2]

Arie W. Kruglanksi[40]

Anton Kurapov[42]

Maja Kutlaca[43]

Nóra Anna Lantos[36]

Edward P. Lemay, Jr.[408]

Cokorda Bagus Jaya Lesmana[43]

Winnifred R. Louis[44]

Adrian Lueders[45]

Najma Iqbal Malik[11]

Anton P. Martinez[46]

Kira O. McCabe[47]

Jasmina Mehulić[35]

Mirra Noor Milla[32]

Idris Mohammed[48]

Erica Molinario[49]

Manuel Moyano[50]

Hayat Muhammad[51]

Silvana Mula[23]

Hamdi Muluk[31]

Solomiia Myroniuk[2]

Reza Najafi[52]

Claudia F. Nisa[13]

Boglárka Nyúl[36]

Paul A. O’Keefe[53]

Jose Javier Olivas Osuna[27]

Evgeny N. Osin[54]

Joonha Park[55]

Gennaro Pica[56]

Antonio Pierro[23]

Jonas H. Rees[57]

Anne Margit Reitsema[2]

Elena Resta[23]

Marika Rullo[58]

Michelle K. Ryan[57, 2]

Adil Samekin[60]

Pekka Santtila[61]

Edyta M. Sasin[13]

Birga M. Schumpe[5]

Heyla A. Selim[62]

Michael Vicente Stanton[63]

Samiah Sultana[2]

Robbie M. Sutton[24]

Eleftheria Tseliou[6]

Akira Utsugi[64]

Jolien Anne van Breen[65]

Kees Van Veen[2]

Alexandra Vázquez[27]

Robin Wollast[43]

Victoria Wai-Lan Yeung[66]

Somayeh Zand[49]

Iris Lav Žeželj[20]

Bang Zheng[67]

Andreas Zick[57]

Claudia Zúñiga[68]

Jocelyn J. Bélanger[13]

Correspondence to the lead author Caspar Van Lissa [C.J.vanLissa@uu.nl]

1. Utrecht University
2. University of Groningen
3. University of Georgia
4. Delft University of Technology
5. University of Amsterdam
6. University of Thessaly
7. International Islamic University Malaysia
8. Pristine University
9. Ankara Science University
10. Universidad Peruana de Ciencias Aplicadas
11. University of Sargodha
12. Sabanci University
13. New York University Abu Dhabi
14. De La Salle University
15. University of Virginia
16. Thammasat University
17. Sungkyunkwan University
18. Heriot Watt University
19. ELTE Eötvös Loránd University
20. University of Belgrade
21. Taras Shevchenko National University of Kyiv
22. Leuphana University of Luneburg
23. University "La Sapienza", Rome
24. University of Kent
25. Alexandru Ioan Cuza University
26. Duke University
27. Universidad Nacional de Educacion a Distancia
28. Setif 2 University
29. University of Bristol
30. Sultan Qaboos University

31. Menoufia University
32. Universitas Indonesia
33. National Chung-Cheng University
34. University of Novi Sad
35. University of Zagreb
36. ELTE Eötvös Loránd University
37. Monash University
38. HCMC University of Education
39. Independent Researcher
40. University of Maryland, College Park
41. Tara Shevchenko National University of Kyiv
42. Durham University
43. Udayana University
44. The University of Queensland
45. University of Limerick
46. The University of Sheffield
47. Carleton University
48. Usmanu Danfodiyo University Sokoto
49. Florida Gulf Coast University
50. University of Cordoba
51. University of Peshawar
52. University of Padova
53. Yale-NUS College
54. HSE University
55. NUCB Business School
56. University of Camerino
57. Bielefeld University
58. University of Siena- Arezzo Campus
59. University of Exeter
60. M. Narikbayev KAZGUU University
61. New York University Shanghai
62. King Saud University
63. California State University East Bay
64. Nagoya University
65. Leiden University
66. Lingnan University
67. Imperial College London
68. Universidad de Chile
69. Wayne State University

Abstract

Before vaccines for COVID-19 became available, a set of infection prevention behaviors constituted the primary means to mitigate the virus spread. Our study aimed to identify important predictors of this set of behaviors. Whereas social and health psychological theories suggest a limited set of predictors, machine learning analyses can identify correlates from a larger pool of candidate predictors. We used random forests to rank 115 candidate correlates of infection prevention behavior in 56,072 participants across 28 countries, administered in March-May 2020. The machine-learning model predicted 52% of the variance in infection prevention behavior in a separate test sample—exceeding the performance of psychological models of health behavior. Results indicated the two most important predictors related to individual-level injunctive norms. Illustrating how data-driven methods can complement theory, some of the most important predictors were not derived from theories of health behavior—and some theoretically-derived predictors were relatively unimportant.

Keywords: Machine learning; COVID-19; Health Behaviors; Social Norms; Public Goods Dilemma

Summary

In the absence of a vaccine or cure, virus containment depended on individual-level compliance with behaviors recommended by the World Health Organization. We used machine learning to identify the most important indicators of compliance, based on a large international psychological survey and country-level secondary data. The most important indicators were not the “usual suspects”, such as personal threat of virus infection, but rather injunctive norms—namely, the belief that one’s community *should* engage in such behavior and that society should take restrictive virus containment measures. People appear who tend to engage in infection prevention behaviors also tend to believe that general compliance is necessary to defeat the pandemic, which extends to endorsement of ‘ought’ norms and support for behavioral mandates. These results highlight the potential to intervene by shaping social norms and expectations.

Introduction

Behavioral measures are crucial in limiting the spread of infectious diseases. This was especially the case in the early phase of the COVID-19 pandemic, between March and May 2020, when no vaccines were available. In this first phase of the pandemic, three infection prevention behaviors were recommended by most governments: frequent hand washing, social distancing, and self-quarantining[1]. The efficacy of these measures for curbing the virus depends on the extent to which individuals engage in these behaviors. The COVID-19 pandemic represented a public health emergency with rich social and system-level data available to evaluate engagement in compliance and focus research and future policy interventions on the most important predictors of such behaviors. Although one approach might be to test whether a specific variable explains important variance in predicting health behaviors. The present work applies machine learning to a large psychological dataset, which was assembled in the early phase of the pandemic and enriched with country-level societal data in order to consider a wider pool of candidate variables. Our primary aim was to identify the most important predictors of infection prevention behavior, given the available data; a secondary aim was to illustrate how inductive methods can help to inform crisis response.

Social and health psychology entered the pandemic with a large toolbox of personal, social, and societal-level theories that may all independently predict individual-level infection prevention behavior to some extent. These individual health theories each involve some overlapping and some distinct predictors. However, when numerous disconnected studies use disparate research methods, levels of analysis, limited samples, and narrow contexts, it is difficult to compare the relative predictive utility of variables indicated by these theories. In other words, when any given study focuses only on the variables that fall within the scope of its theory, it is hard to tell how important the variables are, relative to other variables considered by other theories (or variables not considered at all). Machine learning is a more holistic methodology, as it can assess and compare a large number of potential predictors simultaneously, including theoretically relevant ones, and identify which predictors ultimately explain the most variance in the outcome measure of interest.

The aim of this study is to use machine learning to identify the most important predictors of infection prevention behaviors during the early stages of the COVID-19 pandemic from a multinational, rapid-response survey. We combine multi-national survey data, country-level secondary database integration, and machine learning methods with the practical aim of identifying the most important predictors that could serve as targets for future research and behavioral interventions by governments and organizations such as the WHO. This method offers a holistic evaluation of numerous candidate predictor variables. The candidate variables cover different theoretical domains, so the results might speak to the relative importance of different theories as well as specific predictors. Moreover, the results of this inductive, exploratory approach might suggest promising avenues for future confirmatory research, to investigate direction of causality, and could support the allocation of scientific resources towards the most promising predictors of compliance in future crises that resemble the current pandemic. Results can also provide input for theory development or refinement[2] .

Our study was conducted between March to May of 2020 – that is, in the initial phase of the pandemic, several months before the first COVID-19 vaccine (Pfizer-BioNTech COVID-19) was approved by the US Food and Drug Administration in August of the same year. At the time, there was hope a future vaccine could bring an end to the pandemic, implying that behavioral measures were mainly an interim or short-term solution. However, by 2021, hopes surrounding vaccines had still not fully materialized, partly because the available vaccines waned in efficacy over time and across new virus strains, and because much of the global population remained unvaccinated (e.g., COVID-19 vaccine hesitancy has since become a major area of research[3,4]. By winter 2022, with new virus strains, recurring lockdowns, and the return of behavioral restrictions, the infection prevention behaviors recommended during the initial period of our study remained highly relevant.

Machine Learning Can Identify Candidate Predictors

Machine learning can complement theory-driven approaches by identifying important determinants, or correlates, of a particular outcome, identifying blind spots in existing knowledge, and ranking predictors by their relative importance[2]. Machine learning instead estimates predictive performance in new datasets, and thus, generalizability of the results. Further, it includes checks and balances to prevent spurious findings (i.e., overfitting; see[5]). The random forests algorithm, in particular, is free from certain assumptions of regression/correlation analysis, namely the assumption of linearity, absence of interactions, and normality of residuals. Random forests intrinsically capture non-linear associations and higher-order interaction effects, and can account for multilevel data: The clustering variable can be included as a predictor, which allows for relationships to differ across clusters (e.g., if measurement or associations differ between countries)[6].

Our approach incorporated both individual-level (psychological) predictors and country-level (societal) variables. To identify key individual-level predictors of infection prevention behaviors—at least during the initial phase of the pandemic—we launched a large-scale psychological survey in 28+ countries in the immediate weeks after the World Health Organization (WHO) declared COVID-19 a pandemic. The survey was designed with country-level database integration and machine learning in mind, and a separate team set out to perform machine learning analysis in isolation of any confirmatory analysis. The *a priori* objective was to recruit tens of thousands of survey responses globally, to assess their attitudes towards and to society's prescriptions, and examine how these factors relate to individual infection prevention behaviors. The survey provided individual-level variables, such as basic demographic characteristics (e.g., gender, age, education, religiousness), brief self-report measures of various psychological factors (e.g., subjective states and well-being, work and financial concerns, societal attitudes, COVID-relevant attitudes and beliefs), and individual infection prevention behaviors (e.g., hand washing, avoiding crowds).

Deductive and Inductive Approaches

Deductive research, or hypothesis-testing, is the predominant focus of contemporary behavioral research. It tends to focus on a relatively narrow set of theoretically-derived variables, and the results revolve around statistical inference: Whether the theoretical hypotheses are supported by significant or reliable effects. In deductive research, less emphasis is placed on comprehensiveness or breadth of candidate predictors. Relatedly, the relative importance of different predictors is often of secondary importance, as is the model's predictive performance. Thus, although an advantage of deductive approaches is that they can be used to draw inferences about theoretical hypotheses, they also have specific limitations. These are particularly poignant in the context of the COVID-19 pandemic. To allocate scientific resources effectively in a crisis, it is important to cast a wide net among potential predictors, across different theories, and even include under-theorized factors to unearth potential blind spots in the extant literature. Inductive research–that is, rigorous exploratory work that identifies reliable patterns in data, is more suited to these demands.

In recent years, inductive research has been gaining traction as a technique to complement existing theories by identifying important omissions[2]. In particular, machine learning offers powerful new tools for systematic exploration that can identify relevant predictors and complex relationships that have eluded theoreticians[7]. Machine learning is an approach to data analysis that focuses on maximizing predictive performance. This involves the use of flexible models to find reliable patterns in data. Machine learning models can distill a large set of candidate variables down to the ones that are most important in predicting the outcome of interest, and also indicate the direction and shape of the marginal association between those predictors and the outcome. In a context where predictor variables are likely to be related to each other, machine learning is better suited to manage these complex relationships than, e.g., multiple regressions. Moreover, it incorporates checks and balances to prevent spurious findings [5]. However, it is important to note that inductive and deductive approaches are interwoven, as the set of variables used as input for a machine learning analysis is typically based on theoretical considerations. Thus, as we describe below, we included in our survey a large set of candidate individual- and societal- level indicators, of infection prevention behavior, that were of theoretical interest to our international group of psychology experts.

Relevant Theory

Infection control that relies on individual compliance with health recommendations constitutes a *public good*. The main characteristic of public goods (e.g., clean air) is that people can benefit from it even if they have not contributed to its production or purchase. This creates the temptation to free-ride on the contributions of others [8,9]. The COVID-19 pandemic has some characteristics of a public goods dilemma, in that control of the virus can only be achieved if most members of society contribute to the effort [8,9]. However, a pandemic also differs from many other public goods dilemmas, due to the immediate personal health threat of the virus: engaging in infection prevention behavior not only reduces the societal spread of the infection, it also lowers *individual* infection risk. Accordingly, individual-level psychological factors could predict infection prevention behavior even when individuals feel unobserved [10-12]. Thus, we might expect self-reported individual differences to predict compliance, such as perceived personal infection risk and vulnerability.

