



# Population-scale longitudinal mapping of COVID-19 symptoms, behaviour and testing

William E. Allen<sup>1,2,3,18</sup>  , Han Altae-Tran<sup>1,3,4,18</sup>, James Briggs<sup>1,3,5,18</sup> , Xin Jin<sup>1,2,3,18</sup> , Glen McGee<sup>1,6,18</sup> , Andy Shi<sup>1,6,18</sup> , Rumya Raghavan<sup>1,3,7</sup> , Mireille Kamariza<sup>1,2,3</sup> , Nicole Nova<sup>1,8</sup> , Albert Pereta<sup>1</sup>, Chris Danford<sup>1</sup>, Amine Kamel<sup>1</sup>, Patrik Gothe<sup>1</sup>, Evrhet Milam<sup>1</sup>, Jean Aurambault<sup>1</sup> , Thorben Primke<sup>1</sup>, Weijie Li<sup>1</sup> , Josh Inkenbrandt<sup>1</sup>, Tuan Huynh<sup>1</sup>, Evan Chen<sup>1</sup>, Christina Lee<sup>1</sup>, Michael Croatto<sup>1</sup>, Helen Bentley<sup>1</sup>, Wendy Lu<sup>1</sup> , Robert Murray<sup>1</sup>, Mark Travassos<sup>1,9</sup> , Brent A. Coull<sup>6</sup>, John Openshaw<sup>1,10</sup> , Casey S. Greene<sup>1,11</sup> , Ophir Shalem<sup>1,12</sup>, Gary King<sup>1,13</sup> , Ryan Probasco<sup>1</sup>, David R. Cheng<sup>1</sup>, Ben Silbermann<sup>1</sup>, Feng Zhang<sup>1,3,4,14,15,16</sup>   and Xihong Lin<sup>1,3,6,17</sup>  

**Despite the widespread implementation of public health measures, coronavirus disease 2019 (COVID-19) continues to spread in the United States. To facilitate an agile response to the pandemic, we developed How We Feel, a web and mobile application that collects longitudinal self-reported survey responses on health, behaviour and demographics. Here, we report results from over 500,000 users in the United States from 2 April 2020 to 12 May 2020. We show that self-reported surveys can be used to build predictive models to identify likely COVID-19-positive individuals. We find evidence among our users for asymptomatic or presymptomatic presentation; show a variety of exposure, occupational and demographic risk factors for COVID-19 beyond symptoms; reveal factors for which users have been SARS-CoV-2 PCR tested; and highlight the temporal dynamics of symptoms and self-isolation behaviour. These results highlight the utility of collecting a diverse set of symptomatic, demographic, exposure and behavioural self-reported data to fight the COVID-19 pandemic.**

The rapid global spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the novel virus causing COVID-19 (refs. <sup>1–3</sup>), has created an unprecedented public health emergency. In the United States, efforts to slow the spread of disease have included, to varying extents, social distancing, home-quarantine and treating of infected patients, mandatory facial covering, closure of schools and non-essential businesses, and test–trace–isolate measures<sup>4,5</sup>. The COVID-19 pandemic and ensuing response has produced a concurrent economic crisis of a scale not seen for nearly a century<sup>6</sup>, exacerbating the effect of the pandemic on different socioeconomic groups and producing adverse health outcomes beyond COVID-19. As a result, there is currently intense pressure to safely wind down these measures. Yet, in spite of widespread lockdowns and social distancing throughout the United States, many states continue to exhibit steady increases in the number of cases (<https://www.worldometers.info/coronavirus/>). To understand where and why the disease continues to spread, there is a pressing need for real-time individual-level data on COVID-19 infections and tests, as well as on the behaviour, exposure and

demographics of individuals at the population scale with granular location information. These data will allow medical professionals, public health officials and policy makers to understand the effects of the pandemic on society, tailor intervention measures, efficiently allocate testing resources and address disparities.

One approach to collecting these types of data on a population scale is to use web- and mobile-phone-based surveys that enable large-scale collection of self-reported data. Previous studies, such as FluNearYou, have demonstrated the potential for using online surveys for disease surveillance<sup>7</sup>. Since the start of the COVID-19 pandemic, several different applications have been launched throughout the world to collect COVID-19 symptoms, testing and contact-tracing information<sup>8</sup>. Studies in the United States and Canada (CovidNearYou, <https://covidnearlyou.org/us/en-US/>; and ref. <sup>9</sup>), the United Kingdom (Covid Symptom Study<sup>10,11</sup>, also in the United States) and Israel (PredictCorona<sup>12</sup>) have reported large cohorts of users drawn from the general population with a goal towards capturing information about COVID-19 along a variety of dimensions, from symptoms to behaviour, and have demonstrated

<sup>1</sup>The How We Feel Project, San Leandro, CA, USA. <sup>2</sup>Society of Fellows, Harvard University, Cambridge, MA, USA. <sup>3</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>4</sup>Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>5</sup>Schmidt Science Fellows, Oxford, UK. <sup>6</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. <sup>7</sup>Health Sciences and Technology Program, Massachusetts Institute of Technology and Harvard Medical School, Cambridge, MA, USA. <sup>8</sup>Department of Biology, Stanford University, Stanford, CA, USA. <sup>9</sup>Center for Vaccine Development and Global Health, University of Maryland School of Medicine, Baltimore, MD, USA. <sup>10</sup>Division of Infectious Diseases and Geographic Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. <sup>11</sup>Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. <sup>12</sup>Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. <sup>13</sup>Albert J. Weatherhead III University Professor, Institute for Quantitative Social Sciences, Harvard University, Cambridge, MA, USA. <sup>14</sup>McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>15</sup>Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>16</sup>Howard Hughes Medical Institute, Chevy Chase, MD, USA. <sup>17</sup>Department of Statistics, Harvard University, Cambridge, MA, USA. <sup>18</sup>These authors contributed equally: William E. Allen, Han Altae-Tran, James Briggs, Xin Jin, Glen McGee, Andy Shi. ✉e-mail: [weallen@fas.harvard.edu](mailto:weallen@fas.harvard.edu); [zhang\\_f@mit.edu](mailto:zhang_f@mit.edu); [xlin@hsph.harvard.edu](mailto:xlin@hsph.harvard.edu)

some ability to detect and predict the spread of disease<sup>10–12</sup>. This field has rapidly evolved since the beginning of the pandemic, with many analyses of these datasets focusing on COVID-19 diagnostics (that is, symptoms, test results, medical background)<sup>9</sup>, care seeking<sup>13</sup>, contact tracing<sup>14</sup>, patient care<sup>15</sup>, effects on healthcare workers<sup>16</sup>, hospital attendance<sup>13</sup>, cancer<sup>17</sup>, primary care<sup>18</sup>, clinical symptoms<sup>19</sup> and triage<sup>20</sup>. Here, we perform a comprehensive analysis of a source of COVID-19-related information spanning diagnostic and behavioural factors sampled from the general population during the beginning of the pandemic in the United States. We investigate exposure, demographic and behavioural factors that affect the chain of transmission; understand the factors for who has been tested; and study the degree of presence of asymptomatic, presymptomatic and mildly symptomatic cases<sup>21</sup>.

To fill the gap and achieve these goals, we developed How We Feel (HWF; <http://www.howwefeel.org>) (Fig. 1a–d), a web and mobile-phone application for collecting de-identified self-reported COVID-19-related data. Rather than targeting patients with suspected COVID-19 or existing study cohorts, HWF aims to collect data from users representing the population at large. By drawing from a large user base across the United States who learn about the study through word of mouth and government partnerships, these results are complementary to other studies such as the Covid Symptom Study and CovidNearYou that also include sizable US populations and are targeted towards the general public. Users are asked to share information on demographics (gender, age, race/ethnicity, household structure, ZIP code), COVID-19 exposure and pre-existing medical conditions. They then self-report daily how they feel (well or not well), any symptoms they may be experiencing, test results, behaviour (for example, use of face coverings) and sentiment (for example, feeling safe to go to work) (Fig. 1c and Extended Data Fig. 1). To protect privacy, users are not identifiable beyond a randomly generated number that links repeated logins on the same device. A key feature of the app is the ability to rapidly release revised versions of the survey as the pandemic evolves. In the first month of operation, we released three iterations of the survey with increasingly expanded sets of questions (Fig. 1b).

We find symptomatic subjects, healthcare workers and essential workers are more likely to be tested. Due to asymptomatic and mildly symptomatic individuals and heterogeneous symptom presentation, our results show that commonly used symptoms may not be sufficient criteria for evaluating COVID-19 infection. Further, we find that exposure both outside and within the household is a major risk factor for users testing positive and build a predictive model to identify likely COVID-positive users. African-American users, Hispanic/Latinx users, and healthcare workers and essential workers are at a higher risk of infection, after accounting for the effects of pre-existing medical conditions. Finally, we find that even at the height of lockdowns throughout the United States, the majority of users were leaving their homes, and a large fraction were not engaging in social distancing or face protection.

## Results

The app was launched on 2 April 2020 in the United States. As of 12 May 2020, the app had 502,731 users in the United States, with 3,661,716 total responses (Fig. 1b and Supplementary Table 1). In total, 74% of users responded on multiple days, with an average of seven responses per user (Extended Data Fig. 2). Each day, ~5% of users who accessed the app reported feeling unwell (Fig. 1b). The user base was distributed across all 50 states and several US territories, with the largest numbers of users in more populous states such as California, Texas, Florida and New York (Fig. 1d). Connecticut had the largest number of users per state, as the result of a partnership with the Connecticut state government (Fig. 1d). Users were required to be 18 years of age or older and were 42 years old on average (mean, 42.0; s.d., 16.3), including 18.4% in the bracket of 60+,

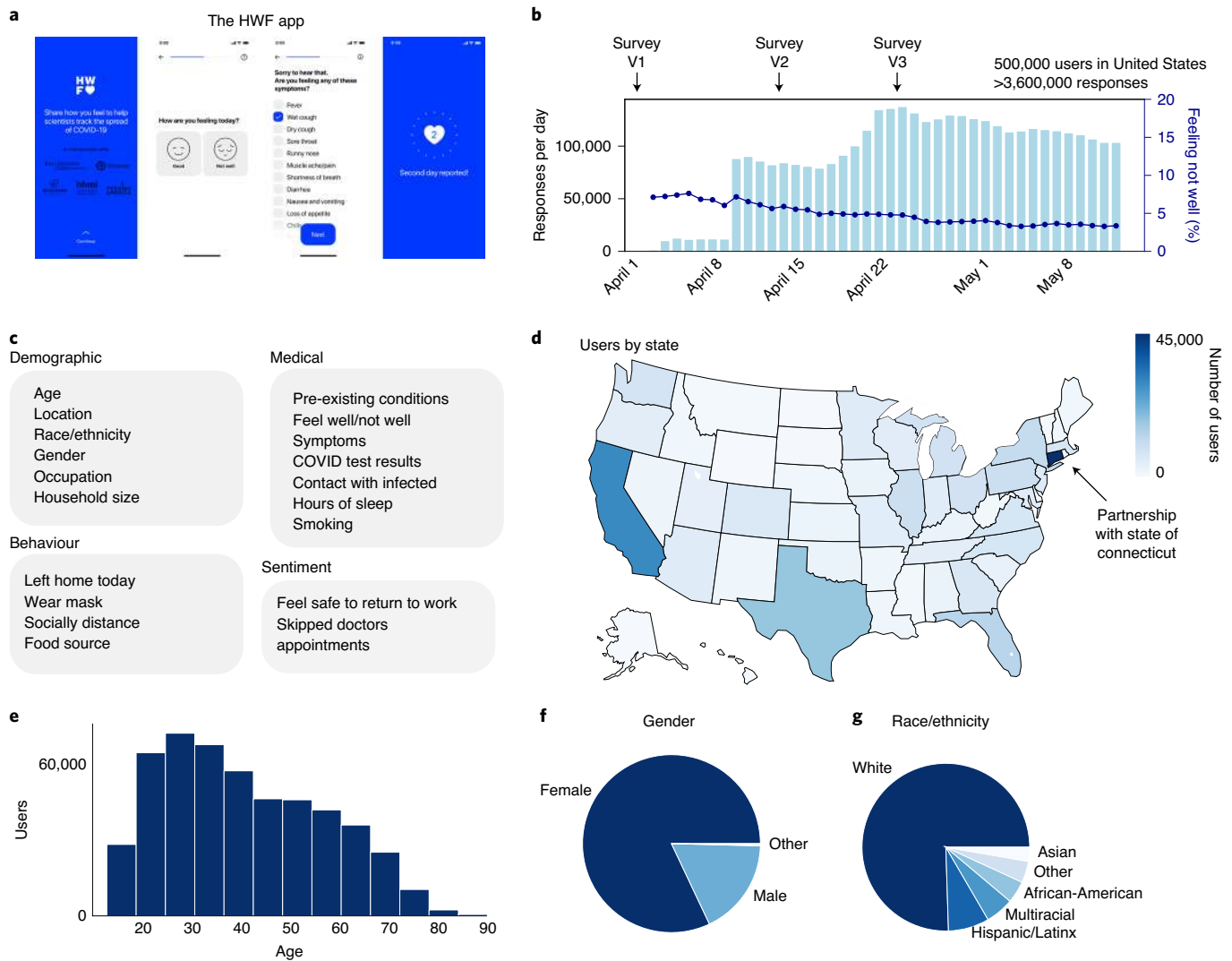
which has experienced the highest mortality rate from COVID-19 (Fig. 1e)<sup>22,23</sup>. Users were primarily female (82.7%) (Fig. 1f) and white (75.5%, excluding 20.3% with missing data) (Fig. 1g). Although the survey ran from April 2 until May 12, users could report test results from earlier than April 2.