Beyond its potential as a public goods dilemma, the COVID-19 pandemic is also a health emergency with profound social, economic and societal ramifications. In practical terms, millions of people were expected to lose their jobs, experience economic hardship, and suffer psychological strains as result of the lockdowns or self-quarantining [13]. More generally, an international group of behavioral scientists proposed various other psychosocial factors that may predict responses to the COVID-19 pandemic[14], ranging from individuals' internal states to their societal attitudes and beliefs. This necessitated research that comprehensively (re-)examined potential predictors of infection prevention behavior, with attention to the broad social, economic, and personal ramifications of the pandemic.

Our survey also included factors directly relevant to the domain of health behavior, such as those suggested by the Health Belief Model[15,16]. According to the Health Belief Model, two conditions must be met to motivate people to engage in COVID-19 infection prevention behavior: They have to believe that they are at risk of contracting the virus, and that engaging in the recommended virus protection behaviors would be effective in reducing that risk[15]. A further assumption of this model is that the effect of perceived effectiveness of a health behavior will be moderated by the perceived costs of engaging in that behavior. If the behavior is too effortful, people might not adopt it, even if they think that doing so would be effective. A second relevant theory is the Theory of Planned Behavior (TPB[17-19]). This more general psychological theory of behavior prediction posits that intentions to engage in a specific behavior would be predicted by three constructs: attitude towards the behavior (advantages and disadvantages), subjective norms (e.g., what is expected of me by important others), and perceived behavioral control (i.e., will I be able to do it).

Despite the potential relevance of health behavior theories, they illustrate the aforementioned tendency of deductive research to focus on a narrow set of theoretical constructs. Other potentially important predictors, not germane to the given theory, might be overlooked. In line with this narrow focus, models based on such theories typically explain limited variance in the outcome variable. For example, a meta-analysis based on 185 independent tests of the TPB found that attitudes, subjective norms and perceived control explain 39% of the variance in intention, with intention accounting for 22% of variance in behavior[18]. Although this descriptive performance is perceived as relatively strong in the field of social science, it still leaves room for potential predictors from other research domains. Thus, rather than focus exclusively on variables that target the health behavior, the present analysis casts a wide net, by including psychological and societal factors that specifically pertain to the COVID-19 domain, as well as other factors whose relevance may generalize across domains.

The Present Study

We sought to distinguish important individual- and societal- level indicators of infection prevention behavior using random forests [6]. The analysis is based on data from a large-scale psychological survey enriched with publicly available country-level secondary data. Random forests was used for its relatively competitive performance, computational inexpensiveness, and ease of interpretation[20]. The expected results consist of an estimate of predictive performance, which indicates how well the final model predicts infection prevention behavior in a new sample; a ranking of predictors based on variable importance, which reflects their relative contribution to the model's predictive performance; and partial dependence plots, which reveal the direction and shape of each predictor's marginal association with the outcome.

The specific approach used in this paper maximized the reliability and generalizability of results in three ways. First, the data were split into a training sample, used to build the model, and a testing sample. The testing (or "hold out") sample is never used in the initial analysis, but rather is used to estimate the generalizability of the final model after analyses on the training sample are complete (*a priori* splitting of the dataset can be verified via the project's public historical record). This procedure helps to determine the model's predictive performance: In a classic deductive analysis, performance is traditionally expressed in terms of R^2, which reflects a theoretical model's *descriptive* performance: the percentage of variance in the outcome explained by the model in the data. In the machine learning literature, by contrast, it is commonplace to estimate *predictive* performance by assessing R^2 in an independent test sample that was not used to estimate the model. Predictive performance reflects the generalizability of a model. Second, part of our global data collection efforts included the recruitment of paid subsamples from 20 countries that were representative of the population's age and gender distribution. Such sampling procedures can improve generalizability to the extent that it includes persons who might otherwise not participate as self-selected volunteers. Third, random forests is a specific machine learning method that includes checks and balances to ensure reliability and generalizability of the results[6]. Random forests analysis accomplishes this by splitting the training data into 1000 bootstrap samples, and estimating a regression tree model on each of these bootstrap samples independently. Each regression tree in turn splits the sample recursively until the post-split groups reach a minimum size. A split is made by determining which predictor (out of a randomly selected subset of predictors) and value of that predictor maximizes the homogeneity of the post-split groups. Thus, a tree resembles a flowchart with relatively homogenous end nodes. Interactions are represented by subsequent splits on different variables; non-linear effects are represented by repeated splits on the same variable; random effects are represented by splits on the cluster variable (country) followed by splits on substantive variables. Naturally, each of these 1000 models will include some spurious findings (overfitting). However, when the predictions from the 1000 models are averaged, these spurious findings tend to balance out, thus leaving only the reliable patterns. Whether this approach is successful in identifying reliable and generalizable patterns can be objectively evaluated based on subsequent predictive performance on the hold-out (test) sample.

Results

For a complete archive of all analysis code and results, including fit tables and figures, see https://github.com/cjvanlissa/COVID19_metadata.

Data Analytic Plan

Prior to analysis, we split our data by randomly assigning 70% of observations to a training set and 30% of observations to a test set[5]. The test set was reserved exclusively for unbiased evaluation of the final model's predictive performance, and was neither used nor examined during model building to prevent cross-contamination. Thus, all models were trained using the training set and evaluated using the test set. We applied a random forest model using the ranger R-package[53]. Random forests offer competitive predictive performance at a low computational cost, intrinsically capture non-linear effects and higher-order interactions, offer a single variable importance metric for multi-level categorical variables (such as country), and afford relatively straightforward interpretation of variable importance and marginal effects of the predictors[6]. With regard to the multilevel structure of the data, random forests inherently accommodate data nested within country, including cross-level interactions where a given predictor has a different effect in different countries.

The forest included 1000 trees. The model had two tuning parameters: the number of candidate variables to consider at each split of each tree in the forest, and the minimum node size. The optimal values for these parameters were selected by minimizing the out-of-bag *mean squared error* (MSE) using model-based optimization with the R-package tuneRanger[54]. The best model considered 31 candidate variables at each split, and a minimum of six cases per terminal node.

The outcome metrics considered in the present study consist of 1) predictive performance, which reflects the model's ability to accurately predict new data; 2) variable importance, which reflects each predictor's relative role in accurately predicting the outcome measure, and 3) partial dependence plots, which indicate the direction and (non)linearity of a specific marginal effect[6]. Predictive performance is, essentially, a measure of explained variance (R^2), except that in the machine learning context, predictive performance is evaluated on the test sample, which was not used to estimate the model. Estimates of R^2 on the training sample should be interpreted as a measure of descriptive performance (i.e., how well the model describes the data at hand), and can be (severely) positively biased when used as an estimate of predictive performance in new data. Given that we had recruited paid subsamples (age-gender representative) in 20 countries, we additionally computed predictive performance for the paid-only portion of the test sample, to better examine the generalizability of our findings to the target population.

The relative importance of predictor variables is based on permutation importance: Each predictor variable is randomly shuffled in turn, thus losing any meaningful association with the outcome, and the mean decrease in the model's predictive performance after permutation, as compared to the un-permutated model, is taken to reflect the (inverse) importance of that variable[6].

The partial dependence plots are generated using the metaforest R-package[4]. Partial dependence plots display the marginal (bivariate) association between each predictor and the outcome[55]. They are derived by computing predictions of the dependent variable across a range of values for each individual predictor, while averaging across all other predictors using Monte Carlo integration.

Total Variance Explained

The random forest model predicted a large proportion of the variance in self-reported infection prevention behaviors in the full test sample ($R^2_{test} = .523$), as well as in the paid subsample ($R^2_{rep} = .586$). As these samples had not been used in model estimation, this indicates that the results are robust. Notably, the high predictive performance on the paid subsample indicates the generalizability of the findings. The explained variance in the training sample was of approximately the same magnitude ($R^2_{train} = .518$). This correspondence between training and testing R^2 indicates that the model successfully learned reliable patterns in the data, and was not overfit.

The top 30 predictors, ranked by relative variable importance, are illustrated in Figure 1, along with an indication of whether the effect is generally positive, negative, or other (e.g., curvilinear). Table 1 serves as the legend for the variables illustrated in Figure 1. Table S3 provides full results of all 115 predictors, rank-ordered by variable importance.

Consistent with expectations, the most important predictors of infection prevention behavior included a mix of individual-level (survey) variables and country-level (database) indices. The shape of the bivariate marginal association between each predictor and the outcome is displayed in the partial dependence plots (Figure 2). Recall that partial dependence plots display the marginal relationship between one predictor and the outcome, while averaging across all other predictors using Monte Carlo integration [55]. Note that the marginal predictions for the two levels of "leave for work" are identical; a denser Monte Carlo integration grid might show a small difference here, but exceeds our computational resources.

Individual-level Predictors

Social Norms. By far the most important predictors of infection prevention behaviors were individual-level beliefs about how other people should behave, and whether society should mandate infection prevention behavior. The two strongest predictors were injunctive norms targeting infection prevention – namely, the belief that people in the community *should* engage in social distancing and self-isolation (ranked 1st), and their endorsement of extraordinary restrictive measures to contain the virus (mandatory quarantines and vaccination; reporting suspected infected individuals, ranked 2nd). The third strongest predictor was a pro-social willingness to protect vulnerable groups from the coronavirus (3rd). Respondents who complied with the norm to engage in infection prevention behaviors indicated that they wanted to do their bit to help other people to cope with the pandemic. Other, related indicators included the descriptive normative belief that people in one's community *do* self-isolate and engage in social distancing (ranked 7th), a pro-social willingness to limit the economic consequences of the coronavirus on others (8th), and support for economic intervention (26th). Partial dependence plots indicate that the injunctive ('should') norm had a positive, approximately exponential, marginal relationship with the outcome measure, whereas the other indicators had positive, approximately linear marginal relationships.

Social and Public Behavior. The next most important indicators were behavioral correlates of the dependent measure, namely, self-reported days in the last week that the individual engaged in social and public contact. Each of these behaviors had a negative, approximately linear relationship with infection prevention behaviors. This included the number of days that respondents reported leaving home (5th), the number of days in the past week they had in-person (face-to-face) contact with people who live outside their home, including "…immigrants" (4th), "…other people in general" (6th), and "…friends and relatives" (20th). Thus, higher in-person contact, which is inadvisable during a pandemic, generally corresponded with less infection prevention behavior. In contrast, online (virtual) contact with friends and relatives—which does not violate social distancing measures—positively predicted infection prevention behavior (ranked 25th).

Personal Psychological Factors. A third set of individual-level predictors thematically pertained to personal and psychological resources and all had positive linear relationships with the outcome variable: a problem-focused coping style (9th), having high hopes that the coronavirus situation would soon improve (11th), and a temporal focus on the present (16th) and/or the future (17th). Consistent with theories of health behavior[44], the perceived personal consequences of coronavirus infection ranked 10th. Relatedly, self-reported knowledge about COVID-19—important for risk-assessment—ranked 28th.

Several individual-level variables rounded out the bottom of the list. These are harder to interpret, because of their lower variable importance and non-conclusive partial dependence plots. Having to leave one's house for work (ranked 29th) had a slight negative association with infection prevention behavior, perhaps because having to leave the house for extrinsic reasons hinders social distancing and self-isolation. The positive association between conspiracy beliefs and infection prevention behavior (ranked 23rd) might seem paradoxical, as one might expect a negative association, if we had specifically measured belief in the conspiracy theory that the virus is a hoax. However, we instead assessed generic conspiracy beliefs[38] – whether respondents believe that politicians do not always disclose the motives behind their decisions, that important things happen without public knowledge, and that government agencies closely monitor citizens. It might be the case that participants who endorse these beliefs tend to take infection prevention into their own hands.

Country-level Predictors

General Societal Conditions. Five (of 9) general societal indices were ranked among the important indicators of infection prevention behaviors. The most important country-level predictor was a WHO/OECD indicator of national health care resources and infrastructure: the number of doctors per 10,000 inhabitants (ranked 12th). Other country-level predictors were the Global Health Security index (ranked 22nd), which pertains to pandemic preparedness and general health security, and two (out of six) World Governance Indicators: political stability (15th) and government effectiveness (27th). Country-level COVID-19 policy stringency (i.e., severity of lockdown conditions) ranked 30th, which potentially illustrates the limits of government lockdowns in compelling individual-level behavior, relative to other predictors.