A major ongoing problem in the United States is the overall lack of testing across the country<sup>24</sup> and disparities in test accessibility, infection rates and mortality rates in different regions and communities<sup>25</sup>. In the absence of population-scale testing, it will be critical during a reopening to allocate limited testing resources to the groups or individuals most likely to be infected to track the spread of disease and break the chain of infection. We therefore first examined who in our user base was currently receiving testing. We analysed 4,759 users who took the Version 3 (V3) survey and who were PCR tested for SARS-CoV-2 (of 272,392 total users) (Fig. 2a and Extended Data Fig. 3a). Of these, 8.8% were PCR positive. The number of tests reported by test date displays a similar trend to the estimated number of tests across the United States, suggesting that our sampling captures the increase in test availability (Fig. 2a). The number of PCR tests per HWF user is highly correlated with external estimates of per-capita tests by state (Fig. 2b and Extended Data Fig. 3b; Pearson correlation 0.77)<sup>26</sup>.

We first examined via logistic regression which factors either collected in the survey or inferred from US Census data by user ZIP code were associated with receiving a SARS-CoV-2 PCR test, regardless of test result. As expected, we observed a higher fraction of tested users from states with higher per-capita test numbers, according to the COVID Tracking Project<sup>26</sup> (Extended Data Fig. 3b). Healthcare workers (odds ratio (OR), 2.94; 95% confidence interval (95% CI), 2.75, 3.15;  $P < 0.001$ ) and other essential workers (OR, 1.39; 95% CI, 1.28, 1.52;  $P < 0.001$ ) were more likely to have received a PCR test compared with users who did not report those professions (Fig. 2c). Users who reported experiencing fever, cough or loss of taste/smell (among other symptoms) had higher odds of being tested compared with users who never reported these symptoms (Fig. 2c). The majority of these symptoms are listed as common for COVID-19 cases by the Centers for Disease Control and Prevention (CDC) (Fig. 2c, starred)<sup>27</sup>. A less-common symptom, reporting a tight feeling in one's chest, was also associated with receiving a PCR-based test (OR, 2.27; 95% CI, 1.93, 2.66;  $P < 0.001$ ). These results suggest that the most commonly reported symptoms are being used as screening criteria for determining who receives a test, potentially missing asymptomatic and mildly symptomatic individuals. This group could include those who are at high risk for infection but do not meet the testing eligibility criteria.

To obtain a global view of self-reported symptom patterns, we applied an unsupervised manifold learning algorithm to visualize how symptoms were correlated across users (see Methods). As expected, we found that symptom presentation separated broadly by feeling well versus feeling unwell (Fig. 2d and Extended Data Fig. 4). Users who felt unwell were concentrated in a single cluster indicating similar overall symptom profiles, which was characterized by high proportions of common COVID-19 symptoms as defined by the CDC<sup>27</sup> (Fig. 2e), and contained the vast majority of responses from users with both positive (+) and negative (–) SARS-CoV-2 PCR tests (Fig. 2f). Thus, COVID-19 symptoms tend to overlap with symptoms for other diseases and do not necessarily predict positive test results.

This overlap suggests that commonly used symptoms may not be sufficient criteria for evaluating COVID-19 infection. It has previously been reported that many people infected with SARS-CoV-2 are asymptomatic, mildly symptomatic or in the presymptomatic phase of their presentation<sup>28–30</sup> and therefore unaware that they are infected. In our dataset, on the day of their test, most users (73%) that tested PCR positive for SARS-CoV-2 reported feeling unwell with the common symptoms listed by the CDC (dry cough, shortness of breath,



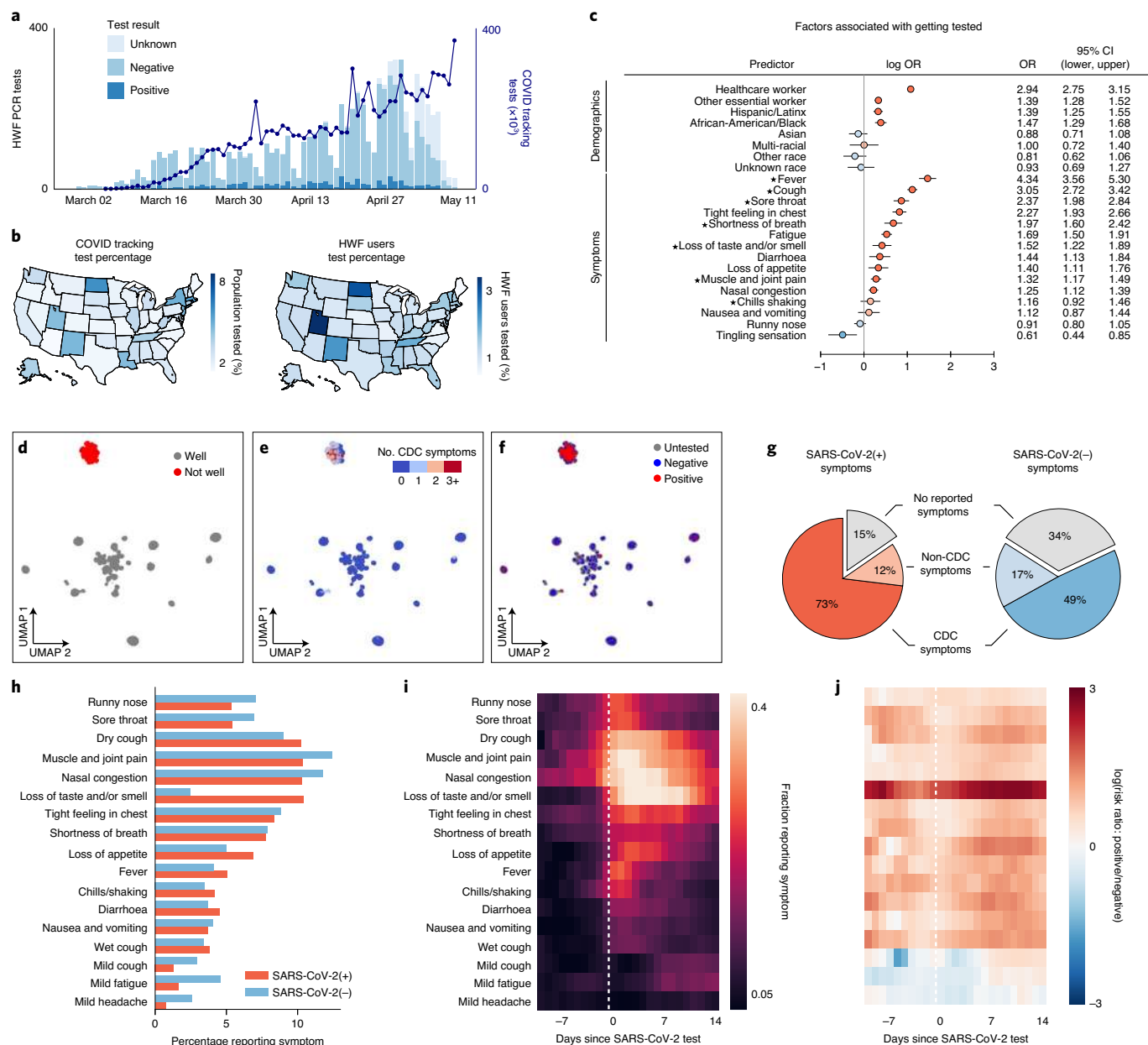
**Fig. 1 | The HWF application and user base.** **a**, The HWF app: longitudinal tracking of self-reported COVID-19-related data. **b**, Responses over time, as well as percentage of users reporting feeling unwell, with releases of major updates to survey indicated. **c**, Information collected by the HWF app. **d**, Users by state across the United States. **e**, Age distribution of users. Note: users had to be older than 18 to use the app. **f**, Distribution of self-reported gender. **g**, Distribution of self-reported race or ethnicity. Users were allowed to report multiple races. ‘Multiracial’ means the user indicated more than one category. ‘Other’ includes American Indian/Alaskan Native and Hawaiian/Pacific Islander, as well as users who selected ‘Other’.

chills/shaking, fever, muscle/joint pain, sore throat, loss of taste/smell). However, 11.5% of positive users reported feeling unwell and exclusively reported symptoms not listed as common for COVID-19 by the CDC on the day of their test, and 15.4% reported feeling no symptoms at all (Fig. 2g). Because of the commonly used symptom- and occupation-based screening criteria for receiving a PCR test and under-testing, this total of 36.9% probably underestimates the true fraction of asymptomatic, presymptomatic and mildly symptomatic cases, which in Wuhan, China, was estimated to be ~87% (ref. 21), and in the United States was estimated to be >80%. A large number of asymptomatic cases were also observed in serological studies<sup>31,32</sup>. In total, 48.9% of users testing negative for SARS-CoV-2 reported feeling unwell with the most common COVID-19 symptoms, compared with an expected false-negative rate of 20–30% for PCR-based tests of symptomatic patients<sup>33</sup>, again suggesting symptom presentation overlap with other diseases (Fig. 2g).

We investigated the symptoms that were most predictive of COVID-19 by exploring the distribution and dynamics of symptoms in PCR test (+) and (–) users around the test date. PCR test

(+) users reported a higher rate of common COVID-19 symptoms, including dry cough, fever, loss of appetite, and loss of taste and/or smell, than PCR test (–) users (Fig. 2h). Many PCR-tested users longitudinally reported symptoms in the app in an interval extending  $\pm 2$  weeks from their test date (Extended Data Fig. 5). We used these data to examine the time course of symptoms among those who tested positive. In the days preceding a test, dry cough, muscle pain and nasal congestion were among the most commonly reported symptoms. Reported symptoms peaked in the week following a test and declined thereafter (Fig. 2i). Taking the ratio of the symptom rates at each point in time between PCR test (+) and (–) users showed that the most distinguishing feature in users who tested positive was loss of taste and/or smell, as has been previously reported<sup>11</sup> (Fig. 2j).

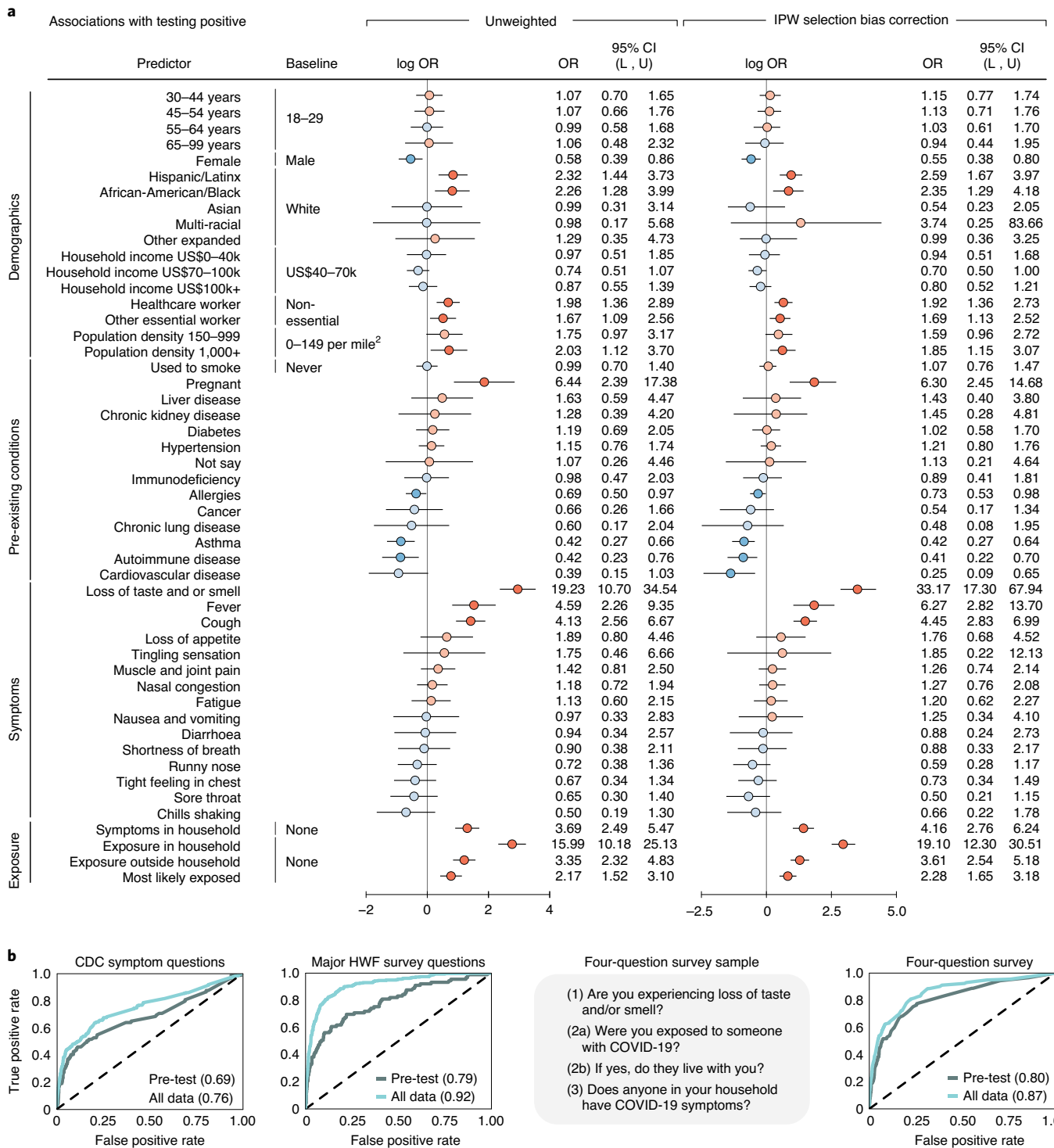
We next investigated medical and demographic factors associated with testing PCR positive for acute SARS-CoV-2 infection, focusing on 3,829 users who took the V3 survey within  $\pm 2$  weeks of their reported PCR test date (315 positive, 3,514 negative) (Fig. 3a and Supplementary Tables 2–6). These users are a subset of all of



**Fig. 2 | SARS-CoV-2 PCR testing and symptoms.** **a**, Stacked bar plot of user-reported test results over time, overlaid with official number of tests across the United States based on COVID Tracking Project data ( $n = 4,759$  users who took the V3 survey and reported a test result, of 277,151 users). **b**, Left: map of per-capita test rates across the United States. Right: map of COVID-19 tests per number of users by state. **c**, Associations of professions and symptoms with receiving a SARS-CoV-2 PCR test, adjusted for demographics and other covariates (see Methods). Common symptoms listed by the CDC are starred ( $n = 4,759$  users with a reported test within 14 d of a survey response of 277,151 users). **d-f**, UMAP visualization of 667,651 multivariate symptom responses among HWF users that reported at least one symptom. Colouring indicates responses according to users feeling well (**d**), the reported number of COVID-19 symptoms listed by the CDC (**e**), and the COVID-19 test result among tested users (**f**). **g**, Proportions of patients positive for COVID-19 (red) and negative for COVID-19 (blue) experiencing CDC common symptoms (dark), only non-CDC symptoms (light) or no symptoms (grey) on the day of their test.  $n = 1,170$  positive users and 8,892 negative users who reported a test result between 2 April and 12 May 2020. **h**, Histogram of reported symptoms among COVID-19-tested users. **i**, Longitudinal self-reported symptoms from users that tested positive for COVID-19. Dates are centred on the self-reported test date. **j**, Ratio of symptoms comparing users that tested positive versus those that tested negative for COVID-19.

the users who reported taking a test in the V3 survey, as some of the reported test results were outside of this time window. To correct for selection bias of receiving a PCR test when studying the risk factors of a positive test result, we incorporated the probability of receiving PCR tests as inverse probability weights (IPWs) into our logistic model of PCR test result status (+/-) (see Methods)<sup>34</sup>. As with the analysis of who received a test, the reported symptom of loss

of taste and/or smell was most strongly associated with a positive test result (OR, 33.17; 95% CI, 17.3, 67.94;  $P < 0.001$ ). Other symptoms associated with testing positive included fever (OR, 6.27; 95% CI, 2.82, 13.70;  $P < 0.001$ ) and cough (OR, 4.45; 95% CI, 2.83, 6.99;  $P < 0.001$ ). Women were less likely to test positive than men (OR, 0.55; 95% CI, 0.38, 0.80;  $P = 0.002$ ), and both Hispanic/Latinx users (OR, 2.59; 95% CI, 1.67, 3.97;  $P < 0.001$ ) and African-American/



**Fig. 3 | SARS-CoV-2 PCR test result associations and predictions.** **a**, Factors associated with respondents receiving and reporting a positive test result, as determined through logistic regression. Left: results from unweighted model. Right: results from model incorporating selection probabilities via IPWs. Reference categories are indicated where relevant; when not indicated, the reference is not having that specific feature. log ORs and their confidence intervals are plotted, with red indicating positive association and blue indicating negative association. Darker colours indicate confidence intervals that do not cover 0. Population density and neighbourhood household income were approximated using county-level data. L, lower bound; U, upper bound of 95% CIs;  $n = 3,829$  users (315 positive, 3,514 negative) who took the V3 survey within  $\pm 2$  weeks of receiving a test. **b**, Prediction of positive test results using  $\pm 2$  weeks of data from the test date, using fivefold cross-validation, shown as ROC curves. The XGBoost model was trained on different subsets of questions: CDC symptom questions, using just the subset of COVID-19 symptoms listed by the CDC; all survey questions, using the entire survey; four-question survey, using a reduced set of four questions that were found to be highly predictive. Numerical values are AUC;  $n = 3,829$  users.

Black users (OR, 2.35; 95% CI, 1.29, 4.18;  $P=0.004$ ) were more likely to test positive than white users, highlighting potential racial disparities involved with COVID-19 infection risk. The odds of testing positive were also higher for those in high-density neighbourhoods (OR, 1.85; 95% CI, 1.15, 3.07;  $P=0.014$ ). Healthcare workers (OR, 1.92; 95% CI, 1.36, 2.73;  $P<0.001$ ) and other essential workers (OR, 1.69; 95% CI, 1.13, 2.52;  $P=0.01$ ) also had higher odds of testing positive compared with non-essential workers. Pregnant women were substantially more likely to test positive (OR, 6.30; 95% CI, 2.45, 14.68;  $P<0.001$ ). However, we note that this result is based on a small sample of 48 pregnant women included in this analysis (9 test positive, 39 test negative) and is unstable, subject to potentially high selection bias. Performing this analysis with and without correction for selection bias produced similar results (Fig. 3a). As a further sensitivity analysis, we reran the analyses excluding users from the states of California and Connecticut, the state containing most users (Extended Data Fig. 7a), and correcting for broader demographic differences using US Census data (Extended Data Fig. 7b), obtaining similar results to the uncorrected model in both cases. Finally, we performed Firth-corrected logistic regression to check for bias in our testing model related to the large fraction of users testing negative, and obtained similar results to our uncorrected model (Extended Data Fig. 8).

Motivated by previous studies that reported that high cluster transmissions occurred in families in China, Korea and Japan<sup>35–37</sup>, we explored household and community exposures as risk factors for users testing PCR positive. The odds of testing positive were much higher for those who reported within-household exposure to someone with confirmed COVID-19 than for those who reported no exposure at all (see Methods) (OR, 19.10; 95% CI, 12.30, 30.51;  $P<0.001$ ) (Fig. 3a and Supplementary Table 5). This is stronger than comparing the odds of testing positive among those who reported exposure outside their household versus no exposure at all (OR, 3.61; 95% CI, 2.54, 5.18;  $P<0.001$ ). Further, the odds of testing PCR positive are much higher for those exposed in the household versus those exposed outside their household or not exposed at all, after adjusting for similar factors (OR, 10.3; 95% CI, 6.7, 15.8;  $P<0.001$ ) (Supplementary Table 10). These results are consistent with previous findings that indicate a very high relative risk associated with within-household infection<sup>36,38–41</sup>. This is compatible with the findings that other closed areas with high levels of congregation and close proximity, such as churches<sup>42</sup>, food-processing plants<sup>43</sup> and nursing homes<sup>44</sup>, have shown similarly high risks of transmission.

Developing models to predict who is likely to be SARS-CoV-2(+) from self-reported data has been proposed as a means to help overcome testing limitations and identify disease hotspots<sup>11,12</sup>. We used data from the 3,829 users who used the app within  $\pm 2$  weeks of their reported PCR test results to develop a set of prediction models that were able to distinguish positive and negative results with a high degree of predictive accuracy on cross-validated data (Fig. 3b). We used the machine learning method XGBoost, which outperformed other classification methods (Extended Data Fig. 6). For each user, we predicted their test results either using data before the test (pre-test), which would be most useful in predicting COVID-19 cases in the absence of molecular testing, or using data before and after the test (all data) as a benchmark for the best possible prediction we could make using all available data. We considered: (1) a symptoms-only model, which included only the most common COVID-19 symptoms listed by the CDC; (2) an expanded model, which further incorporated other features observed in the survey; and (3) a minimal-features model, which retained only the four most predictive features (loss of taste and/or smell, exposure to someone with COVID-19, exposure in the household to someone with confirmed COVID-19 and exposure to household members with COVID-19 symptoms) (see Methods and Supplementary Tables 11–14). The symptoms-only model achieved a cross-validated area

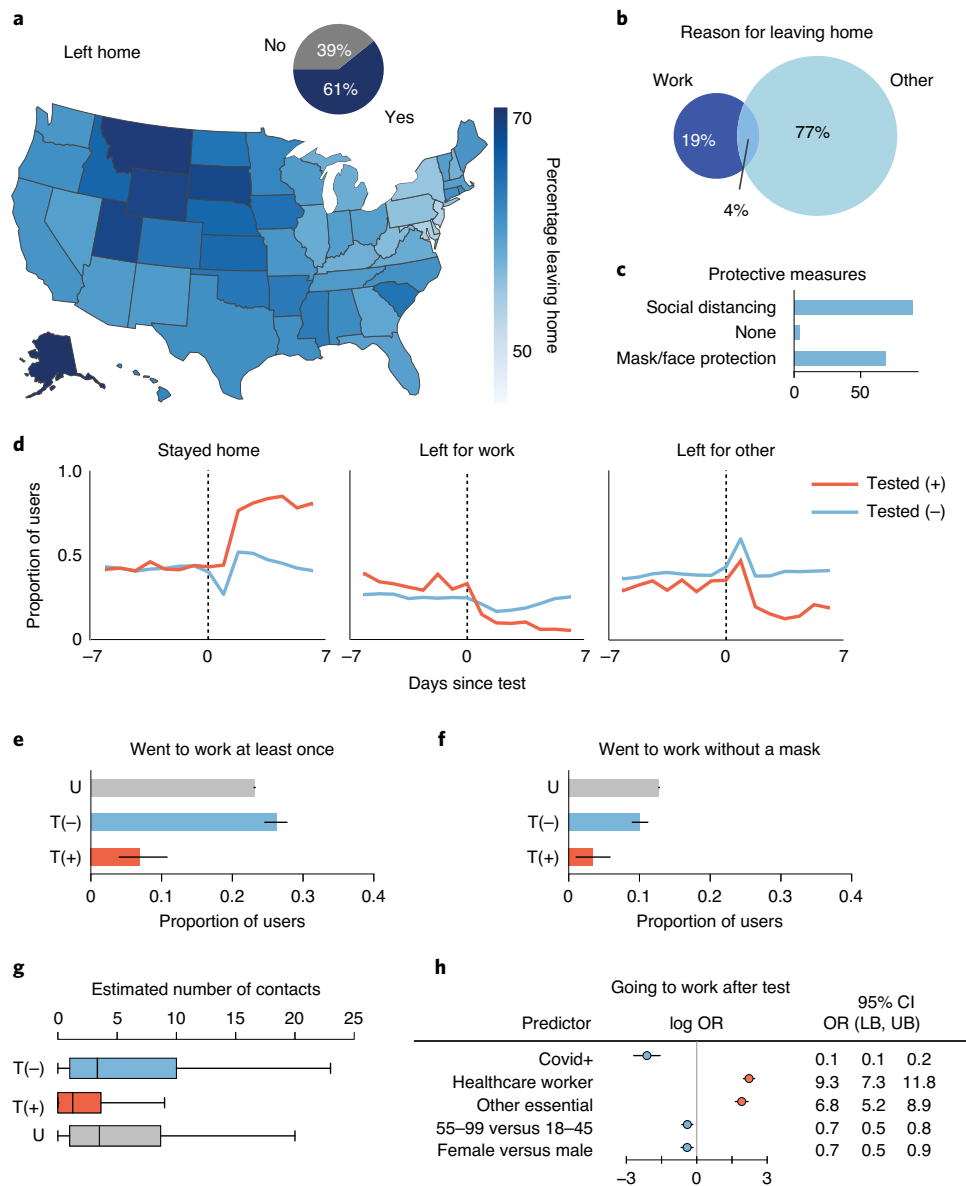
under the receiver operating characteristic (ROC) curve (AUC) of 0.76 using data before and after a test, and AUC 0.69 using just the pre-test data. Expanding the set of features to include other survey questions substantially improved performance (cross-validated AUC 0.92 all data, 0.79 pre-test). In the minimal-features model, we were able to retain high accuracy (cross-validated AUC 0.87 all data, AUC 0.80 pre-test) despite only including four questions, one referring to a symptom and three referring to potential contact with known infected individuals. Restricting the observed inputs to the 1,613 users (89 positive, 1,524 negative) who answered the survey in the 14 d before being tested limited the sample size and reduced the overall accuracy, but the relative performance of the models was similar (Fig. 3b).