COVID-19 Conditions. All three indicators of objective COVID-19 virus spread conditions in participants' countries at the time of participation were important indicators of infection prevention behavior: the cumulative number of confirmed COVID-19 cases (ranked 14th), deaths (19th) and recoveries (21st). All three patterns were negative, indicating that self-reported infection prevention behavior was lower among respondents who lived in countries with higher virus case counts, deaths, and recoveries on the day that they responded to the survey.

The Effect of Time

As our study covered a span of several weeks, time could be included as a predictor, operationalized as the calendar date of each survey response. The effect of time was negative (13th), indicating that self-reported infection prevention behavior generally decreased between March and May 2020.

Discussion

The present study used machine learning to identify and rank predictors of infection prevention behavior among a wide set of potential candidates. After training on one sample, the resulting random forests model predicted over 50% of the variance in self-reported infection prevention behavior in a second (test) sample. This exceeds the standards for explained variance of social and health psychological theories, thus indicating that this data-driven approach can complement theoretical models. Moreover, whereas theoretical models typically focus on a limited narrow set of relevant variables, the present machine learning analysis identified additional, undertheorized predictors (e.g., temporal focus), thus offering complementary insights.

Who Complies with Infection Prevention Behavior?

A coherent picture emerged from our analysis of the type of person that showed early compliance with the recommended set of infection prevention behaviors. The underlying pattern of individual-level indicators could point to an intuitive understanding that infection control is a public good and, a conviction that the only way of virus mitigation involves widespread compliance with recommended behaviors. The compliant individuals appear to understand that factors such as personal risk (which was not indicated as highly important) is managed through similar efforts from others. If everybody engaged in infection prevention behavior, the number of infected people in society would be reduced. Furthermore, if the people who did contract the virus maintained physical distancing, they would be less likely to infect others. This would explain why the strongest correlates of infection prevention behavior were beliefs that others in the community should engage in social distancing and self-isolation and that society should take restrictive measures to enforce that behavior, such as mandatory quarantine, reporting people suspected to be infected, and (eventually) mandatory vaccination. Endorsement of such measures implies the prioritization of infection control over concerns about people's liberties and autonomy.

The descriptive normative belief, that other people in the community *do* engage in social distance and self-isolation, also emerged as a relatively important predictor. It makes sense that individuals might be less motivated to comply if they were among a community of non-compliers. Furthermore, according to their self-reports about their own behavior, compliant individuals did not engage in behavior that would be inconsistent with self-protection, such as leaving their homes or having personal contacts with other people. If they had contacts with their family and friends, it was not in face-to-face meetings, but online.

The findings also point to the idea that people who comply with recommended infection prevention behaviors are forward-looking problem-solvers. That is, they tended to engage in a problem-focused coping style, focus on the present and the future (rather than dwell on the past), and maintained high hopes that the coronavirus situation would soon improve. This optimistic view is important because these individuals were likely aware of the costs of these infection prevention behaviors and perhaps needed psychological resources to alleviate these costs. In this vein, other important predictors were a pro-social willingness to self-sacrifice to protect vulnerable groups from the virus, to limit the economic consequences of the coronavirus on such groups, and to support collective interventions in the economy such as tax increases. These results might also help understand the tension between members of society who do and do-not engage in updated recommendations. Given that the largest predictor of infection prevention behaviors—at least those originally recommended by the WHO—is the injunctive normative belief that one *should* participate in the behaviors, people who do not engage in those behaviors are likely to be seen as immoral, or at the very least norm-violators. Supportive of this, a large British survey indicated in September 2020 — three months after the WHO started to universally recommended mask wearing — that 58% of the mask wearers in Britain had severely negative attitudes towards those who did not wear masks and 68% of Brits who complied with lockdown rules had strong negative views about lockdown rule breakers. In fact, significant minorities who kept to the rules said that they "hated" those who did not[56].

Aside from individual-level factors, several country-level indicators emerged as important predictors. This pattern of results is noteworthy for several reasons. First, because it means that there are meaningful between-country differences in compliance, which are partly explained by country-level characteristics. Second, the absence of the variable "country" from the top predictors indicates that there are no remaining between-country differences in compliance to be explained, once the effect of the included country-level predictors is accounted for. Thus, it is unlikely that other between-country differences – such as collectivism/individualism – have a meaningful effect over and above a country's health care resources (e.g., number of doctors) and pandemic severity. Third, whereas it could be argued that the effect of individual-level predictors might be inflated due to common method bias, this explanation can be ruled out for the country-level predictors. The fact that these factors were among the most important predictors thus speaks to the robustness of the findings.

The findings regarding country-level predictors further suggest that infection control is a societal-level challenge, in that individual-level compliance with infection prevention recommendations is more likely in a society that has the political stability and health care infrastructure to take effective action to contain the virus and treat people who have become infected. The findings regarding country-level indicators are consistent with this analysis: government stability and effectiveness, pandemic preparedness and health care resources (i.e., number of doctors), pandemic preparedness and lockdown stringency, were all relatively important indicators of infection prevention behavior.

Respondents in countries with higher confirmed COVID-19 infections, deaths, and recoveries reported less infection prevention behavior themselves. Such findings might suggest reverse causality, as a country is likely to experience increased pandemic severity if its citizens do not endorse infection prevention behaviors. Alternatively, it is possible that higher virus counts demotivate infection prevention efforts—though, this assumes widespread individual-level knowledge about virus rates. Given that self-reported knowledge about COVID-19 was an important positive indicator, it is more plausible that, in a society in which there is less compliance, there will be more infections, deaths, and recoveries.

Finally, one worrisome association is that time since the start of the pandemic, operationalized as date of participation, emerged as an important negative predictor of personal health behavior. This suggests that, even in the early phase of the pandemic, there was already a decrease in compliance with government advice. It could be that with time, self-isolation and social distancing became unbearable for many people. This is consistent with the notion of ‘COVID-fatigue’, and highlights the need to investigate what factors might promote more sustained adherence to infection prevention behaviors.

Unexpected Absences from Top Indicators

It is interesting to consider some of the other 85 variables that were not among the top indicators. From a health psychological perspective, it is surprising that the perceived personal likelihood of getting infected was not among the important predictors. Though, the perceived personal consequence of infection was ranked 10th. According to the Health Belief Model[15], perceived vulnerability and severity are both central to health threat appraisal. The fact that the perceived severity of getting infected was a highly ranked predictor, but perceived infection risk was not, might suggest that people's behavior is more strongly driven by expected consequences than probability. Alternatively, the link between compliance and infection risk might be smaller because people implicitly recognize that this risk is largely outside of their control, to the extent that the pandemic constitutes a public goods dilemma.

Several other, theoretically relevant variables that were absent from the most important predictors, included loneliness and boredom, emotional and affective states experienced during the last week, subjective well-being, various forms of psychological and financial strain, and job insecurity. It is important to note, however, that the present analysis does not rule out the importance of these personal factors for other outcomes, nor does it serve as evidence for a null-effect.

No demographic variables emerged as especially important, even though several are associated with increased risk of complications from COVID-19. For instance, elderly people are at higher risk to die from a COVID-19 infection and are therefore strongly advised to take great care[21]. Furthermore, there is reason to assume that social distancing and self-isolation present more of a dilemma to young rather than elderly people, especially those on a pension. For young people, the costs of social distancing and self-isolation are typically higher and – because they usually recover more easily from a COVID-19 infection – the rewards of those infection prevention behaviors are smaller. Consistent with this argument, the media have framed the pandemic as a potential “intergenerational conflict of interest”, where the young bear the brunt of the cost of containment measures, whereas the elderly enjoy most of its benefits. It is therefore noteworthy that our analysis did not identify age as an important predictor. However, this finding is consistent with preregistered research that similarly found no support for the “intergenerational conflict of interest” hypothesis[57].

Limitations, Strengths, and Future Directions

An important strength of this study is that the questionnaire used was designed by an interdisciplinary consortium of scientists from different countries. This resulted in a questionnaire with a broader scope than those guided by a singular theoretical perspective. It makes the resulting data ideally suited for a machine learning analysis that can distill the most important predictors from many potential candidates. However, despite this broad scope, it is important to acknowledge that this study covered only a small fraction of available psychological and societal factors. Similar studies are recommended to identify other important predictors of virus prevention behaviors, including related behaviors that emerged later in the pandemic, such as the wearing of face coverings and vaccination.

Another strength is the very large international sample, which made it possible to apply machine-learning methods to identify important patterns in the data. Additionally, the availability of an age-gender representative subsample improved the generalizability of the findings. Finally, a noteworthy strength is that the variance explained by the model was consistently high, and approximately the same, in the sample used to train the model (R^2_{train} = .52), and in the testing sample used to estimate the robustness of the findings (R^2_{test} = .52), and in the age- and gender-representative testing sample used to estimate generalizability of the findings to the target population (R^2_{rep} = .59). This indicates that the model captured reliable patterns in the data, without overfitting noise and spurious effects, and has high generalizability.

There are also limitations in the methods and sampling. A methodological trade-off was made due to the urgency of the crisis: In order to respond rapidly to the pandemic onset in March 2020, with a large-scale cross-national study, while relying on volunteer efforts and limited funding, the choice was made to use exclusively self-report measures, which are easily translated and administered to large-scale samples at low cost. Of course, the use of self-report measures risks introducing variance due to the subjective nature of self-reports, and common method bias between self-reported predictors and the outcome. A second methodological limitation—one shared with all non-experimental research, is the question of causality. For some of the included predictors, causal mechanisms may be known or suggested by theory; for others, future research will be needed to examine whether causal relations exist; and for others still, causality might be unlikely. We have taken care to discuss the associations observed through the lens of past theory. Since causality cannot be inferred from these results, the primary contribution of this study is the rapid reduction of a large number of candidate predictors to a smaller subset of those most strongly associated with the outcome of interest. This allows researchers to prioritize the most likely candidate predictors for future research, and helps policy makers focus their efforts on the most influential predictors for which causal mechanisms are known or suspected. Conversely, it is also useful to know which factors are *not* strongly associated with virus prevention behaviors, as policies that target these factors are unlikely to be effective. For some variables, causality might be unlikely, but these might still be helpful from a descriptive point of view, or to decide who to target in interventions, or to contextualize the relative importance of other variables.

A third limitation pertains to the sampling: Although efforts were made to recruit age-gender representative subsamples, even these subsamples will not be strictly representative of the target population. Moreover, they could be otherwise biased by other, potentially unknown characteristics—including the different virus strains, and shifting societal responses of the pandemic. Nonetheless, the approximately stable model performance across all samples reduces the likelihood that generalizability to the target population would be substantially different.

The analysis of this study uses deductive methods maximize predictive performance, typically explain more variance than purely deductive approaches, and in the case of random forests, intrinsically capture non-linear effects and higher order interactions, including between-country differences in effects. However, the results are harder to interpret than the parameters (e.g., regression coefficients and *p* values). We should note that the variables included in the PsyCorona survey were guided by theory, and thus our approach combines inductive and deductive approaches. Thus, although our application of machine learning is useful for gaining preliminary insights, it also capitalizes on a rich history of theorizing about what drives engagement in health behavior. However, although our study includes potentially important variables and theoretical areas, it is neither exhaustive nor conclusive. Inductive analysis can complement theories or provide an impetus for the development of new hypotheses, but the output is not yet a comprehensive theory.

Nevertheless, the present research contributes to the literature by offering a large scale cross-national psychological survey, enriched by database integration, and analyzed using machine learning.

Given that external enforcement of infection prevention behaviors is difficult, recommendations are most likely effective if they are internalized by individuals and supported by societal-level factors. The picture that emerges from this analysis is that early compliance with infection prevention behavior recommendations is partly psychological and partly contextual. Our findings suggest a strong emphasis on norms—both injunctive and descriptive—and the societal conditions enabling these norms.

Although the data collected describe infection prevention behaviors during the beginning of the pandemic, they may be useful for understanding later patterns of behavior (e.g, low vaccine rates) or future crises that involve a combination of personal and societal risk. Health behavior theories tend to focus on the intrapersonal factors that predict behavior, possibly because these seem proximal to the health behaviors of interest. However, our data suggest these proximal factors may predict less variance in behavior than broader considerations of communal behavior. Future models may benefit from considerations of perceptions of norms in conjunction with personal risk when they are applied to other health behaviors as well.

Conclusions

We began with an assumption that control of the pandemic is analogous to a public goods dilemma, in that COVID-19 is a social challenge that, in the absence of a vaccine at the time of the study, could only be addressed if enough individuals engaged in infection prevention behavior. In accordance with this assumption, social beliefs and societal factors, rather than exclusively personal psychological states, emerged as the main predictors in our analysis.

Experimental Procedures

Resource Availability

Lead Contact. The lead contact for this paper is Dr. Caspar van Lissa, who may be contacted at C.J.vanLissa@uu.nl.