The fact that a fraction of SARS-CoV-2(+) users report no symptoms or only less-common symptoms (Fig. 2g) raises the possibility that many infected users might behave in ways that could spread disease, such as leaving home while unaware that they are infectious. In spite of widespread shelter-in-place orders during the sample period, we found extensive heterogeneity across the United States in the fraction of users reporting leaving home each day, with 61% of the responses from 24 April to 12 May indicating the user had left home that day (Fig. 4a). The majority (77%) of these users reported leaving for non-work reasons, including exercising; 19% left for work (Fig. 4b). Of people who left home, a majority of users, but not all, reported social distancing and using face protection (Fig. 4c). Different states had persistently different levels of people wearing masks and leaving home (Extended Data Fig. 9). This incomplete shutdown with partial adherence, and lack of total social and physical protective measures, coupled with insufficient isolation of infected cases, may contribute to continued disease spread.

Given the large number of people leaving home each day, it is important to understand the behaviour of people who are potentially infectious and therefore likely to spread SARS-CoV-2. To this end, we further analysed the behaviour of people reporting to be PCR test (+) or (–). There was an abrupt, large increase in users reporting staying home after receiving a positive test result (Fig. 4d,e). Many, but not all, PCR test (+) users reported staying home in the 2–7 d after their test date (7% still went to work,  $n=14$  of 203 users), whereas 23% ( $n=62,483$  of 269,833 users) of untested and 26% ( $n=664$  of 2,533 users) of PCR test (–) users left for work (Fig. 4d,e). Similarly, 3% of PCR test (+) ( $n=7$  of 203 users) users reported going to work without a mask, in contrast with untested (12.7%,  $n=34,481$  of 269,833 users) and PCR test (–) (10%,  $n=255$  of 2,533 users) users (Fig. 4f). Positive individuals reported coming into close contact with a median of 1 individual over 3 days in contrast to individuals who tested negative or were untested, who typically came into close contact with a median of 4 people within 3 days (Fig. 4g). Regression analysis suggested that healthcare workers (OR, 9.3; 95% CI, 7.3, 11.8;  $P<0.001$ ) and other essential workers (OR, 6.8; 95% CI, 5.2, 8.9;  $P<0.001$ ) were much more likely to go to work after taking a positive or negative test, and PCR-positive users were more likely to stay home (OR, 0.1; 95% CI, 0.1, 0.2;  $P<0.001$ ) (Fig. 4h and Supplementary Table 15).

## Discussion

Using individual-level data collected from the HWF app, we showed that incorporating information beyond symptoms—in particular, household and community exposure—is vital for identifying infected individuals from self-reported data. This finding is particularly important for risk assessment at the early stage of transmission (for example, during the latent and presymptomatic periods when subjects have not developed symptoms yet), so that high-risk subjects can have priorities for being tested and quarantined and close contacts can be traced, to block the transmission chain early on. Our results show that vulnerable groups include subjects with household and community exposure, healthcare workers and essential workers,



**Fig. 4 | Behavioural factors potentially contributing to COVID-19 spread.** **a**, Proportion of responses indicating users leaving home across the United States (map) or overall (inset pie chart) ( $n=1,934,719$  responses from 279,481 users). **b**, Percentage of responses for users reporting work or other reason for leaving home ( $n=1,176,360$  responses from 244,175 users). **c**, Reported protective measures taken by users upon leaving home per response ( $n=1,176,360$  responses from 244,175 users). **d**, Time course of proportion of SARS-CoV-2 PCR-tested (+) or (-) users staying home, leaving for work and leaving for other reasons ( $n=4,396$  total users who reported being tested positive or negative in the V3 survey and responded on at least 1 d within  $\pm 1$  week of being tested). **e, f**, Proportion of users SARS-CoV-2 PCR-tested (+) or (-), or untested, going to work at least once (**e**) ( $n=14$  of 203 positive, 664 of 2,533 negative, 62,483 of 269,833 untested), and going to work without a mask (**f**) ( $n=7$  of 203 positive, 255 of 2,533 negative, 34,481 of 269,833 untested), who responded within the 2-7 d post-test for tested (T) or the 3 weeks since last check-in for untested (U). Healthcare workers and other essential workers are compared with non-essential workers as the baseline. **g**, Average reported number of contacts per 3 d in the 2-7 d after their test date. T(+),  $n=138$  users; T(-),  $n=2,269$  users; U,  $n=254,751$  users. LB, lower bound; UB, upper bound. **h**, Logistic regression analysis of factors contributing to users going to work in the 2-7 d after their COVID-19 test ( $n=678$  users going to work of 2,736 users with definitive test outcome and survey responses in the 2-7 d after their test date).

and African-American and Hispanic/Latinx users. They are at higher risk of infection and should have priorities for being tested and protected. Our findings also show statistically significant racial disparity after adjusting for the effects of pre-existing medical conditions, which needs to be addressed.

We find evidence among our users for several factors that could contribute to continued COVID-19 spread despite widespread implementation of public health measures. These include a

substantial fraction of users leaving their homes on a daily basis across the United States; users who claim to not socially isolate or return to work after receiving a PCR test (+) result; self-reports of asymptomatic, mildly symptomatic or presymptomatic presentation; and a much higher risk of infection for people with within-household exposure.

That said, we note several limitations of this study. The HWF user base is inherently a non-random sample of voluntary users of

a smartphone app, and hence our results may not fully generalize to the broader US population. In particular, the study may be subject to selection bias by not capturing populations without internet access, such as low-income or minority populations, who may be at elevated risk, and over-representation of females. Our results are based on self-reported survey data, and hence may suffer from misclassification bias—particularly those based on self-reported behaviours. Moreover, a relatively small percentage of subjects received PCR testing. As shown in Fig. 2, the subjects who were tested were more likely to be symptomatic, healthcare workers and essential workers, and people of colour. Naïve regression analysis of test results using responses of subjects who were tested could be subject to selection bias. To mitigate this, we have attempted to correct for these selection biases via the inverse probability weighting approach by estimating the selection probability, the probability of receiving tests, using the observed covariates (see Methods). Some residual bias may persist if there remain some unobserved factors related to underlying disease status and receiving a test, or if the selection model is misspecified. What is more, the HWF user base may not be representative of the broader US population. Although our regression analysis conditioned on a wide range of covariates to account for possible selection bias, if any unobserved factors associated with underlying disease status are also related to using the app—for example, health literacy and access to the internet, particularly in vulnerable groups such as low-income families—the results may be subject to additional selection bias.

Although there is enormous economic pressure on states, businesses and individuals to be able to return to work as quickly as possible, our findings highlight the ongoing importance of social distancing, mask wearing and large-scale testing of symptomatic, asymptomatic and mildly symptomatic people, exposure assessment and, potentially, even more rigorous ‘test-trace-isolate’ approaches<sup>45–48</sup> as implemented in several states, such as Massachusetts, New York, New Jersey and Connecticut, which have bent the infection curve<sup>45–48</sup>. Applying predictive models on a population scale will be vitally important to provide an ‘early warning’ system for timely detection of a second wave of infections in the United States and for guiding an effective public policy response.

As testing resources are expected to continue to be limited, HWF results could be used to identify which groups should be prioritized, or potentially to triage individuals for molecular testing based on predicted risk. HWF’s integration of behavioural, symptom, exposure and demographic data provides a powerful platform to address emerging problems in controlling infection chains, to rapidly assist public health officials and governments with developing evidence-based guidelines in real-time and to stop the spread of COVID-19.

## Methods

**Ethics statement.** The HWF application was approved as exempt by the Ethical & Independent Review Services LLP IRB (Study ID 20049–01). The analysis of HWF data was also approved as exempt by Harvard University Longwood Medical Area Institutional Review Board (IRB) (Protocol no. IRB20-0514) and the Broad Institute of MIT and Harvard IRB (Protocol no. EX-1653). Informed consent was obtained from all users and the data were collected in de-identified form.

**Open-source software.** We used the following open-source software in the analysis:

- Numpy: <https://www.numpy.org> (ref. 49)
- Matplotlib: <https://www.matplotlib.org> (ref. 50)
- Pandas: <https://pandas.pydata.org/> (ref. 51)
- Scikit-learn: <https://scikit-learn.org/stable/index.html> (ref. 52)
- SciPy: <https://www.scipy.org> (ref. 53)
- Statsmodels: <https://www.statsmodels.org/stable/index.html> (ref. 54)
- R: <http://www.r-project.org> (ref. 55,56)
- Tidyverse: <http://www.tidyverse.org> (ref. 57)
- Data.table: <https://CRAN.R-project.org/package=data.table> (ref. 58)
- sampleSelection<sup>59</sup>

**Application.** The HWF application was developed in React Native (<https://reactnative.dev/>), using Google App Engine (<https://cloud.google.com/appengine>) and Google BigQuery (<https://cloud.google.com/bigquery>) for the backend, and launched on the Android and iOS platforms. Users were identified only with a device-specific randomly generated number. Users below the age of 18 were not allowed to use the application.

**Inclusion criteria.** If a user logged in multiple times in a day, only the first time was retained. We excluded any users who responded to a survey version on one day and then on a later day responded to an older survey version. We excluded any users who reported different genders on different days, and we excluded any observations with missing feeling, gender or smoking history.

Before survey V3, users responded only whether or not they received a COVID-19 test, and we assumed that they received a PCR test. In survey V3, users reported the type of test they received, and we excluded antibodies tests from analyses.

**Logistic regression: receiving a test.** The HWF app allows users to report previous COVID-19 test information, including test date, test type (swab versus antibody), test result (positive, negative or unknown), location of test and reason for receiving the test (Fig. 2). A user may report that the test result is not yet known, and then update this information in future check-ins. A test was considered to be ‘unique’ if it was reported by the same user with the same test date (including ‘NA’ (not available),  $n=11$ ) and type. For this analysis, ‘swab’ tests were assumed to be PCR-based tests for SARS-CoV-2. Tests with a reported test date before 1 January 2020 were excluded. Before V3, users were not asked about their test type. Tests from the same user with the same test date may have been missing a reported test type in earlier check-ins, but the user may have filled in this information at later check-ins; in this case, we consider this to be the same test and assign the reported test type. For each unique test, all test information (including result) from the user’s most recent check-in was used.

We compared testing data from HWF with the COVID Tracking Project (<https://covidtracking.com/>) for all 50 states and the District of Columbia. For comparison with HWF data used in this analysis, we extracted COVID Tracking Project data until 11 May 2020. Tests with a ‘not yet known’ test result were excluded from this analysis. In Extended Data Fig. 6, the left panel compares the number of unique swab tests divided by the number of unique users in HWF with the total tests per state (totalTestResults) reported by the COVID Tracking Project divided by the state population as estimated by the 2010 Census (<https://pypi.org/project/CensusData/>). The right panel compares the proportion of unique swab tests in HWF with a positive result with the proportion of tests in the COVID Tracking Project with a positive result.

For the analysis of who received a test, the outcome was 1 if a user reported a swab test, or 0 otherwise. We fit a logistic regression model using demographics, professions, exposure and symptoms, among other covariates. Time-varying measures (for example, symptoms) were averaged over their V3 survey responses. Analysis was conducted with the statsmodel package (v.0.11.1) in Python<sup>54,55</sup>. We reported the log ORs and ORs, along with corresponding 95% CIs. Supplementary Table 3 lists the covariates used in the selection (who received a test) regression model, as well as the estimated coefficients, 95% CIs and *P* values.

## Uniform manifold approximation and project for dimension deduction

(UMAP). Of the 3,661,716 survey responses collected by HWF up until 12 May 2020, 667,651 reported having at least one symptom (excluding ‘feeling\_not\_well’) from the set of 25 symptom questions asked across all surveys. Only these responses were used for UMAP analysis (Fig. 2d–f). Each of the 25 queried symptoms was treated as a binary variable. The input data were therefore a matrix of 667,651 survey responses with 25 binary symptom variables. UMAP was applied to this matrix following McInnes and Healy<sup>60</sup> using the Python package `umap-learn` with parameters: `n_neighbors=1000`, `min_dist=0.5`, `metric='hamming'`. The resulting two-dimensional embedding was plotted with different colourmaps for each response in Fig. 2. The distributions of all 25 symptoms are shown individually in Extended Data Fig. 4.

**Asymptomatic analysis.** Status of each symptom was categorized as a CDC symptom, a non-CDC symptom or asymptomatic (Fig. 2g). The CDC symptoms were defined as patients that reported feeling well or unwell with a dry cough, shortness of breath, chills/shaking, fever, muscle/joint pain, sore throat or loss of taste/smell. The non-CDC symptoms were defined as patients that reported feeling well or unwell with any symptoms that were not defined by the CDC, including abdominal pain, confusion, diarrhoea, facial numbness, headache, irregular heartbeat, loss of appetite, nasal congestion, nausea/vomiting, tinnitus, wet cough, runny nose and so on.

We restricted analysis to the subset of patients for which we observed symptom data on their test date. For each user that tested positive or negative, we categorized participants into three groups: {CDC symptoms, Non-CDC symptoms, Asymptomatic}. Participants were grouped into CDC symptoms if they reported any CDC symptoms and participants that reported only non-CDC symptoms were grouped in the Non-CDC symptoms category. Participants were considered



asymptomatic if they reported none of the above symptoms. Proportions were reported and graphically represented for each group in Fig. 2g.

**COVID-19 symptoms and dynamics.** In the HWF survey data up to 12 May 2020, a total of 8,429 unique users reported the result of a quantitative PCR (qPCR) COVID-19 test (1,067 positive, 7,362 negative) (Fig. 2h–j). For surveys V1–2, we assumed that all tests were qPCR tests since antibody tests were rare before 24 April. In the V3 survey (24 April to 12 May 12) the test type was explicitly asked. Among qPCR-tested users, each response was assigned a date in days relative to the self-reported test date. The aggregate fraction of responses reporting each symptom was visualized in a histogram in Fig. 2h. The aggregate fraction of responses reporting each symptom at each timepoint among users that tested positive was visualized in a heatmap in Fig. 2i. Figure 2j shows the element-wise log ratio of the positive-test and negative-test heatmaps. That is, each element =  $\log(\text{fraction positive responses reporting symptom at time } t / \text{fraction negative responses reporting symptom at time } t)$ . The heatmaps were smoothed by taking the average for each symptom within a sliding window of  $\pm 1$  d for visualization.

**Logistic regression: test results.** A large number of risk factor survey questions were added in V3 of the survey, so we restricted analysis to V3 survey data for the purposes of identifying risk factors associated with SARS-CoV-2(+) test results (Fig. 3a). User responses were selected using a symmetric 28-d window around the last reported COVID-19 swab test date for any given user. Users that had no reported test outcome, or reported both positive and negative outcomes in different responses, were removed. Users who identified as ‘other’ in the gender response were dropped due to small sample size. Median neighbourhood household income was estimated by mapping user ZIP codes to corresponding ZCTAs (ZIP code tabulation areas) from the census, and then using the American Community Survey 5-year average results from 2018 to infer a neighbourhood household income (B19013\_001E). Population density was calculated at the county level for each user based on data from the Yu Group at University of California at Berkeley<sup>61</sup>.

Race was a categorical variable, with distinct groups: ‘white’; ‘African-American’; ‘Hispanic/Latinx’; ‘Asian’; ‘multiracial’ for those who marked two or more race categories; ‘other’ for those who marked ‘other’, ‘Native American’ or ‘Hawaiian/Pacific Islander’; and ‘unknown’ for those who did not disclose their race. A given food source was marked as ‘True’ if the user had indicated the use of that food source over any response within the given time window.

Because the HWF app asks for a separate set of symptoms depending on whether or not the user reports feeling ‘well’, there is not a one-to-one correspondence between symptoms reported by those feeling ‘well’ and ‘not well’. We excluded symptoms that were only present among those feeling ‘well’ or only among those feeling ‘not well’. For symptoms reported by both those who were ‘well’ and ‘not well’, we combined them into single symptoms. Supplementary Table 2 shows the variables merged using the ‘any’ function. Each symptom’s responses were then averaged over all available responses over the 28-d window. Similarly, distribution of sleep was averaged across the time window.

Multiple logistic regression was performed using statsmodels with the binary response outcome being the swab test outcome (positive coded as 1, negative as 0) to estimate coefficients, which were converted to ORs using exponentiation. Supplementary Table 4 lists the covariates used in this outcome regression model, as well as the estimated coefficients, 95% CIs and *P* values.

To mitigate selection bias inherent in restricting the analysis to those who have received a test, we used several inverse probability weighting adjustments. The probability of selection was estimated via the logistic regression analysis of who received a test (Fig. 2c). These estimated selection probabilities were incorporated into the outcome model via inverse probability weighting, and we reported confidence intervals based on robust (sandwich-form) standard errors and bootstrap standard errors. As inverse probability weighting can be sensitive to very small selection probability, we truncated them at several different values, using 0.1 and 0.9; and 0.05 and 0.95. The results using the truncated IPW selection probabilities at 0.1 and 0.9 are reported in Fig. 3. The result using truncated IPW selection probabilities at 0.05 and 0.95 were similar. Supplementary Table 5 lists the covariates used in the outcome regression model with IPW truncation at 0.1 and 0.9, as well as the estimated coefficients, and 95% CIs. Confidence intervals were obtained by bootstrapping the entire model selection process with 2,000 replicates. Specifically, for each bootstrap replicate, the entire dataset was resampled with replacement, a new selection/probability model was fitted for who gets a test, followed by a new IPW model fit using the inferred propensities from the bootstrap sample. Coefficient estimates for the IPW models across the bootstrap samples were used to generate the confidence intervals and mean value of the coefficient.

For additional sensitivity analysis, we used the bivariate probit model with sample selection used in econometrics to simultaneously estimate a selection (who gets tested) equation and an outcome (who tests positive) equation incorporating the selection probability as an additional covariate. Due to possible collinearities, not all features could be used in both the selection model and the outcome model. Specifically, profession could only be included in the selection model, and thus should be interpreted with caution. Supplementary Table 6 lists the covariates

used in the full information maximum likelihood estimates of the selection and outcome regression model, as well as the estimated coefficients, 95% CIs and *P* values. Qualitatively, the trends observed in the simultaneous selection/outcome model fitting are similar to those found in the two-step selection + IPW outcome logistic models.

To address sample bias in the user distribution in comparison with the distribution of individuals in the United States, we employed a poststratification correction for non-probability sampling models as an additional analysis. Poststratification using age, gender, ethnicity and location was performed on the testing selection model which generates the IPWs for the testing positive model. The United States was subdivided into the nine major census regions (see Supplementary Table 7). A joint distribution of estimated population over age, gender, ethnicity and region was obtained from the American Community Survey 5-year estimates from 2018. The corresponding distribution of users was generated across the same variables, and the ratio between each cell in the census distribution and the user distribution was used as the corresponding inverse probability weight in the testing selection model. The testing selection model thus should represent a user’s probability of getting tested from a corrected user base distribution matching major US Census demographics. The census-corrected testing selection model was used to generate IPWs for the subsequent testing positive model and was otherwise performed in the same way as that calculated using only the probability of receiving a test, as calculated using the HowWeFeel samples. Bootstrapping was performed on the entire process. The coefficient estimates for the poststratification testing model are shown in Supplementary Table 8, while estimates and confidence intervals for the subsequent poststratified IPW test outcome model are shown in Supplementary Table 9. A comparison of results with and without poststratification can be found in Extended Data Fig. 9. A comparison of the census-based poststratification-corrected models with the uncorrected models can be found in Extended Data Fig. 7. Performing census-based poststratification correction yields model coefficients and confidence intervals that are similar compared with when no census-based poststratification is performed.

To assess whether or not the states with the largest numbers of users bias the results, we also performed a comparison between the selection and outcome models with IPW correction with and without users from California and Connecticut (Extended Data Fig. 7). When removing California and Connecticut data, coefficient estimates from the selection and outcome models remain largely similar, suggesting limited bias due to California and Connecticut. Moreover, there is an overall increase in confidence interval widths of the outcome model, reflecting an overall increase in variance. Together, this comparison suggests that the California and Connecticut user base adds observations without adding substantial bias that may make the overall sample and corresponding analyses unrepresentative of the entire US population.

**Household transmission analysis.** In the HWF survey V3, users were first asked if they were exposed to someone with confirmed COVID-19. If they answered ‘yes’, then they were asked if that person lived in their household. We removed users who answered something other than ‘yes’ to the first question and who answered the second question. Additionally, we restricted the analysis to users who reported a negative or positive COVID-19 swab test and those who reported two or more household members. The outcome of interest was the binary outcome of testing positive on the COVID-19 swab test. The exposure of interest was the binary variable of having a household member test positive for COVID-19; we grouped respondents who answered ‘no’ with those who did not answer the question regarding household members.

The rest of the analysis proceeded similarly to the analysis for Fig. 3a, including the covariates used and the symptom collapsing strategy for each user across their responses within the 2-week window before the test and 2-week window after the test. We also performed sensitivity analysis using symptoms before the test. The difference between this analysis and that in Fig. 3a is that the reference group for household exposure was any other exposure or no exposure, whereas the reference group for household exposure and for other exposure in Fig. 3a is no exposure.

For both the unadjusted and the adjusted analyses, we performed logistic regression without and with the covariates. Supplementary Table 10 shows that the 95% CIs were calculated on the log OR scale and then exponentiated to obtain ORs.

**Sensitivity analysis: Firth regression.** Because of the small number of users in the user base who received a SARS-CoV-2 PCR test (1.7%) and the small number of tested users who received a positive test (8.2%), it is possible for standard logistic regression to be biased. To address this issue, we performed sensitivity analysis with Firth regression<sup>62</sup>, as implemented in the logistf R package (<https://cran.r-project.org/package=logistf>). We found very little difference between the Firth regression results and the logistic regression results presented in the paper (Extended Data Fig. 8), indicating that the imbalance of tested users or users who tested positive was not so severe as to bias the results.

**Prediction models.** XGBoost was compared across different featurizations and subsets of the data to assess the predictiveness of the algorithm on the HWF test result data (Fig. 3c). Two datasets were generated according to the data selection

and featurization used in the regression analysis of COVID-19 swab test outcomes, with the difference between the two sets being the time span used for the window, and the inclusion of additional features not used for inference. In the pre-test dataset, the window was selected such that only responses from 14 days before the test up until the day before the last reported test were included for analysis. The post-test dataset, on the other hand, is identical to the regression analysis dataset, using data from 14 days before and after the last reported test. The features for the different feature sets are shown in Supplementary Tables 11–13. Mask wearing and social isolation were computed as time averages of the responses to these questions. Models were trained and tested using five-fold cross-validation over the datasets. Within each fold, an additional threefold cross-validation was performed on the training set to optimize model hyperparameters before testing on the test set of that fold (see Supplementary Table 14 for grid-search coordinates). Test set AUCs from each fold were averaged to form a final AUC estimate. Final ROC curves were computed using the combined test set scoring and test set labels from each fold.

In addition to the models shown in the main text, we tested a range of classifiers, feature sets and data aggregation strategies for their performance at predicting COVID-19 test results from HWF survey data (shown in Extended Data Fig. 6). Input data were restricted to V3 survey data collected between 24 April and 12 May, and to qPCR-tested users who responded within –10 and +14 d of their test: a total of 3,514 negative tests and 315 positive tests. Three different feature sets, each consisting of a series of binary input variables from the HWF survey, were used: 56 symptoms, 77 additional features or all 133 features together. Note that this featurization differs slightly from the featurization used in the logistic regression in Fig. 3a, the goal of which was estimation and inference rather than prediction. Each of the 3,829 qPCR-tested users responded between 1 and 25 times within the time window of analysis. To account for time and sparse response rates, we binned data across time in four different ways: (1) average response for each feature in the 9 d preceding the test data (pre-test); (2) average response from –10 to +14 d (average); (3) binning the data into 3 weeks ([–10, –1], [0, 7], [8, 14]) and averaging each separately, creating a separate time-indexed feature label for each time bin (week\_bins\_avg); or (4) imputing the response for days with no data by backfilling, then forward filling, then proceeding as in point ‘(3)’ (week\_bins\_imp). The classifiers were implemented from the scikit-learn and XGBoost Python packages with the following parameter choices: LogisticRegression(), LassoCV(max\_iter=2000), ElasticNetCV(max\_iter=2000), RandomForestClassifier(n\_estimators=100), MLPClassifier(max\_iter=2000), XGBClassifier(). Hyperparameters for cross-validation (CV) methods were automatically optimized by grid-search using fivefold cross-validation. Mean AUC was calculated for each classifier using fivefold cross-validation.

**Post-test behaviour analysis.** Users with post-test information (in the 2–7 d) after their test date (or hypothetical test date for untested users) were collected and analysed (Fig. 4d–g). All featurization on this post-test window was identical to that of the selection/test outcome models. For computing whether a user went to work at least once, all responses for which users either leaving the house or not from V3 were used, and if any response for a user contained a ‘yes’ answer to leaving the house for work, the user was marked as leaving the home for work. Similar analysis was performed for leaving to work without a mask by marking the user as a ‘yes’ if they reported they were going to work and separately reported not using a mask when leaving the house that day. Proportions of each behaviour across the three populations (tested positive, tested negative and untested) were computed, and were bootstrapped with 2,000 replicates to generate confidence intervals.

Estimated number of contacts was performed similarly, except using the average value over individual user responses across the 2–7 d after their test.