Materials Availability. The full survey is available in the supplemental material, as well as codebooks and translation procedures for all languages (tables S1 & S2). All analysis code is available in an online repository (https://github.com/cjvanlissa/COVID19_metadata), which also includes a full historical record since the start of the project. This can be used to verify that the analysis proceeded transparently and straightforwardly; the random seed used to select participants for the test sample was established before access to data was obtained, and testing data were never used for model training.

Data and Code Availability. The data and code used in this analysis are available at DOI: 10.5281/zenodo.5948816

PsyCorona Survey: Recruitment and Item Selection

The survey was translated from English into 29 other languages by bilingual members of the international research team. It was distributed online during the early phase of the pandemic (March-May 2020), with most participants completing the survey in March and April (see figure S1 for daily frequencies). Parallel sampling strategies were employed: convenience sampling, snowball sampling, and paid sampling. Given that age and gender were identified early as population vulnerability characteristics to the virus[21,22], the self-selected samples were supplemented with paid subsamples that were representative of a given country's population distribution of age and gender. The panel firms Qualtrics Panels and WJX achieved age-gender representative samples in 20 countries (n ~ 1000 per country): Argentina, Australia, Brazil, Canada, China, France, Germany, Italy, Japan, Netherlands, Philippines, Romania, Russia, Serbia, South Africa, South Korea, Spain, Turkey, United Kingdom, and the United States. Four additional countries only achieved gender representativeness, due to insufficient access to the 55+ age group in Greece, Indonesia, Saudi Arabia, and Ukraine. These paid subsamples were used to improve the generalizability of the model.

In order to maximize project feasibility (e.g., each item was translated into 30 languages), increase survey breadth, and reduce participant burden, we used brief measures of each construct. Where possible, survey items were selected from established scales. Because the set of variables relevant to the pandemic (e.g., norms about handwashing, endorsement of stringent regulations for violating quarantine) did not exist prior to the pandemic, we crafted face valid items to assess these constructs.

Although the PsyCorona study was designed and implemented prior to Van Bavel and colleagues'[14] discussion of candidate domains of inquiry for pandemic behavior, it touches on nearly all of these topics, including navigating threats, stress and coping, science communication, moral decision-making, and political leadership. The survey covered three overarching themes. The first theme included personal factors that could affect individuals' capacity to respond to the virus, such as psychological coping and outlook, loneliness and deprivation, subjective emotional states, well-being, employment, and financial (in)security. The second theme pertained to social attitudes and norms, including general beliefs and attitudes about society, economic considerations, migrant attitudes and prejudice, perceived and preferred social norms for infection prevention, and endorsement of extraordinary virus containment and its economic rescue measures. The third theme pertained to virus-relevant personal concerns, values, and tendencies, including social contact and leaving the home, as well as the dependent variable of interest: self-reported engagement in voluntary, infection prevention behaviors recommended by the WHO. Personal factors adapted or informed by prior work included affective states (incl. valence and arousal[23]); boredom[24]; coping and avoidance[25,26]; financial strain[27]; loneliness[28]; neuroticism[29]; happiness and well-being[30-32]; time perception, management, and temporal focus[33,34], working conditions and job insecurity[35-37]. The social attitudes and norms domain included generic conspiracy beliefs and paranoia[38,39]; immigrant attitudes[40-42]; norm perceptions and preferences (adapted[43]); societal discontent and disempowerment[44-45]. Virus-relevant personal concerns included perceived norms (both descriptive and injunctive, adapted[46]); virus-relevant beliefs and perceived knowledge, virus exposure risk and economic risk, and severity of virus and economic consequences (adapted[46,47]); trust in government pandemic communication and response (adapted[43,48,49]), and attitudes towards prosocial responses and extraordinary societal responses[48]. This list is not exhaustive; see table S3 for a full list and item details and our OSF page for a full list of references for each item (https://mfr.de-1.osf.io/render?url=https://osf.io/7kfj5/?direct%26mode=render%26action=download%26mode=render).

Key demographic variables, such as age, gender, education level, and religiousness were included as predictors. Country of residence was included as a categorical predictor. A summary table of all variables entered as predictors is available in (table S3. Psychometric properties of scales, including reliability and the range of factor loadings, are available in table S5. There was no evidence of multicollinearity among the continuous individual-level predictors, with all variance inflation factors between 1.11 - 2.66.

Infection Prevention Behavior. Through May 2020, a set of three infection prevention behaviors were advised across most countries and contexts: washing hands, avoiding crowds, and self-isolation/self-quarantine (wearing a face covering was not universally recommended by the WHO until June 2020[50]). Participants were presented with a single screen that read, *"to minimize my chances of suffering from coronavirus, I..."* and indicated their agreement to *"1. ...wash my hands more often"*, *"2. ...avoid crowded spaces"*, and *"3. ...put myself in quarantine/self-isolate"*, each rated on a seven-point scale rated -3 (strongly disagree) to 3 (strongly agree). To ensure items could be combined into a unidimensional scale, we conducted Horn's parallel analysis[51]. Only one component had an Eigenvalue exceeding randomly permuted data. This component explained 70% of the variance in the three items, which is high. The three factor loadings were high and approximately equal in size (range: .78 - .89), indicating that it is justifiable to combine these three items into a mean score representing infection prevention behaviors (M = 2.20, SD = 1.00, α = .75). Note that the items were specifically framed to assess the behavioral intent to reduce the risk of infection, consistent with theories of health behavior that people engage in self-protective actions because they are perceived as instrumental for threat reduction[46].

Data Enrichment and Data Cleaning

We enriched the individual-level PsyCorona data with publicly available country-level datasets. These datasets were selected due to their international relevance for affording, shaping, or guiding individual-level behavioral responses to the virus: First, pandemic severity, as indicated by the number of cases, deaths, and recovered patients. Second, pandemic-related policies including both preexisting policies and ongoing governmental response to the COVID-19 pandemic. Third, pandemic preparedness. Table 2 presents an overview of the databases. The time range in data collection afforded variability in the degree to which people in a given country were seeing cases and/or engaging in different containment policies. Where applicable, respondent's country-level data were matched to their date of participation (e.g., confirmed cases, lockdown severity). Altogether, there were 115 predictors (80 survey factors, 35 country-level factors).

We subsequently cleaned the data in several steps. First, to ensure that there was enough data on country-level, we excluded observations from countries that accounted for less than 1% of total observations. The final sample included *N* = 56,072 respondents across 28 countries (see table S4 for samples that remained in the data). Second, we excluded any columns and rows from the data that had a proportion of missing values of more than 20%. Third, we computed mean scores for multi-item scales using the tidySEM R-package[52]. For instance, responses to all 4 items on job insecurity[37] were summarized by creating a single composite score for job insecurity. Scales with low reliability were excluded (*Cronbach's alpha* < .65). See table S5 for scale descriptive statistics, including reliability and range of factor loadings.

Acknowledgements

"This research received support from the New York University Abu Dhabi (VCDSF/75-71015), the University of Groningen (Sustainable Society & Ubbo Emmius Fund), and the Instituto de Salud Carlos III (COV20/00086) co-funded by the European Regional Development Fund (ERDF) 'A way to make Europe'"

Author Contribution Statement

The study was designed by Caspar J. van Lissa, Wolfgang Stroebe, Michelle R. vanDellen, N. Pontus Leander, Maximilian Agostini, Jocelyn Belanger, Ben Gutzkow, and Jannis Kreienkamp. The manuscript was written by Caspar J. van Lissa, Wolfgang Stroebe, Michelle R. vanDellen and N. Pontus Leander. Data analyses were performed by Caspar J. van Lissa, Tim Draws, Andrii Grygoryshyn, and Clara S. Vetter. Georgios Abakoumkin, Jamilah Hanum Abdul Khaiyom, Vjolica Ahmedi, Handan Akkas, Carlos A. Almenara, Mohsin Atta, Sabahat Cigdem Bagci, Sima Basel, Edona Berisha Kida, Allan B. I. Bernardo, Nicholas Buttrick Phatthanakit Chobthamkit, Hoon-Seok Choi, Mioara Cristea, Sára Csaba, Kaja Damnjanović, Ivan Danyliuk, Arobindu Dash, Daniela Di Santo, Karen M. Douglas, Violeta Enea, Daiane Gracieli Faller, Gavan J. Fitzsimons, Alexandra Gheorghiu, Ángel Gómez, Joanna Grzymala-Moszczynska, Ali Hamaidia, Qing Han, Mai Helmy, Joevarian Hudiyana, Bertus F. Jeronimus Ding-Yu Jiang, Veljko Jovanović, Željka Kamenov, Anna Kende, Shian-Ling Keng, Tra Thi Thanh Kieu, Yasin Koc, Kamila Kovyazina, Inna Kozytska, Joshua Krause Arie W. Kruglanksi, Anton Kurapov, Maja Kutlaca, Nóra Anna Lantos, Edward P. Lemay, Jr., Cokorda Bagus Jaya Lemsmana, Winnifred R. Louis, Adrian Lueders Najma I. Malik, Anton P. Martinez, Kira O. McCabe, Jasmina Mehulić, Mirra Noor Milla, Idris Mohammed, Erica Molinario, Manuel Moyano, Hayat Muhammad, Silvana Mula, Hamdi Muluk, Solomiia Myroniuk, Reza Najafi, Claudia F. Nisa, Boglárka Nyúl, Paul A. O’Keefe, Jose Javier Olivas Osuna, Evgeny Osin, Joonha Park, Gennaro Pica, Antonio Pierro, Jonas Rees, Anne Margit Reitsema, Elena Resta, Marika Rullo, Michelle K. Ryan, Adil Samekin, Pekka Santtila, Edyta M. Sasin, Birga M Schumpe, Heyla A. Selim, Michael Vicente Stanton, Samiah Sultana, Robbie M. Sutton, Eleftheria Tseliou, Akira Utsugi, Jolien Anne van Breen, Kees Van Veen, Alexandra Vázquez, Robin Wollast, Victoria Wai-lan Yeung, Somayeh Zand, Iris Lav Žeželj[20], Bang Zheng, Andreas Zick, Claudia Zúñiga[67]contributed to project design, data collection, translation, and review of the manuscript.

Competing Interests

The authors declare no competing interests.

Data Re-use Disclosure Statement

The PsyCorona data were made available for theory-testing studies by the researchers who helped to collect the data. Portions of the PsyCorona data have been previously reported in specific hypothesis tests[57-61]. This machine learning analysis was planned *a priori* as part of the onset of PsyCorona, is the only paper that uses inductive analysis, and is based on the total dataset.

References

1. Omeife, H. O. Coronavirus: Distancing and handwashing could lower flu rates, too. *MedicalXpress* https://medicalxpress.com/news/2020-04-coronavirus-distancing-handwashing-flu.html (2020).

2. Van Lissa, C. J. Metaforest: Exploring Heterogeneity in Meta-Analysis using Random Forests (0.1.2) [R-package]. (2018).

3. Aw, J., Seng, J.J.B., Seah, S.S.Y., & Low, L.L. (2021). COVID-19 vaccine hesitancy – a scoping review of the literature in high-income countries. Vaccines, 9, https://doi.org/10.3390/vaccines9080900

4. Wang, Q., Yang, L., Jin, H. & Lin, L. (2021). Vaccination against COVID-19: A systematic review and meta-analysis of acceptability and its predictors. Preventive Medicine, 150, 106694.

5. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning:*

Data mining, inference, and prediction. (Springer, 2009).

6. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

7. Brandmaier, A. M., Prindle, J. J., McArdle, J. J. & Lindenberger, U. Theory-guided exploration with structural equation model forests. *Psychol. Methods* **21**, 566–582 (2016).

8. Stroebe, W. & Frey, B. S. Self-interest and collective action: The economics and psychology of public goods. *Br. J. Soc. Psychol.* **21**, 121–137 (1982).

9. Olson, M. *The logic of collective action: public goods and the theory of groups*. (Harvard University Press, 2009).

10. Oosterhoff, B., Palmer, C. A., Wilson, J. & Shook, N. Adolescents' motivations to engage in social distancing during the COVID-19 pandemic: Associations with mental and social health. *J. Adolesc. Heal.* **67**, 179–185 (2020).

11. Deutsch, M. & Gerard, H. B. A study of normative and informational social influence upon individual judgments. *J. Abnorm. Soc. Psychol.* **51**, 629–636 (1955).

12. Hagger, M. S. *et al.* Autonomous and controlled motivational regulations for multiple health-related behaviors: Between-and within-participants analyses. *Heal. Psychol. Behav. Med.* **2**, 565–601 (2014).

13. Brooks, S. K. *et al.* The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *Lancet* **395**, 912–920 (2020).

14. Van Bavel, J. J. *et al.* Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* **4**, 460–471 (2020).

15. Janz, N. K. & Becker, M. H. The Health Belief Model: A decade later. *Health Educ. Q.* **11**, 1–47 (1984).

16. Abraham, C. & Sheeran, P. The Health Belief Model. in *Predicting health*

behaviour: Research and practice with social cognition models (eds. Connor, M. & Norman, P.) 28–80 (Open University Press, 2005).

17. Ajzen, I. *Attitudes, personality, and behavior*. (Open University Press, 2005).

18. Armitage, C. J. & Conner, M. Efficacy of the Theory of Planned Behaviour: A meta-analytic review. *Br. J. Soc. Psychol.* **40**, 471–499 (2001).

19. Robin, R., Mceachan, C., Conner, M., Taylor, N. J. & Lawton, R. J. Prospective prediction of health-related behaviours with the Theory of Planned Behaviour: a meta-analysis. *Health Psychol. Rev.* **5**, 97–144 (2011).

20. Strobl, C., Malley, K. & Tutz, G. An introduction to recursive Partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forestst. *Psychol. Methods* **14**, 323–348 (2009).

21. COVID-19 guidance for older adults. *Centers for Disease Control and Prevention* https://www.cdc.gov/aging/covid19-guidance.html (2020).

22. Wenham, C., Smith, J., Morgan, R. & Group, W. COVID-19: the gendered impacts of the outbreak. *Lancet* **395**, 846–848 (2020).

23. Russell, J. A. A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178 (1980).

24. Eastwood, J. D., Fahlman, S. A., Mercer-Lynn, K. B. & Flora, D. B. Development and validation of the Multidimensional State Boredom Scale. *Assessment* **20**, 68–85 (2013).

25. Carver, C. S., Scheier, M. F. & Weintraub, J. K. Assessing coping strategies: a theoretically based approach. *J. Pers. Soc. Psychol.* **56**, 267–283 (1989).

26. Sexton, K. A. & Dugas, M. J. The cognitive avoidance questionnaire: validation of the English translation. *J. Anxiety Disord.* **22**, 355–370 (2008).

27. Selenko, E. & Batinic, B. Beyond debt, a moderator analysis of the relationship between perceived financial strain and mental health. *Soc. Sci. Med.* **73**, 1725–1732 (2011).

28. Hughes, M. E., Waite, L. J., Hawkley, L. C. & Cacioppo, J. T. A short scale for measuring loneliness in large surveys: Results from two population-based studies. *Res. Aging* **26**, 655–672 (2004).

29. Hahn, E., Gottschling, J. & Spinath, F. M. Short measurements of personality – Validity and reliability of the GSOEP Big Five Inventory (BFI-S). *J. Res. Pers.* **46**, 355–359 (2012).

30. Abdel-Khalek, A. Measuring happiness with a single-item scale. *Soc. Behav. Pers.* **34**, 139–150 (2006).

31. Hershfield, H. E., Mogilner, C. & Barnea, U. People who choose time over money are happier. *Soc. Psychol. Personal. Sci.* **7**, 697–706 (2016).

32. Seligman, M. *Flourish: A new understanding of happiness, well-being, and how to achieve them.* (Nicholas Brealey Publishing, 2011).

33. Macan, T. H. Time management: Test of a process model. *J. Appl. Psychol.* **79**, 381–391 (1994).

34. Shipp, A. J., Edwards, J. R. & Lambert, L. S. Conceptualization and measurement of temporal focus: The subjective experience of the past, present, and future. *Organ. Behav. Hum. Decis. Process.* **110**, 1–22 (2009).

35. Konovsky, M. A. & Cropanzano, R. Perceived fairness of employee drug testing as a predictor of employee attitudes and job performance. *J. Appl. Psychol.* **76**, 698–707 (1991).

36. Porath, C., Spreitzer, G., Gibson, C. & Garnett, F. G. Thriving at work: Toward its measurement, construct validation, and theoretical refinement. *J. Organ.*

Behav. **33**, 250–275 (2012).

37. Van der Elst, T., De Witte, H. & De Cuyper, N. The Job Insecurity Scale: A psychometric evaluation across five European countries. *Eur. J. Work Organ. Psychol.* **23**, 364–380 (2014).

38. Bruder, M., Haffke, P., Neave, N., Nouripanah, N. & Imhoff, R. Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy Mentality Questionnaire. *Front. Psychol.* **4**, 225 (2013).

39. Schlier, B., Moritz, S. & Lincoln, T. M. Measuring fluctuations in paranoia: Validity and psychometric properties of brief state versions of the Paranoia Checklist. *Psychiatry Res.* **241**, 323–332 (2016).

40. User's Guide and Codebook for the ANES 2016 Time Series Study. *Election Studies* https://electionstudies.org/wp-content/uploads/2018/12/anes_timeseries_2016_userguidecodebook.pdf (2019).

41. ESS Round 7 Source Questionnaire. *American National Election Studies* https://www.europeansocialsurvey.org/docs/round7/fieldwork/source/ESS7_source_main_questionnaire.pdf (2014).

42. Zavala-Rojas, D. Thermometer Scale (Feeling Thermometer). in *Encyclopedia of Quality of Life and Well-Being Research* (ed. Michalos, A. C.) 6633–6634 (Springer, 2014). doi:10.1007/978-94-007-0753-5_1028.

43. Gelfand, M. *Rule makers, rule breakers: Tight and loose cultures and the secret signals that direct our lives*. (Scribner, 2019).

44. Gootjes, F., Kuppens, T., Postmes, T. & Gordijn, E. Disentangling Societal Discontent and Intergroup Threat: Explaining Actions towards Refugees and towards the State. *Int. Rev. Soc. Psychol.* **34**, 1–14 (2021, in press).

45. Leander, N. P., Chartrand, T. L. & Wood, W. Mind your mannerisms: Behavioral mimicry elicits stereotype conformity. *J. Exp. Soc. Psychol.* **47**, 195–201 (2010).

46. Stroebe, W. *Social psychology and health*. (Open University Press, 2011).

47. Stroebe, W., Leander, N. P. & Kruglanski, A. W. Is it a dangerous world out there? The motivational bases of american gun ownership. *Personal. Soc. Psychol. Bull.* **43**, 1071–1085 (2017).

48. Van Zomeren, M., Postmes, T. & Spears, R. Toward an integrative social identity model of collective action: a quantitative research synthesis of three socio-psychological perspectives. *Psychol. Bull.* **134**, 504–535 (2008).

49. Stroebe, W., Kreienkamp, J., Leander, N. P. & Agostini, M. Do Canadian and U.S. American handgun owners differ? *Can. J. Behav. Sci.* (2020, in press) doi:10.1037/cbs0000243.

50. Advice on the use of masks in the context of COVID-19: Interim guidance. *World Health Organization* https://www.who.int/publications/i/item/advice-on-the-use-of-masks-in-the- community-during-home-care-and-in-healthcare-settings-in-the-context-of-the-novel-coronavirus-(2019-ncov)-outbreak (2020).

51. Horn, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**, 179–185 (1965).

52. Van Lissa, C. J. TidySEM: Generate tidy SEM-syntax (0.1.0.5). (2020).

53. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. (2015).

54. Probst, P., Wright, M. & Boulesteix, A.-L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* (2019) doi:10.1002/widm.1301.

55. Friedman, J. H. (2001), “Greedy Function Approximation: A Gradient Boosting Machine,” Annals of Statistics, 29, 1189–1232. [273,274]

56. Covid lockdown rules more divisive than Brexit, survey finds. *Guardian* https://www.theguardian.com/world/2020/sep/11/covid-lockdown-rules-more-divisive-than-brexit-survey-finds (2020).

57. Jin, S. *et al.* Intergenerational conflicts of interest and prosocial behavior during the COVID-19 pandemic. *Pers. Individ. Dif.* **171**, 1–8 (2021).

58. Han, Q. *et al.* Trust in government regarding COVID-19 and its associations with preventive health behaviour and prosocial behaviour during the pandemic: a cross-sectional and longitudinal study. *Psychol. Med.* 1–11 (2021, in press) doi:10.1017/S0033291721001306.

59. Nisa, C. F. *et al.* Lives versus livelihoods? Perceived economic risk has a stronger association with support for COVID-19 preventive measures than perceived health risk. *Sci. Rep.* **11**, 1–12 (2021).

60. Romano, A. *et al.* Cooperation and trust across societies during the COVID-19 pandemic. *J. Cross. Cult. Psychol.* 1–21 (2021, in press) doi:10.1177/0022022120988913.

61. Han, Q. *et al.* Associations of risk perception of COVID-19 with emotion and mental health during the pandemic. *J. Affect. Disord.* **284**, 247–255 (2021, in press).

Figure 1. Machine learning results for self-reported personal infection prevention behavior. Variables ranked in order of relative importance

Figure 2. Partial dependence plots depicting bivariate associations between each variable and infection prevention behaviors

Table S3: Full list of variables included in predictive modeling (in rank order of importance). Note: Additional variable descriptive statistics, references, and sources are available at https://osf.io/kxtjf/.

Table 1. Brief descriptions of the top 30 predictors listed in Figure 1. Full variable descriptions are in the supplemental material

	Variable	Brief description
1	Should social distance	Injunctive norm (Right now, people in my area..."-...should self-isolate and engage in social distancing.”)
2	Covid restrictive measures	Support for severe collective virus containment measures (3 items: mandatory quarantines, mandatory vaccinations, report people suspected to be infected with COVID-19)
3	Covid prosocial	Pro-social willingness to protect vulnerable groups from the coronavirus (4 items)
4	Contact immigrants	Days of in-person (face-to-face) contact with immigrants
5	Home.leave.often	How many days in the last week did you leave your home?
6	Contact people	Days of in-person (face-to-face) contact with other people in general
7	Do social distance	Descriptive norm (Right now, people in my area..."-...do self-isolate and engage in social distancing.”)
8	Econ prosocial	Pro-social willingness to protect vulnerable groups from economic consequences of the coronavirus (3 items)
9	Problem solving	Problem-focused coping style (3 items)
10	Consequence contracting	How personally disturbing would it be if… “You were infected with coronavirus”
11	Covid hopeful	“I have high hopes that the coronavirus situation will soon improve”
12	c_doctors_per10k	Number of doctors per 10,000 residents (Country-level; WHO)
13	Date	Date of survey participation (March 19-May 25).
14	c_confirmed	Number of confirmed coronavirus infections (Country-level; Johns Hopkins CSSE)
15	c_political stability	Political stability and absence of violence/terrorism (Country-level; WGI)
16	Focus_present	Temporal focus on the present moment
17	Focus_future	Temporal focus on the future
18	Online_immigrants	Days of online (virtual) contact with immigrants in the past week
19	c_deaths	Number of confirmed COVID-19 deaths (Country-level; Johns Hopkins CSSE)
20	Contact friends	Days of in-person (face-to-face) contact with friends & relatives in the past week
21	c_recovered	Number of confirmed COVID-19 recoveries (Country-level; Johns Hopkins CSSE)
22	c_ghs	Global health security index: pandemic preparedness and health security (Country-level). Source: Global Health Security Index
23	Conspiracy	Generic conspiracy beliefs (3 items)
24	Societal discontent	Concern about direction of society (3 items)
25	Online friends	Days of online (virtual) contact with friends & relatives in the past week
26	Econ. Restrictive measures	Support for extraordinary governmental intervention in economy (3 items)
27	c_govt. effectiveness	Government effectiveness (Country-level; WGI)

28	Covid knowledge	”How knowledgeable are you about the situation regarding the coronavirus?“
29	Leave for work	"In the past week, did you leave your house for work?” (binary)
30	c_stringency	Government COVID response tracker, measured across 17 policy indicators (Country-level): Source: OxCGRT

Notes: Full details of each measure are provided in table S3, as well as the survey codebook (https://osf.io/qhyue/?view_only=d60116c8090d4ec696bfaa9ea14b9432). Country-level variables are denoted with a c_ at the beginning of each variable name.

Table 2. Summary of country-level databases

Database	Description
1. Johns Hopkins University COVID-19 Data Repository Center for Systems Science and Engineering (CSSE).	Number of confirmed COVID-19 infections, deaths, and recoveries by date per country.
2. Global Health Security (GHS) Index	Country-level ratings of pandemic preparedness and general health security.
3. World Health Organization (WHO) and Organization for Economic Cooperation and Development (OECD)	Country-level health care resources and health infrastructure.
4. World Bank: Worldwide Governance Indicators (WGI)	Per-country data on aggregate ratings of: Voice and accountability, regulatory quality, political stability and absence of violence, rule of law, government effectiveness, and control of corruption.
5. Oxford COVID-19 Government Response Tracker (OxCGRT)	Governmental responses and policies with respect to COVID-19 by date per country.