Logistic analysis was performed to understand the effect of PCR test result on user behaviour in the 2–7 d after the test, adjusting for other potential covariates. Supplementary Table 15 lists the covariates used in the unadjusted outcome regression model, as well as the estimated coefficients, 95% CIs and *P* values.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

This work used data from the How We Feel project (<http://www.howwefeel.org/>). The data are not publicly available but researchers can apply to use the resource. Researchers with an appropriate IRB approval and data security approval to perform research involving human subjects using the HowWeFeel data can apply to obtain access to data used in the analysis.

## Code availability

The analysis code developed for this paper can be found online at <https://github.com/weallen/HWFPaper20>.

Received: 8 June 2020; Accepted: 5 August 2020;  
Published online: 26 August 2020

## References

- Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
- Wölfel, R. et al. Virological assessment of hospitalized patients with COVID-2019. *Nature* **581**, 465–469 (2020).
- Sanche, S. et al. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerg. Infect. Dis. J.* **26**, 1470–1477 (2020).
- Schuchat, A. Public health response to the initiation and spread of pandemic COVID-19 in the United States, February 24–April 21, 2020. *MMWR. Morb. Mortal. Wkly. Rep.* **69**, 551–556 (2020).
- Kraemer, M. U. G. et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).
- Chen, H., Qian, W. & Wen, Q. The impact of the COVID-19 pandemic on consumption: learning from high frequency transaction data. SSRN <https://doi.org/10.2139/ssrn.3568574> (2020).
- Smolinski, M. S. et al. Flu near you: crowdsourced symptom reporting spanning 2 influenza seasons. *Am. J. Public Health* **105**, 2124–2130 (2015).
- Segal, E. et al. Building an international consortium for tracking coronavirus health status. *Nat. Med.* **26**, 1161–1165 (2020).
- Lapointe-Shaw, L. et al. Syndromic surveillance for COVID-19 in Canada. Preprint at *medRxiv* <https://doi.org/10.1101/2020.05.19.20107391> (2020).
- Drew, D. A. et al. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. *Science* eabc0473 (2020).
- Menni, C. et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat. Med.* **26**, 1037–1040 (2020).
- Rossmann, H. et al. A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys. *Nat. Med.* **26**, 634–638 (2020).
- Lochlainn, M. N. et al. Key predictors of attending hospital with COVID-19: an association study from the COVID Symptom Tracker App in 2,618,948 individuals. Preprint at *medRxiv* <https://doi.org/10.1101/2020.04.25.20079251> (2020).
- Azad, M. A. et al. A first look at contact tracing apps. Preprint at *arXiv* <https://arxiv.org/abs/2006.13354v3> (2020).
- Krausz, M., Westenberg, J. N., Vigo, D., Spence, R. T. & Ramsey, D. Emergency response to COVID-19 in Canada: platform development and implementation for eHealth in crisis management. *JMIR Public Heal. Surveill.* **6**, e18995 (2020).
- Nguyen, L. H. et al. Risk of COVID-19 among front-line health-care workers and the general community: a prospective cohort study. *Lancet* [https://doi.org/10.1016/S2468-2667\(20\)30164-X](https://doi.org/10.1016/S2468-2667(20)30164-X) (2020).
- Lee, K. A. et al. Cancer and risk of COVID-19 through a general community survey. Preprint at *medRxiv* <https://doi.org/10.1101/2020.05.20.20103762> (2020).
- Mizrahi, B. et al. Longitudinal symptom dynamics of COVID-19 infection in primary care. Preprint at *medRxiv* <https://doi.org/10.1101/2020.07.13.20151795> (2020).
- Keshet, A. et al. The effect of a national lockdown in response to COVID-19 pandemic on the prevalence of clinical symptoms in the population. Preprint at *medRxiv* <https://doi.org/10.1101/2020.04.27.20076000> (2020).
- Shoer, S. et al. Who should we test for COVID-19? A triage model built from national symptom surveys. Preprint at *medRxiv* <https://doi.org/10.1101/2020.05.18.20105569> (2020).
- Hao, X. et al. Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature* **584**, 420–424 (2020).
- Verity, R. et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect. Dis.* **20**, 669–677 (2020).
- Onder, G., Rezza, G. & Brusaferro, S. Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *JAMA* **323**, 1775–1776 (2020).
- Maxmen, A. Thousands of coronavirus tests are going unused in US labs. *Nature* **580**, 312–313 (2020).
- Rader, B. et al. Geographic access to United States SARS-CoV-2 testing sites highlights healthcare disparities and may bias transmission estimates. *J. Travel Med.* <https://doi.org/10.1093/jtm/taaa076> (2020).
- How to use the data. *The COVID Tracking Project*. (Accessed 17 May 2020). <https://covidtracking.com/about-data>
- Centers for Disease Control and Prevention*. (accessed July 1, 2020) Coronavirus Disease 2019 (COVID-19); <https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html>
- Wei, W. E. et al. Presymptomatic transmission of SARS-CoV-2—Singapore, January 23–March 16, 2020. *MMWR Morb. Mortal. Wkly Rep.* **69**, 411–415 (2020).
- Linton, N. M. et al. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *J. Clin. Med.* **9**, 538 (2020).
- Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science* **368**, 489–493 (2020).

31. Sutton, D., Fuchs, K., D'Alton, M. & Goffman, D. Universal screening for SARS-CoV-2 in women admitted for delivery. *N. Engl. J. Med.* **382**, 2163–2164 (2020).
32. Baggett, T. P., Keyes, H., Sporn, P.-C. N. & Gaeta, J. M. Prevalence of SARS-CoV-2 infection in residents of a large homeless shelter in Boston. *JAMA* **323**, 2191–2192 (2020).
33. Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D. & Lessler, J. Variation in false-negative rate of reverse transcriptase polymerase chain reaction-based SARS-CoV-2 tests by time since exposure. *Ann. Intern. Med.* **172**, 577–582 (2020).
34. Griffith, G., Morris, T. T., Tudball, M., Herbert, A. & Mancano, G. Collider bias undermines our understanding of COVID-19 disease risk and severity. Preprint at *medRxiv* <https://doi.org/10.1101/2020.05.04.20090506> (2020).
35. World Health Organization. Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19). <https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf> (2020).
36. Nishiura, H. et al. Closed environments facilitate secondary transmission of coronavirus disease 2019 (COVID-19). Preprint at *medRxiv* <https://doi.org/10.1101/2020.02.28.20029272> (2020).
37. Park, Y. J. et al. Contact tracing during coronavirus disease outbreak, South Korea. *Emerg. Infect. Dis. J.* <https://doi.org/10.3201/eid2610.201315> (2020).
38. He, X. et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020).
39. Wang, Z., Ma, W., Zheng, X., Wu, G. & Zhang, R. Household transmission of SARS-CoV-2. *J. Infect.* **81**, 179–182 (2020).
40. Jing, Q.-L. et al. Household secondary attack rate of COVID-19 and associated determinants in Guangzhou, China: a retrospective cohort study. *Lancet Infect. Dis.* [https://doi.org/10.1016/S1473-3099\(20\)30471-0](https://doi.org/10.1016/S1473-3099(20)30471-0) (2020).
41. Bi, Q. et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *Lancet Infect. Dis.* **20**, 911–919 (2020).
42. County, S. et al. High SARS-CoV-2 attack rate following exposure at a choir practice. *MMWR Morb. Mortal. Wkly Rep.* **69**, 606–610 (2020).
43. Gibbins, J. D. et al. COVID-19 among workers in meat and poultry processing facilities. *MMWR Morb. Mortal. Wkly Rep.* **69**, 557–561 (2020).
44. McMichael, T. M. et al. Epidemiology of COVID-19 in a long-term care facility in King County, Washington. *N. Engl. J. Med.* **382**, 2005–2011 (2020).
45. Pan, A. et al. Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China. *JAMA* **323**, 1915–1923 (2020).
46. Clark, G. et al. *COVID-19 Pandemic: Some Lessons Learned So Far* (UK House of Commons Science and Technology Committee, 2020).
47. Finberg, H. V. Ten weeks to crush the curve. *N. Engl. J. Med.* **382**, e37 (2020).
48. Kim, J. Y. It's not too late to go on offense against the coronavirus. *New Yorker* <https://www.newyorker.com/science/medical-dispatch/its-not-too-late-to-go-on-offense-against-the-coronavirus> (20 April, 2020).
49. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
50. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
51. McKinney, W. Data structures for statistical computing in Python. in *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 51–56 (2010).
52. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
53. Virtanen, P. et al., and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
54. Seabold et al. Statsmodels: econometric and statistical modelling with Python. in *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 92 (2010).
55. R Core Team. *R: a Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
56. Xie, Y. *Dynamic Documents with R and knitr*. 2nd edn (Chapman and Hall/CRC, 2015).
57. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
58. Dowle, M. & Srinivasan, A. data.table: Extension of 'data.frame'. R package version 1.12.8 <https://cran.r-project.org/package=data.table> (2019).
59. Toomet O., Henningsen, A. Sample Selection Models in R: Package sampleSelection. *J. Stat. Software.* <https://doi.org/10.18637/jss.v027.i07> (2008).
60. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* <https://arxiv.org/abs/1802.03426> (2018).
61. Altieri, N., et al. Curating a COVID-19 data repository and forecasting county-level death counts in the United States. Preprint at *arXiv* <https://arxiv.org/abs/2005.07882> (2020).
62. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993).

## Acknowledgements

The How We Feel Project thanks operational volunteers A. Simon, R. Seidman, A. Ranganathan, C. O'Neil-Hart, D. Adler, D. Silbermann, J. Chou, L. Determann, M. Terry, R. Macrae, R. Barretto, R. Conway, S. Shenai, T. Falzone and Y. Shimabukuro. We also thank A. Tang for graphic design support. We thank the HWF participants who took our survey and allowed us to share our analysis. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The How We Feel Project is a non-profit corporation. Funding and in-kind donations for the How We Feel Project came from B. and D. Silbermann, F. Zhang and Y. Shi, L. Harp McGovern, D. Cheng, A. Azhir, K.H. Yoon and the Bill & Melinda Gates Foundation. X.L. acknowledges support from Harvard University and NCI grant no. R35-CA197449-05. F.Z. is supported by the Howard Hughes Medical Institute, the McGovern Foundation, and J. and P. Poitras and the Poitras Center.

## Author contributions

W.E.A., H.A.-T., J.B., X.J., G.M., A.S., R.R., N.N. and M.K. contributed to analysis. W.E.A., H.A.-T., J.B., X.J. and G.M. performed the majority of data cleaning, data analysis and figure production, and wrote the manuscript with F.Z. and X.L. A.S. and R.R. performed household transmission and symptom-type analysis. W.E.A. coordinated the analysis effort. P.G., E.M., J.A., T.P., W. Li, A.P., C.D., A.K., J.I., T.H., E.C., C.L., M.C., H.B., W. Lu, R.M., R.P. and B.S. designed and implemented the How We Feel application. B.A.C., M.T., J.O., C.S.G., O.S., G.K., B.A.C. and D.R.C. designed the survey, and provided feedback on app design and analysis. B.S. and F.Z. initiated the project. F.Z. and X.L. supervised all aspects of the work.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41562-020-00944-2>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41562-020-00944-2>.

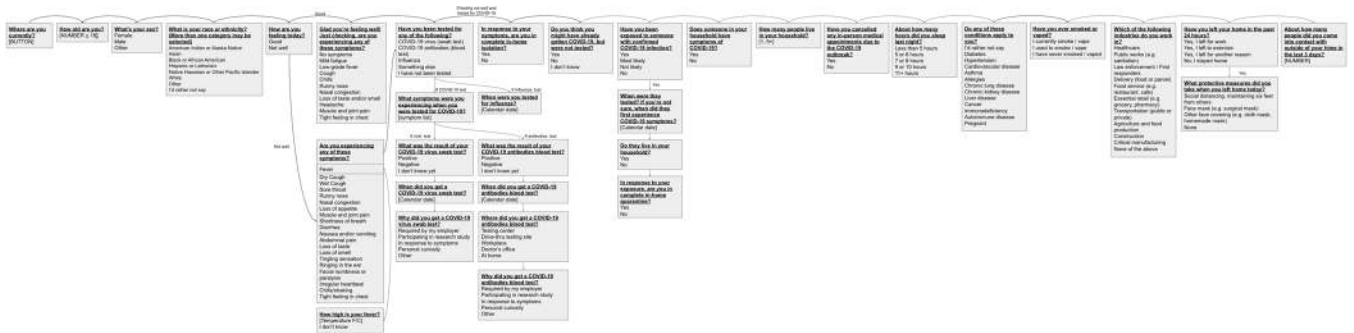
**Correspondence and requests for materials** should be addressed to W.E.A., F.Z. or X.L.

**Peer review information** Primary handling editor: Charlotte Payne.

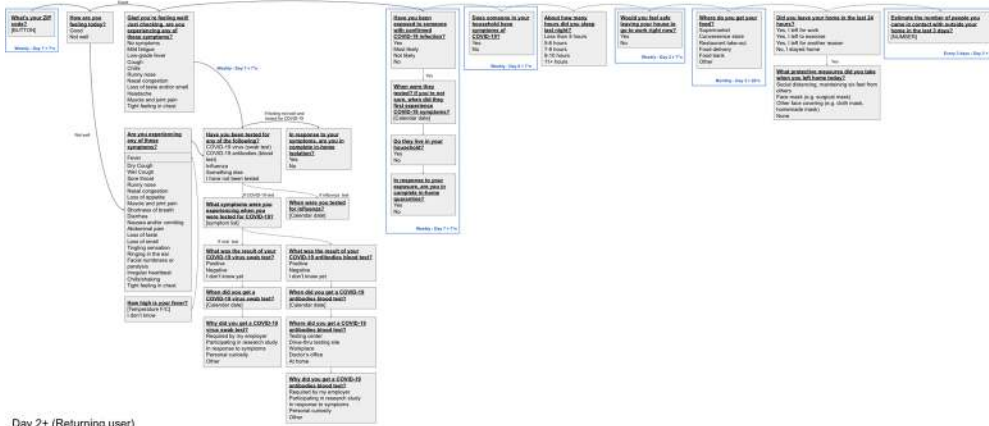
**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

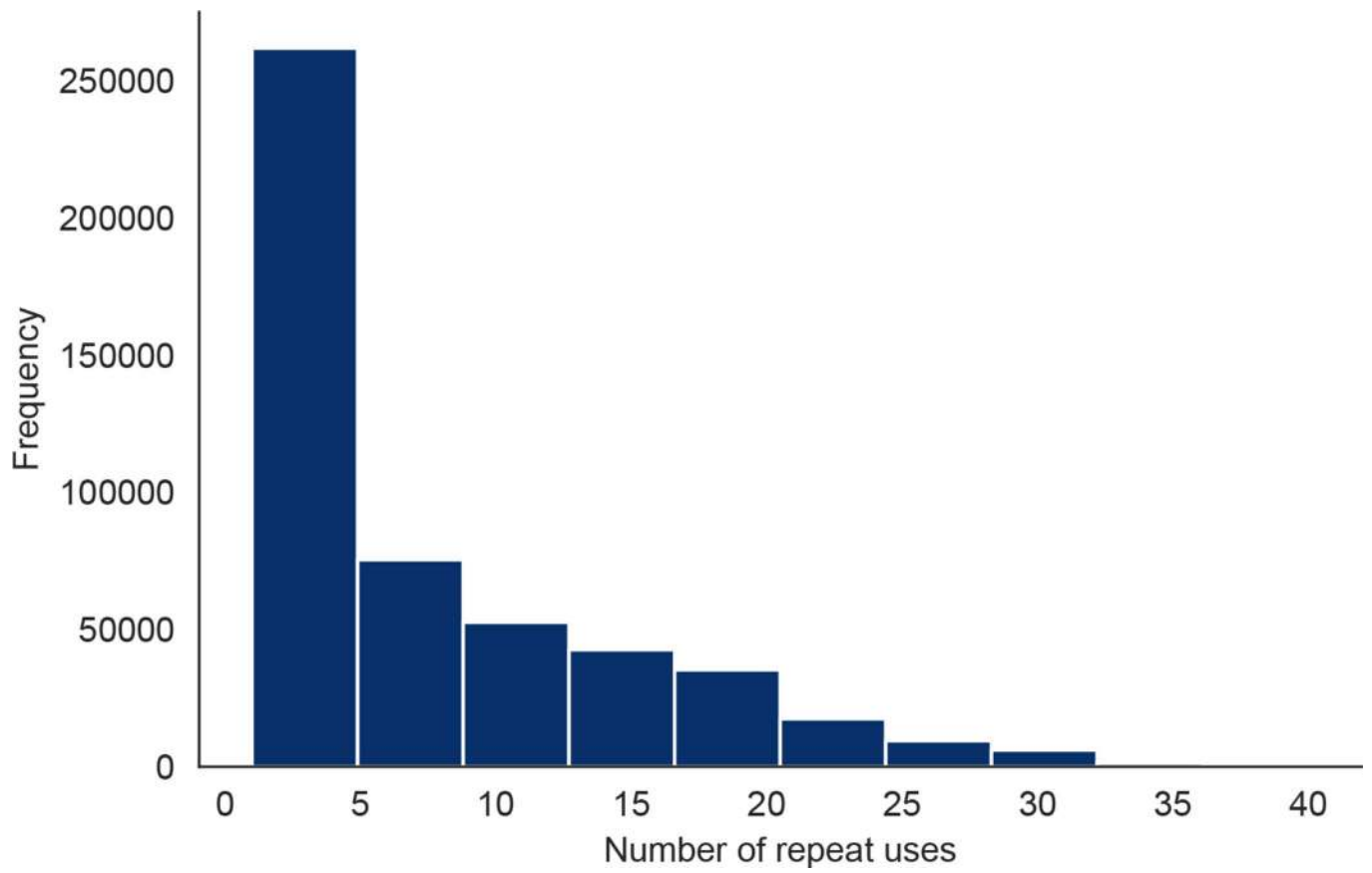


Day 1 (First-time user)

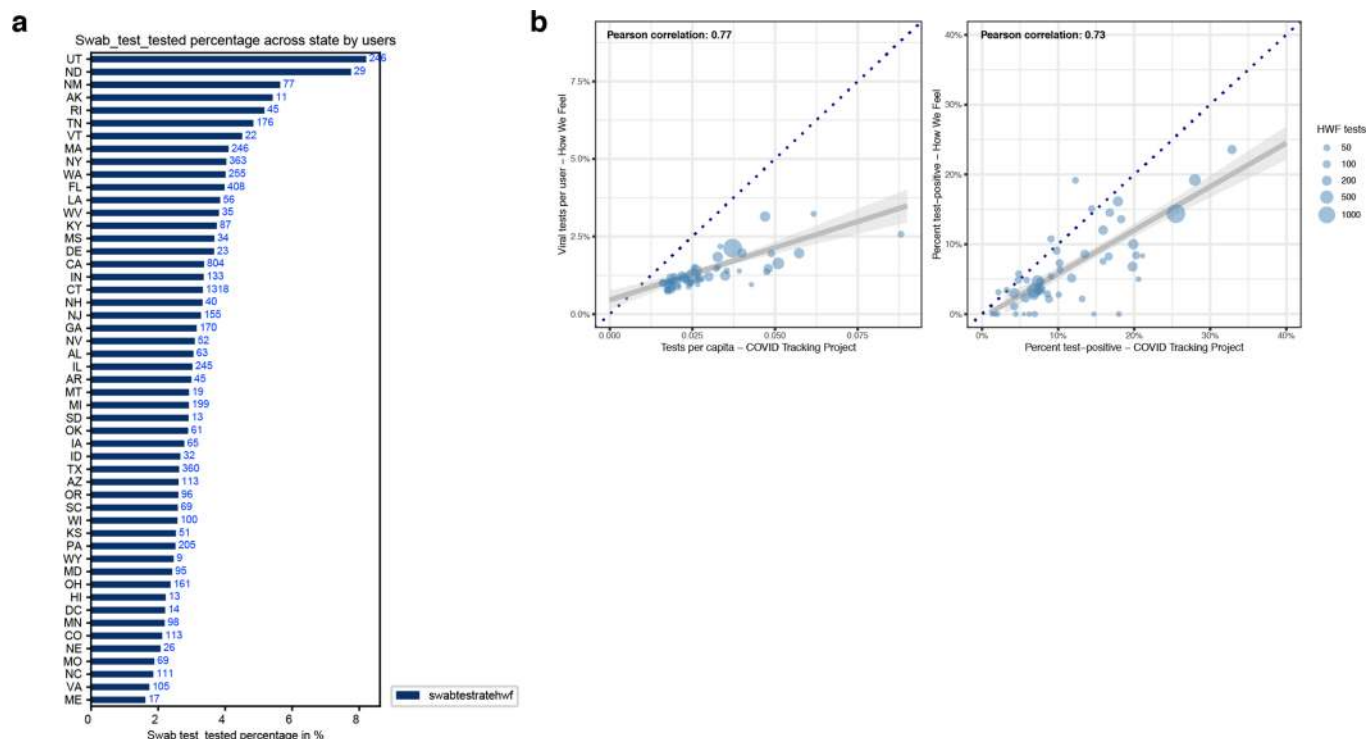


Day 2+ (Returning user)

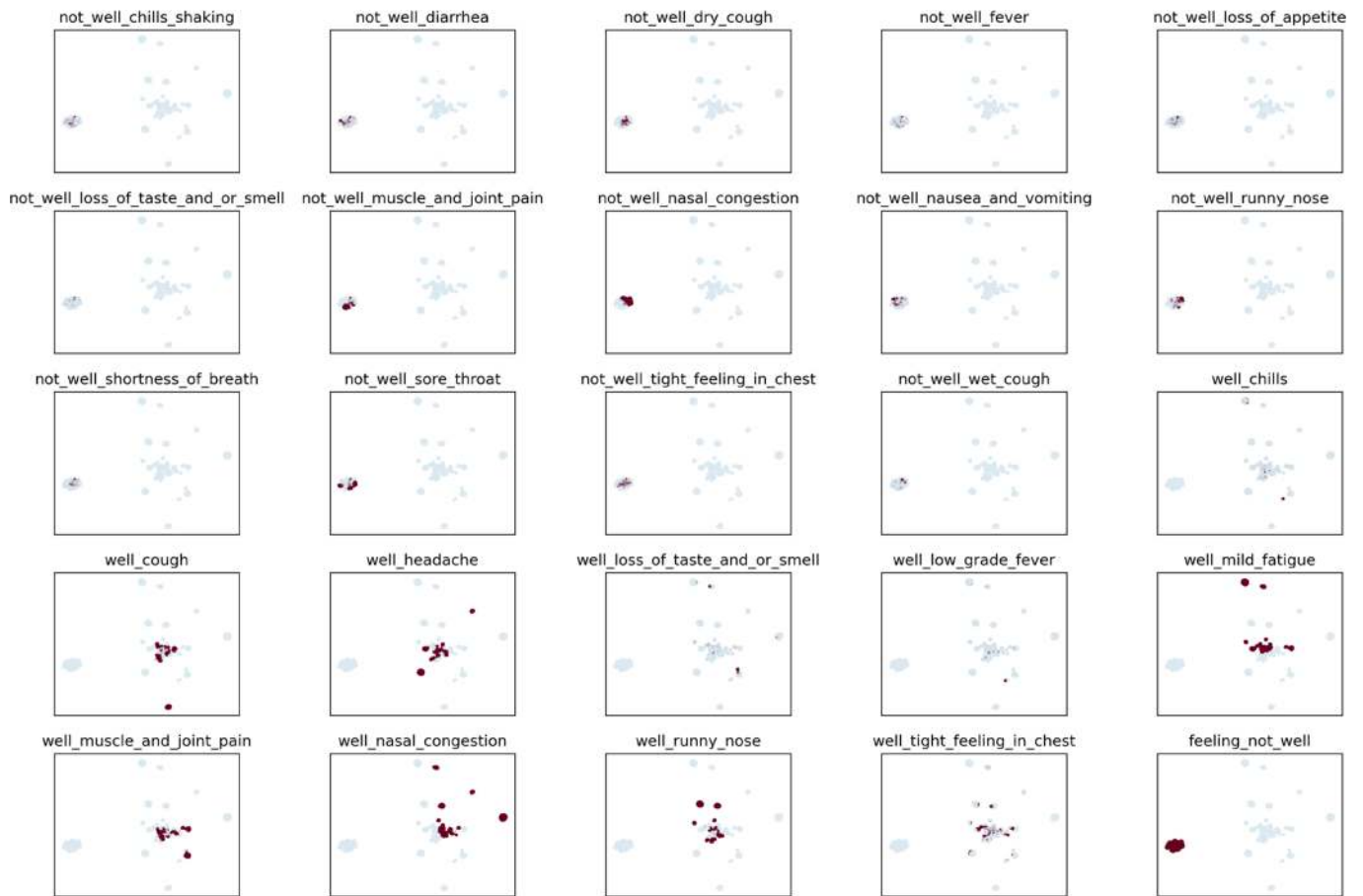
Extended Data Fig. 1 | HWF Survey Structure. Flow of questions through the HWF survey V3 for both first time users and returning users.