1. Available at https://github.com/CSSEGISandData/COVID-19 [52].
2. Available at https://www.ghsindex.org/.
3. Available at https://apps.who.int/gho/data/node.main.HWF and https://stats.oecd.org/index.aspx?queryid=30183.
4. Available at http://info.worldbank.org/governance/wgi/
5. Available at https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker.

Figure 1. Machine learning results for self-reported personal infection prevention behavior.

Variables ranked in order of relative importance

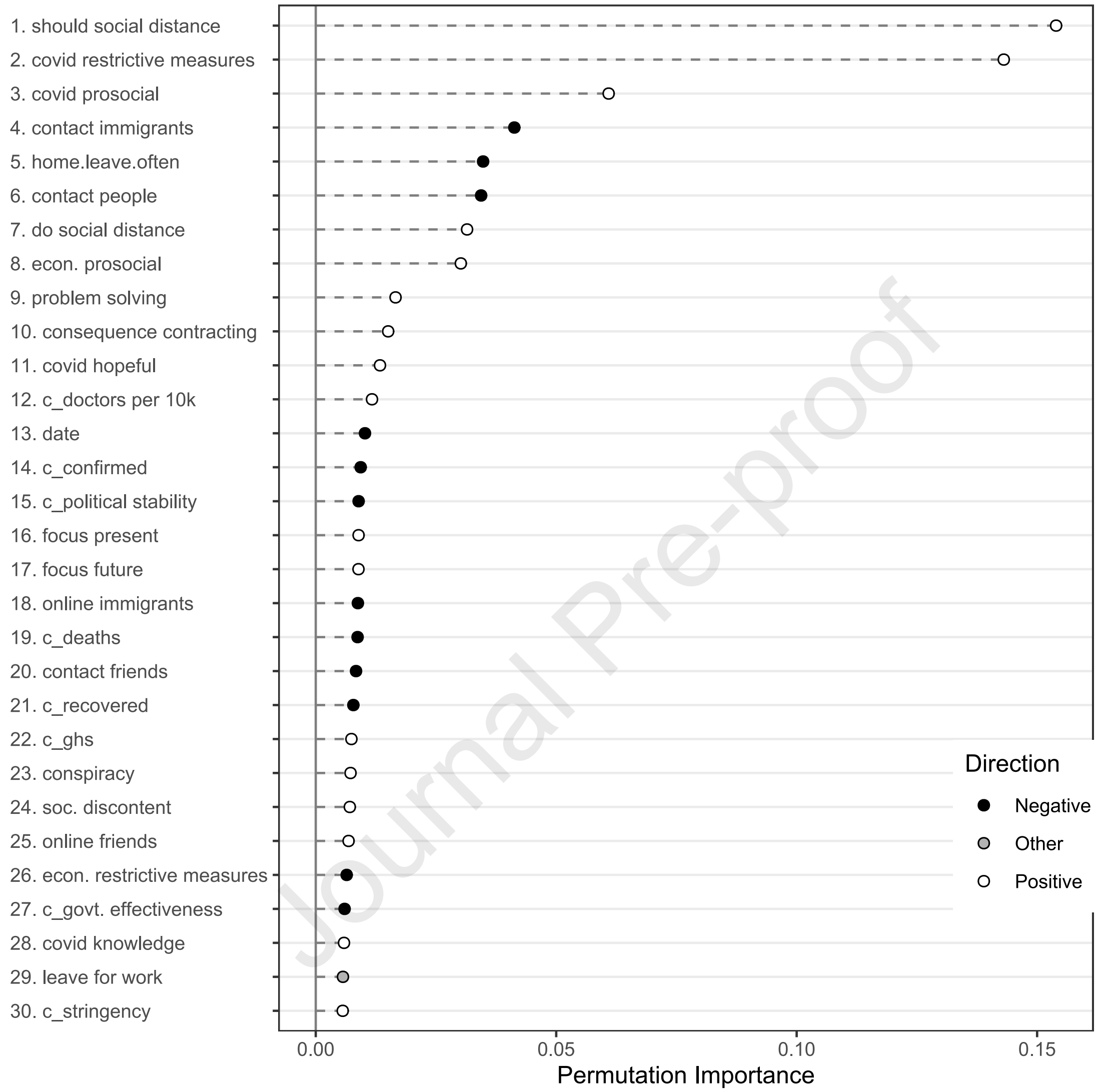
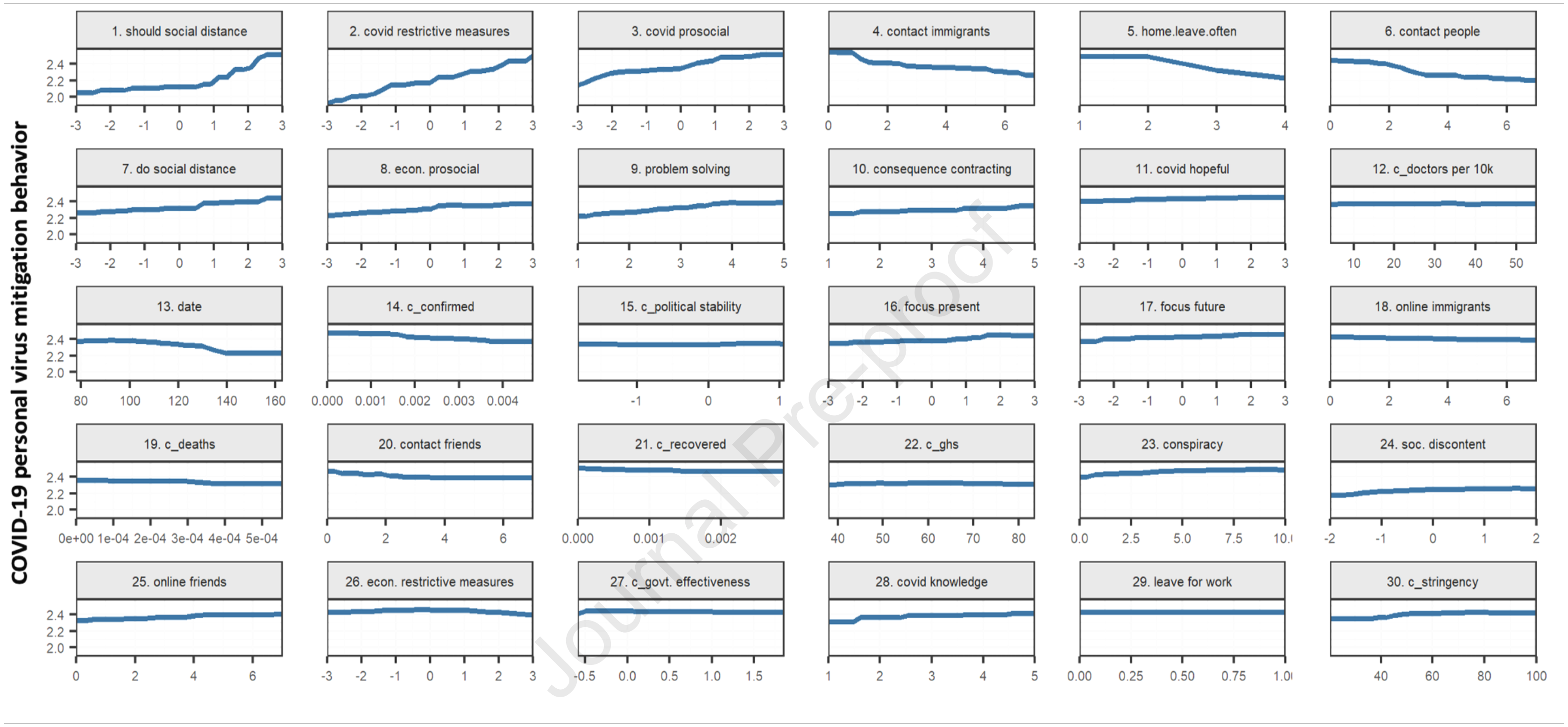


Figure 2. Partial dependence plots depicting bivariate associations between each variable and infection prevention behaviors



Variable	Rank of Importance	Permutation Importance	M (SD)	Item/Construct Description
Should social distance	1	0.154	2.024 (1.288)	Injunctive norm (Right now, people in my area..."-...should self-isolate and engage in social distancing)
COVID restrictive measures	2	0.143	1.480 (1.340)	Support for restrictive collective virus containment measures (mandatory quarantines, mandatory vaccinations, report people suspected to be infected with COVID-19)
COVID prosocial	3	0.061	0.840 (1.230)	Willingness to behave pro-socially to protect vulnerable groups from COVID 19
Contact immigrants	4	0.041	0.518 (1.470)	In the past week, how often respondent had in-person (face-to-face) contact with immigrants
home.leave.often	5	0.035	2.367 (1.044)	How many days in the last week that respondent left their home
Contact people	6	0.034	1.901 (2.181)	In the past week, how often respondent had in-person (face-to-face) contact with other people in general
Do social distance	7	0.031	1.307 (1.509)	Descriptive norm (Right now, people in my area..."-...do self-isolate and engage in social distancing."
Econ. prosocial	8	0.03	0.630 (1.330)	Willingness to behave prosocially to protect others from economic consequences of the coronavirus
Problem solving	9	0.017	3.700 (0.850)	Problem focused coping
Consequence contracting	10	0.015	3.946 (1.227)	How disturbing it would be for respondent to be infected with coronavirus
COVID hopeful	11	0.013	1.216 (1.517)	Positive outlook: I have high hopes that the coronavirus situation will improve
C_doctors per 10k	12	0.012	28.537 (11.625)	Number of doctors per 10,000 residents.
Date	13	0.01		Data of survey participation (March 19-May 25).
C_confirmed	14	0.009	0.0003 (0.0007)	Confirmed coronavirus infections scaled to proportion of population (country-level). Source: Johns Hopkins database.
C_political stability	15	0.009	0.085 (0.644)	Political stability (country-level). Source: World Governance Indicators. From Source: *“Political Stability and Absence of Violence/Terrorism measures perceptions of the likelihood of political instability and/or politically motivated violence, including terrorism.”*
Focus present	16	0.009	1.312 (1.209)	Temporal focus on the present moment
Focus future	17	0.009	1.408 (1.295)	Temporal focus on the future
Online immigrants	18	0.009	0.768 (1.812)	In the past week, how often respondent had online (virtual) contact with immigrants
C_deaths	19	0.009	0.000 (0.000)	Confirmed COVID-19 deaths scaled to proportion of population (country-level). Source: Johns Hopkins
Contact friends	20	0.008	1.974 (2.382)	In the past week, how often respondent had in-person (face-to-face) contact with friends & relatives
C_recovered	21	0.008	0.0001 (0.0003)	Confirmed COVID-19 recoveries scaled to proportion of population (country-level). Source: Johns Hopkins.
C_GHS	22	0.007	63.531 (14.028)	Global health security index: pandemic preparedness and health security (Country-level). Source: Global Health Security Index
Conspiracy	23	0.007	6.560 (2.100)	Endorsement of generic conspiracy beliefs about politicians and government
Soc. Discontent	24	0.007	0.640 (0.770)	Personal worry and discontent with society
Online friends	25	0.007	4.390 (2.478)	In the past week, how often respondent had online (virtual) contact with friends & relatives
Econ. restrictive measures	26	0.006	-0.100 (1.410)	Attitude toward restrictive governmental intervention in economy (e.g., increased government spending, authority, and taxation)
C_govt. effectiveness	27	0.006	0.800 (0.766)	Government effectiveness (country-level). Source: World Governance Indicators. From source: "Government effectiveness captures perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies."
COVID knowledge	28	0.006	3.725 (0.846)	Perception that one is knowledgeable about the situation regarding the coronavirus
Leave for work	29	0.006	0.205 (0.404)	In the past week, how often did respondent leave your home - to go to work (coded 0 if they did not leave the house that week)
C_stringency	30	0.006	78.670 (12.906)	Lockdown severity (country-level) from Oxford Policy Response Tracker: Original Stringency Index. From source: *"records the strictness of “lockdown” policies that primarily restrict people’s behaviour"*
C_nurses per 10k	31	0.005	83.660 (43.070)	Nurses per 10,000 residents (country-level).
Loose norms	32	0.005	6.089 (2.443)	Preference for loose (vs. tight) norms
C_regulatory quality	33	0.005	0.800 (0.802)	Regulatory quality (country-level). Source: World Governance Indicators. From source: *"Regulatory quality captures perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development."*
C_air travel	34	0.005	0.014 (0.011)	Country-level air departures, divided by population size (Country-level index)
Country	35	0.005		[country categorical coding]
C_Govt. response	36	0.005	71.700 (10.859)	Oxford Policy Response Tracker. Overall government response index. From source: *"an overall government response index (which records how the response of governments has varied over all indicators in the database, becoming stronger or weaker over the course of the outbreak"*
Clear messages	37	0.005	4.243 (1.446)	Belief that one is getting clear, unambiguous messages about what to do about the coronavirus
COVID efficacy	38	0.005	0.852 (1.609)	Belief that one's country of residence is able to effectively fight the coronavirus

C_accountability	39	0.005	0.615 (0.796)	Voice and accountability (country-level). World Governance indicators. From source: *"Voice and accountability captures perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media."*
C_control corruption	40	0.005	0.592 (0.895)	Control over corruption (country-level). From source: *"Control of corruption captures perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests."*
Paranoia	41	0.005	3.930 (2.190)	State assessment of suspiciousness of other people
C_Containment health index	42	0.005	74.946 (11.689)	Oxford Policy Response Tracker: Containment health index. From source: *"a containment and health index (which combines “lockdown” restrictions and closures with measures such as testing policy and contact tracing, short term investment in healthcare, as well investments in vaccine"*
Neuroticism	43	0.004	0.080 (1.300)	Personality trait of neuroticism (brief indicator)
C_rule of law	44	0.004	0.678 (0.853)	Rule of law (country-level). World Governance indicators. from source: *"Rule of law captures perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence.”*
Leave for social leisure	45	0.004	0.056 (0.230)	In the past week, how often respondent left the home... for leisure purposes with others (e.g., meeting up with friends, seeing family, going to the cinema, etc). (coded 0 if they did not leave the house that week)
C_tourism expenditures	46	0.004	6.434 (2.644)	Country-level index of International tourism expenditures as percentage of total imports
Life satisfaction	47	0.004	4.123 (1.236)	Sense of personal life satisfaction
Sense of purpose	48	0.004	0.830 (1.572)	Belief that one's life has a clear sense of purpose.
C_health expenditures	49	0.004	9.540 (4.393)	Country-level index of Current health expenditure (CHE) as percentage of gross domestic product (GDP)
Econ. hope	50	0.003	0.572 (1.744)	Extent to which respondent has high hopes that the coronavirus situation will soon improve.
Refocus attention	51	0.003	3.100 (0.910)	coping style - refocus attention (cognitive avoidance)
Migrant threat	52	0.003	5.430 (2.320)	Perceived symbolic & realistic threats from migrants
Disempowerment	53	0.003	-0.010 (0.860)	Perceived disempowerment in society
Happiness	54	0.003	6.337 (2.025)	Personal sense of happiness
C_Population size	55	0.003	119539528.990 (118058899.796)	Country-level index of population size
Relat. Satis.	56	0.003	6.997 (2.193)	Satisfaction with one's personal relationships
Financial strain	57	0.003	0.120 (1.050)	Perceived financial strain
Strict measures	58	0.003	4.110 (1.393)	Perceptions that community is developing strict rules in response to the coronavirus
Focus past	59	0.003	0.673 (1.662)	Temporal focus on the past
Econ. efficacy	60	0.003	0.242 (1.755)	Belief that one's country is able to effectively handle the economic and financial consequences of coronavirus.
Rigid norms	61	0.003	5.556 (2.512)	Preference for flexible vs. rigid societal norms
C_Economic support index	62	0.003	53.852 (27.853)	Oxford Policy Response Tracker: Economic support index. From source: *"an economic support index (which records measures such as income support and debt relief)"*
Consequence economic	63	0.003	3.881 (1.133)	How disturbing it would be for respondent to suffer economic consequences due to the coronavirus situation
Feeling migrants	64	0.002	59.295 (24.182)	Feeling thermometer towards migrants
C_Close transport	65	0.002	0.567 (0.495)	Closures of public transportation on date of survey (country-level). Source: OxCGRT
Gender	66	0.002	1.400 (0.500)	Self-reported gender
Job insecurity	67	0.002	-0.420 (1.080)	Perceived risk of losing one's current job
Organized measures	68	0.002	3.864 (1.397)	Perceptions that one's community is well organized in responding to the coronavirus
Conform norms	69	0.002	5.662 (2.561)	Preference to treat norm violators kindly or harshly
anxious	70	0.002	2.743 (1.249)	Anxious affect in the past week
Infection risk	71	0.002	3.569 (1.441)	Perceived likelihood of personally becoming infected with coronavirus
Punitive measures	72	0.002	3.476 (1.593)	Perceptions that community punishes people who deviate from the rules that have been put in place in response to the coronavirus
Loneliness	73	0.002	2.390 (1.020)	Feelings of loneliness, isolation, and being left out
COVID close	74	0.002	5.428 (1.120)	Social network exposure to virus (e.g., whether oneself, a family member, friend, or community member is known to be infected).
Online people	75	0.002	2.852 (2.643)	In the past week, how many days respondent had online (video or voice) contact with other people in general.
Economic risk	76	0.002	4.439 (1.794)	Perceived likelihood of personally experiencing economic consequences due to the virus
Consequence cancel	77	0.002	3.231 (1.354)	How disturbing it would be for respondent to cancel plans due to the COVID-19
C_restrict gatherings	78	0.002	0.641 (0.480)	Governmental restriction of private gatherings on date of survey (country-level). Source: OxCGRT
Age	79	0.001	2.996 (1.607)	Age (ranges based on U.S. census categories)
nervous	80	0.001	2.598 (1.217)	Nervous affect in the past week
Leave for errands	81	0.001	0.418 (0.493)	In the past week, how often did respondent leave the home... to run errands (coded 0 if did not leave the house in the last week)

Econ. knowledge	82	0.001	3.259 (0.982)	Perception that one is knowledgeable about the economic consequences of the coronavirus
Consequence routines	83	0.001	3.254 (1.266)	How disturbing it would be for respondent to change their life routines due to COVID-19
C_Contact tracing	84	0.001	1.165 (0.624)	Government policy on contact tracing after a positive diagnosis (country-level). Source: OxCGRT
Employ. Status	85	0.001	2.886 (1.665)	Current work situation (unemployed, student, partially or full-time employed, etc.)
Close friend	86	0.001	0.767 (0.642)	Does the respondent have someone with whom they can discuss very personal matters
C_Intl. travel restrictions	87	0.001	3.453 (0.817)	Restrictions on international travel on date of survey (country-level). Source: OxCGRT
excited	88	0.001	2.128 (1.106)	Excited affect in the past week
exhausted	89	0.001	2.480 (1.222)	Exhausted affect in the past week
energetic	90	0.001	2.542 (1.096)	Energetic affect in the past week
C_testing	91	0.001	1.743 (0.916)	Government policy on access to testing (country-level). Source: OxCGRT
calm	92	0.001	2.903 (1.097)	Calm affect in the past week
C_stay home	93	0.001	0.601 (0.490)	Stay-at-home recommendations on date of survey (country-level). Source: OxCGRT
relaxed	94	0.001	2.725 (1.116)	Relaxed affect in the past week
depressed	95	0.001	2.247 (1.200)	Depressed affect in the past week
Education	96	0.001	4.341 (1.434)	Education level
C_Debt relief	97	0.001	1.140 (0.664)	Governmental freezing of financial obligations (country-level). Source: OxCGRT
bored	98	0.001	2.737 (1.324)	Bored affect in the past week
content	99	0.001	2.643 (1.101)	Content affect in the past week
inspired	100	0.001	2.411 (1.142)	Inspired affect in the past week
Leave for leisure	101	0.001	0.186 (0.389)	In the past week, how often did respondent leave the home - for leisure purposes alone (e.g., running, going for a walk, etc.) (coded 0 if they did not leave the house that week)
C_restrict mobility	102	0.001	0.555 (0.497)	Governmental regulations of internal travel on date of survey (country-level). Source: OxCGRT
Leave for other	103	0.001	0.224 (0.417)	In the past week, how often did respondent leave the home - for other reasons (coded 0 if did not leave the house in the last week)
National identity	104	0.001	2.976 (1.471)	Self-other overlap applied between oneself and current country of residence
C_Invest healthcare	105	0	3741033575.756 (29604619032.280)	Short-term spending on healthcare system (country-level). Source: OxCGRT
C_close workplace	106	0	0.689 (0.463)	Workplace closings in country of residence on date of survey (country-level) Source: OxCGRT
Religious	107	0	0.493 (0.500)	Is respondent religious
C_school close	108	0	0.694 (0.461)	School closings in country of residence on date of survey (country-level) Source: OxCGRT
Natural born	109	0	0.900 (0.300)	Natural born citizen of country of residence
Is immigrant	110	0	0.083 (0.275)	Immigrant status in country of residence
C_Cancel events	111	0	0.701 (0.458)	Recommendations regarding public events on date of survey (country-level); Source: OxCGRT
C_Fiscal measures	112	0	30256900268.366 (238457290900.869)	Announced economic stimulus spending (country-level). Source: OxCGRT
Is citizen	113	0	0.949 (0.219)	Citizen of country of residence
C_Invest vaccines	114	0	391086.486 (11487519.495)	Public spending on COVID-19 vaccine development (country-level). Source: OxCGRT
C_International support	115	0	16711202.552 (3575531900.277)	Offers of COVID-19 related aid to other countries (country-level). Source: OxCGRT

Note. Some tables are too large to fit into a document and are thus linked separately. For the machine learning analysis, all analysis code and results are provided online at: DOI: 10.5281/zenodo.5948816. The PsyCorona survey details, including translation procedures and codebook in 30 languages, is available on the Open Science Framework at https://osf.io/qhyue/

Translation Procedures

Table S1. *PsyCorona Scale Translation Procedure*

Language	Translators	Backward Translation	Translation & Revision Team	Translation software	Other Translation Methods
Albanian	2	Yes	Yes	Yes	Backward translation used for some items only to check meaning.
Arabic	2	No	Yes	Yes	
Bengali	1	No	No	Yes	
Croatian	2	No	Yes		
Dutch	2	No	Yes	Yes	
English	*N/A*	*N/A*	*N/A*		*Survey and scales were developed in English*
Farsi	3	No	Yes	Yes	Backward translation used for some items only to check meaning, Initial survey with 3 translators, follow-up survey by 1 translator
French	2	No	Yes	Yes	
German	3	No	Yes	Yes	Searched for scales in papers and databases
Greek	2	No	Yes		
Hindi	1	No	No	Yes	
Hungarian	4	No	Yes		
Indonesian	2	No	Yes		Assisted using online dictionaries
Italian	2	No	Yes	Yes	
Japanese	2	No	Yes		
Korean	2	No	Yes		
Malay	2	No	Yes		Assisted using online dictionaries
Polish	4	Yes	Yes	Yes	
Portuguese	1	No	No	Yes	
Romanian	2	Yes	Yes		

Russian	2	No	Yes		
Serbian	2	No	Yes		
Simplified Chinese	2	No	Yes		
Spanish	4	Yes	Yes		
Thai	2	Yes	Yes		
Traditional Chinese	2	No	Yes		
Turkish	3	No	Yes		
Ukrainian	2	No	Yes	Yes	Backward translation used for some items only to check meaning.
Urdu	2	Yes			
Vietnamese	2	Yes	Yes	Yes	Assisted using online dictionaries

Note: Translators = Number of translators who worked on this scale. Backward Translation = One person translated the measure from English to the language, and a different person translated the scale from the language back to English to check for scale meaning. Translation & Revision Team = One person translated the scale from English to the language, and a second person revised this translation. Alternatively, each person translated the scale and worked together during the revision. Translation Software = Translators used a translation software in the process (e.g., Google Translate). Other Translation Methods = other methods used in the translation of the survey.

Table S2. *Potentially relevant translation issues*

Language	Translation Issues
Albanian	• We were careful to choose semantically correct translations over more literal ones aiming to accommodate cultural differences. In some cases it was needed to add more words for correct understanding. • "Online vs. offline contact" was translated as "online vs. direct contact" • The item about "belief in one God/more than one God" was translated as "belief in one God" as all the religions in Kosovo and Albania are monotheistic. While these kinds of beliefs were not separated for ex. in two items but were within one item, it may have been confusing for the subjects so the translation was adapted culturally. • Items pertaining to political orientation (left/right wing) may not be relatable due to the terminology used. Longer descriptions may have been needed to explain the terms and ensure they are correctly understood.
Arabic	• Some words/phrases were changed or removed to accommodate regional religious beliefs. • There wasn't a word for 'local community' in Arabic so used the term 'society' instead. • In the question where there is a distinction between should and do isolate/social distance myself, and want/have to … a formatting error caused the wrong term to be bolded in some items, but the wording was the same.
Bengali	*None*
Croatian	• Difficulty translating formidability items as the word formidability does not translate well. We adjusted the translation for better understanding. • QID536 - "The events in my life are mainly determined by own actions" - we translated this as "The events in my life are mainly under my control" as this is more semantically correct.
Dutch	• Formidability was translated as 'powerful' as the Dutch word for formidability is rarely used. • "Online vs. offline contact" was translated as "online vs. face-to-face contact"
English	*N/A*
Farsi	• Multiple questions did not translate well. • Attitude about politics is a relatively western way to categorize people into groups.
French	*None*
German	• Some items were hard to translate. E.g. 'community' does not translate well. • Semantically correct translations were sometimes chosen over more literal ones to accommodate cultural differences.
Greek	• "Online" in the item "In the past 7 days, how many days did you have **online** (video or voice) contact with …" was translated "Internet". • "Community" (in the present context) does not translate well into Greek.
Hindi	• Some items were too technical and did not translate well, so simpler translations conveying the meaning were chosen.
Hungarian	• Some items were difficult to translate accurately.
Indonesian	• Some items were difficult to translate accurately due to the inequality of meanings.

Italian	*None*
Japanese	*None*
Korean	*None*
Malay	• Some items were difficult to translate literally due to cultural considerations. E.g., the item about belief in one God/more than one God may be perceived as offensive to Malay Muslim when the item is being written as one item. Agreeing on the item may indicate that the individual believes in either one and this is unacceptable to the majority of Muslims in Malaysia. • Items pertaining to political orientation (left/right wing) may not be relatable to many locally, due to the terminology used. Longer descriptions may be needed to explain the term to ensure the terms could be understood correctly.
Polish	• Tightness-looseness construct was difficult to translate.
Portuguese	*None*
Romanian	*None*
Russian	• Tightness-looseness construct is difficult to express in Russian. • The terms "economic left-right" and "libertarian-authoritarian" make little sense without explanation to most Russians.
Serbian	• Difficulty translating formidability items. • Identification item translated as *I feel close to* instead of *I identify with*.
Simplified Chinese	*None*
Spanish	• Care taken when finding equivalence between standard Spanish and Latin American Spanish.
Thai	• Difficulty in translating cross-cultural research terms. • Some items were difficult to translate literally and accurately. • Questions about bodies were confusing.
Traditional Chinese	• Translated “in my country” to “in the place I live” in order to accommodate both Taiwan and Hong Kong (which is not a country, but a special administrative region).
Turkish	*None*
Ukrainian	• Questions about ‘bodies’ were confusing since the metaphor itself might not have been fully clear for the local population. • The same concerns formidability. Many sentences had to be restructured in order to save the meaning of the question.
Urdu	*None*
Vietnamese	• Some translated items were difficult to express accurately in Vietnamese due to political and social issues (eg. protest/ protesting) and some were not popular to most Vietnamese people (eg. economic left-right or libertarian - authoritarian).

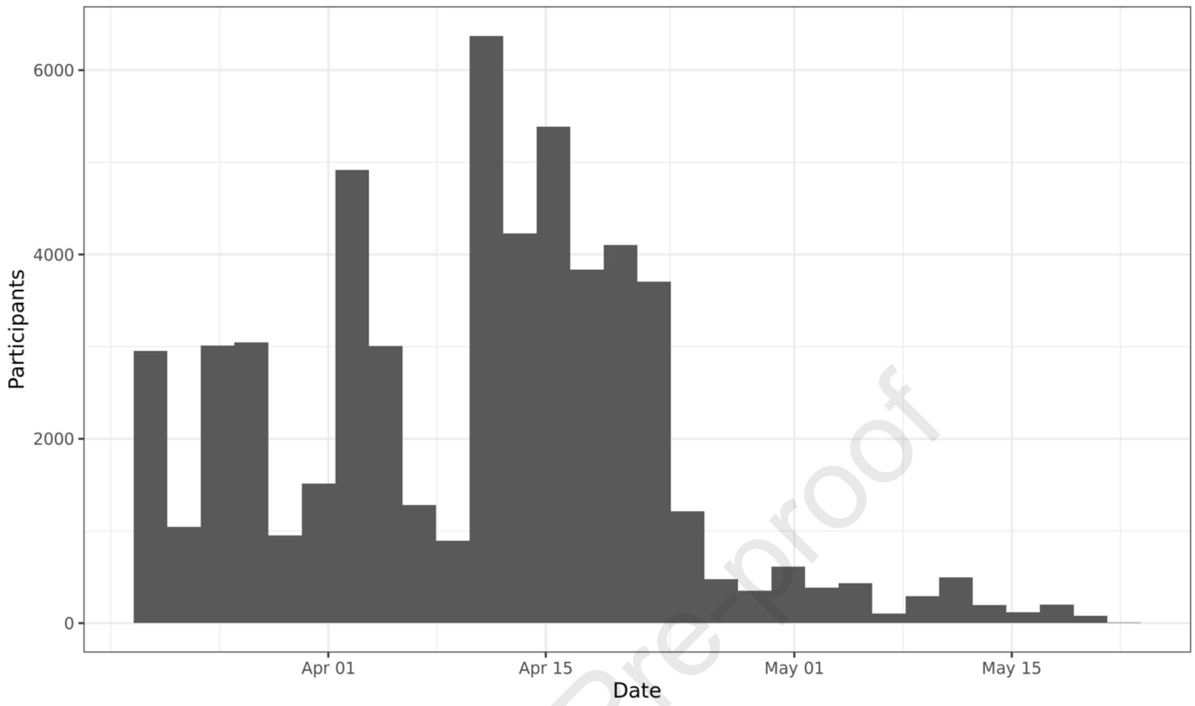
Table S4. Samples in the 28 countries that remained in the data after cleaning.

Country	n	Female	Male	Gender: other	Primary education	Secondary education	Vocational education	Some higher edu	Bachelor's degree	Master's degree	PhD
Ukraine	1433	0.603	0.396	0.001	0.004	0.091	0.134	0.386	0.108	0.22	0.057
Italy	2006	0.602	0.393	0.004	0.006	0.064	0.052	0.503	0.116	0.216	0.043
Greece	2875	0.675	0.323	0.002	0.006	0.017	0.048	0.25	0.379	0.232	0.068
Romania	2704	0.609	0.388	0.003	0.013	0.242	0.032	0.252	0.283	0.154	0.024
Indonesia	2410	0.509	0.486	0.005	0.009	0.351	0.06	0.048	0.37	0.128	0.034
Malaysia	895	0.712	0.286	0.002	0.002	0.056	0.009	0.12	0.534	0.229	0.049
Philippines	1530	0.564	0.425	0.011	0.01	0.077	0.065	0.108	0.555	0.126	0.058
Argentina	1412	0.565	0.431	0.004	0.01	0.233	0.142	0.28	0.241	0.053	0.041
Russia	1438	0.612	0.384	0.003	0.004	0.079	0.195	0.45	0.088	0.133	0.05
USA	11048	0.62	0.373	0.007	0.033	0.094	0.056	0.196	0.389	0.179	0.053
Canada	1538	0.574	0.416	0.01	0.02	0.174	0.109	0.205	0.31	0.141	0.042
Japan	1326	0.474	0.522	0.004	0.002	0.173	0.039	0.334	0.371	0.059	0.02
Egypt	1158	0.841	0.157	0.002	0.007	0.196	0.026	0.477	0.246	0.034	0.012
Netherlands	2409	0.623	0.371	0.007	0.018	0.122	0.183	0.222	0.133	0.234	0.088
Saudi Arabia	1468	0.527	0.463	0.01	0.015	0.192	0.061	0.101	0.493	0.099	0.039
France	1801	0.581	0.414	0.005	0.027	0.145	0.195	0.186	0.111	0.189	0.147
Spain	3203	0.627	0.368	0.006	0.014	0.12	0.158	0.299	0.253	0.106	0.05
Germany	1690	0.565	0.43	0.005	0.011	0.109	0.314	0.179	0.133	0.202	0.053
United Kingd	1935	0.612	0.383	0.005	0.008	0.193	0.132	0.191	0.258	0.158	0.06
South Korea	1452	0.57	0.427	0.003	0.005	0.03	0.015	0.403	0.421	0.097	0.03
Turkey	1826	0.604	0.395	0.002	0.008	0.015	0.208	0.107	0.465	0.153	0.045
Kazakhstan	812	0.562	0.437	0.001	0.001	0.041	0.041	0.302	0.268	0.268	0.079
Australia	1216	0.535	0.46	0.005	0.013	0.22	0.164	0.171	0.296	0.101	0.034
Kosovo	830	0.838	0.162	0	0.004	0.078	0.045	0.299	0.345	0.195	0.034
Brazil	1395	0.577	0.422	0.001	0.02	0.241	0.092	0.339	0.182	0.096	0.029
Poland	718	0.832	0.154	0.014	0.014	0.331	0.059	0.089	0.12	0.331	0.055
Republic of S	2122	0.661	0.337	0.002	0.013	0.17	0.268	0.121	0.248	0.141	0.039
South Africa	1422	0.568	0.429	0.004	0.017	0.189	0.071	0.361	0.283	0.061	0.018

Table S5: Scale descriptive statistics after combining PsyCorona survey items

Subscale	Items	n	mean	sd	min	max	skew	skew_2s	kurt	kurt_2se	Reliabili	Interpret	min_load	max_load
disc	3	55979	0.64	0.77	-2	2	-0.45	-21.51	0.15	3.55	0.68	Questionable	0.41	0.84
jbinsec	4	46018	-0.42	1.08	-2	2	0.38	16.71	-0.57	-12.39	0.81	Good	0.63	0.88
pfs	3	55962	0.12	1.05	-2	2	-0.1	-4.77	-0.66	-15.95	0.85	Good	0.65	0.92
fail	3	55981	-0.01	0.86	-2	2	-0.02	-0.92	-0.18	-4.37	0.66	Questionable	0.48	0.72
lone	3	56005	2.39	1.02	1	5	0.43	20.68	-0.51	-12.33	0.82	Good	0.76	0.82
probsolving	3	55976	3.7	0.85	1	5	-0.48	-23.35	0.15	3.63	0.84	Good	0.77	0.86
posrefocus	3	55973	3.1	0.91	1	5	-0.16	-7.88	-0.17	-4.1	0.85	Good	0.75	0.85
c19proso	4	55979	0.84	1.23	-3	3	-0.55	-26.37	0.16	3.84	0.77	Acceptable	0.56	0.8
c19perbeh	3	55982	2.19	1	-3	3	-1.88	-91.01	4.45	107.54	0.75	Acceptable	0.59	0.95
c19rca	3	55975	1.48	1.34	-3	3	-1.01	-48.75	0.78	18.81	0.71	Acceptable	0.59	0.81
ecoproso	4	55910	0.63	1.33	-3	3	-0.51	-24.82	0.05	1.13	0.86	Good	0.67	0.84
ecorca	3	55900	-0.1	1.41	-3	3	-0.15	-7.13	-0.43	-10.36	0.65	Questionable	0.59	0.68
bordeom	3	55970	1.37	1.29	-3	3	0	0.19	-0.35	-8.53	0.53	Poor	0.07	0.96
migrantthreat	5	55643	5.43	2.32	1	10	-0.18	-8.82	-0.68	-16.25	0.91	Excellent	0.75	0.89
cognitive test	3	55963	1.96	0.38	1	3	-0.78	-37.56	-0.3	-7.2	0.27	Unacceptable	0.22	0.47
neuro	3	55946	0.08	1.3	-3	3	-0.02	-1.21	-0.38	-9.11	0.69	Questionable	0.55	0.86
para	3	55901	3.93	2.19	0	10	0.29	14.02	-0.21	-4.96	0.69	Questionable	0.41	0.91
consp	3	55710	6.56	2.1	0	10	-0.46	-22.25	-0.03	-0.69	0.73	Acceptable	0.47	0.83

Fig. S1. Distribution of participation dates.



Highlights

- We studied predictors of COVID-19 prevention behaviors in a cross-national study.
- The strongest predictors related to injunctive norms.

ETOC

In a study of 56,072 participants from 28 countries, we used a machine learning approach to identify the strongest predictors of COVID-19 infection prevention behavior (pre-vaccine). Few country-level data variables predicted outcomes. Instead, individual psychological variables predicted outcomes. Injunctive norms such as believing people should engage in the behaviors and support for behavioral mandates were the strongest predictors of infection prevention behavior. The results highlight how both data- and theory-driven approaches can increase understanding of complex human behavior.

The Bigger Picture

In the absence of a vaccine or cure, virus containment depended on individual-level compliance with behaviors recommended by the World Health Organization. We used machine learning to identify the most important indicators of compliance, based on a large international psychological survey and country-level secondary data. The most important indicators were not the “usual suspects”, such as personal threat of virus infection, but rather injunctive norms—namely, the belief that one’s community *should* engage in such behavior and that society should take restrictive virus containment measures. People appear who tend to engage in infection prevention behaviors also tend to believe that general compliance is necessary to defeat the pandemic, which extends to endorsement of ‘ought’ norms and support for behavioral mandates. These results highlight the potential to intervene by shaping social norms and expectations.