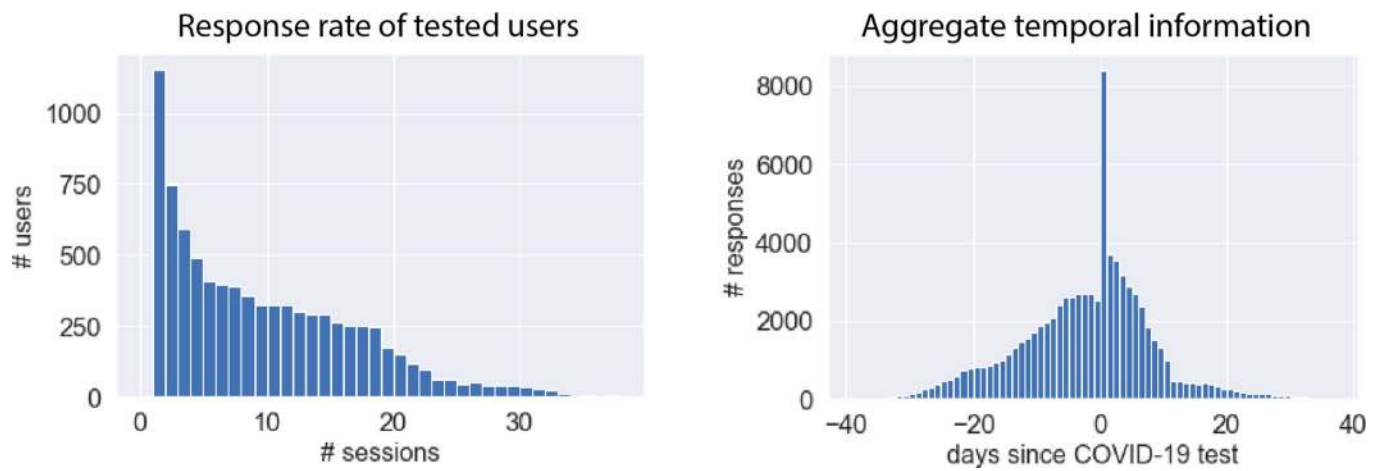
**Extended Data Fig. 2 | Number of Repeat Uses Per HWF User.** The number of times each HWF user checked into the app.



**Extended Data Fig. 3 | Analyses Regarding Receiving PCR-based Viral Tests.** **a**, A univariate plot of the frequency of people receiving a PCR-based viral test in each state. **b**, Correlations of viral tests per person (left) and percent of tests with positive results (right) comparing state-level data from How We Feel to testing data collected by the COVID Tracking Project. Each point represents a state, and the size of the point scales continuously with the total number of viral tests reported to How We Feel. Tests with an unresolved result at time of analysis were excluded. Several sizes shown in legend for reference. The dark blue dotted line is the  $x=y$  line and represents the expectation if sampling was random with respect to testing and test-positive results. The gray line is the best-fit linear regression line (and 95% CI) weighted by the number of viral tests reported to How We Feel.

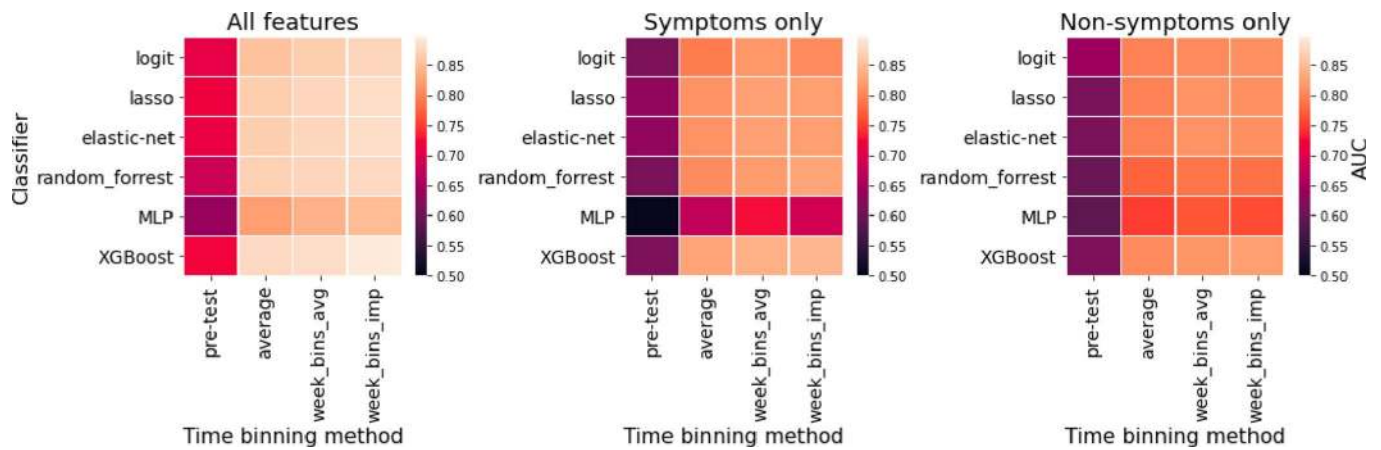


**Extended Data Fig. 4 | UMAP Visualization of Multivariate Self-Reported Symptom Structure.** Plots show individual distributions for 25 self-reported symptoms on the UMAP embedding shown in main text Fig. 2.



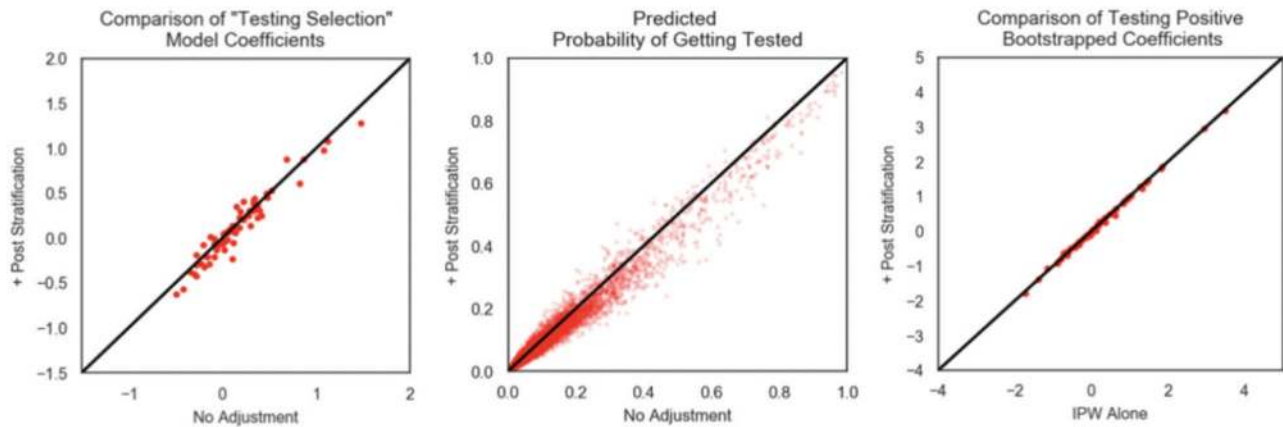
**Extended Data Fig. 5 | HWF Usage Over Time Per COVID-19 Tested User.** Left: Response rate of tested users. COVID-19 HWF users provided between 1 and 39 responses each, with a mean of 9 responses per user. Right: Aggregate temporal information showing number of responses relative to COVID-19 test date. In aggregate, we obtain > 1,843 survey responses each day within a window of 7 days of the COVID-19 test.



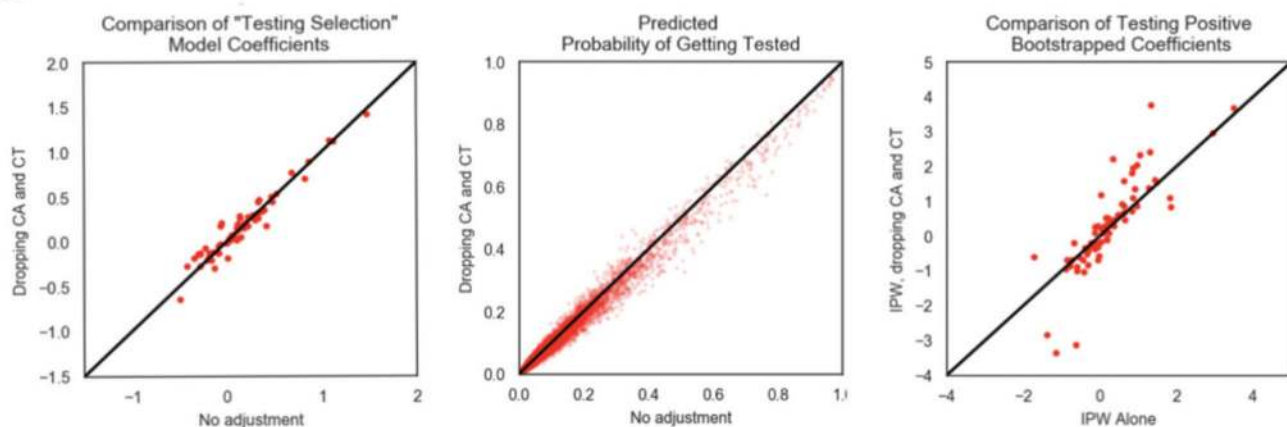


**Extended Data Fig. 6 | COVID-19 Test-Result Prediction Model Comparisons.** Six classified models (heatmap rows) were trained to predict COVID-19 test results from survey data among users tested within the V3 survey (N=3,829; 315 positive; April 24 - May 12), as assessed by cross-validation AUC measurement. Hyperparameters were optimized by grid search. The input survey data was treated in a variety of ways with models trained on either: the average of responses provided before the test (pre-test), the average of responses provided from 10 days before to 14 days after the test (average), the weekly average in this window (week\_bins\_avg), or the weekly average after imputing missing responses by back-filling (week\_bins\_imp). The analysis was performed on three different feature sets: all survey features (N=133), symptoms only (N=56) or non-symptoms only (N=77). The overall most accurate classifier was XGBoost, which was used for the analysis in Fig. 3.

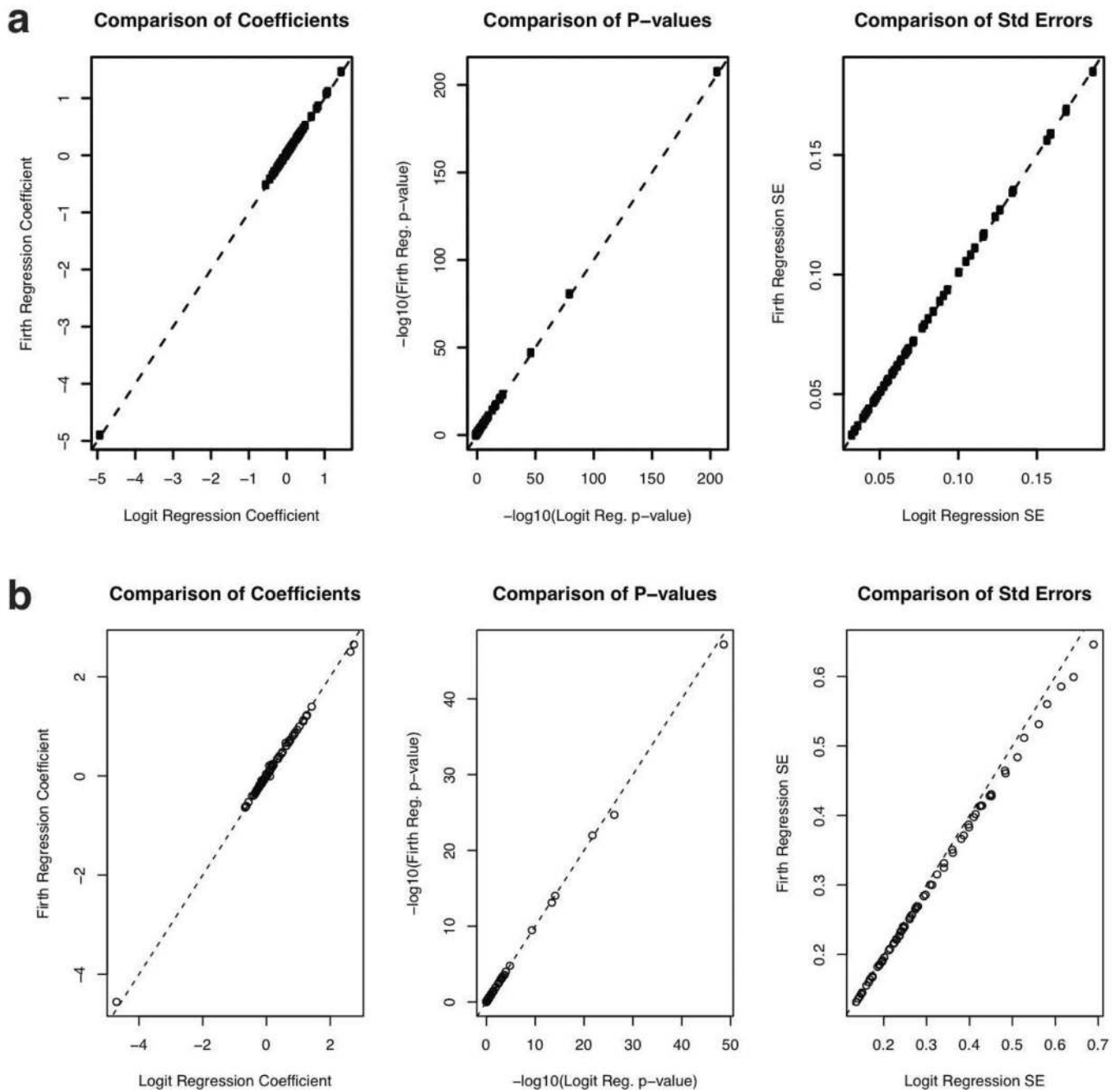
A



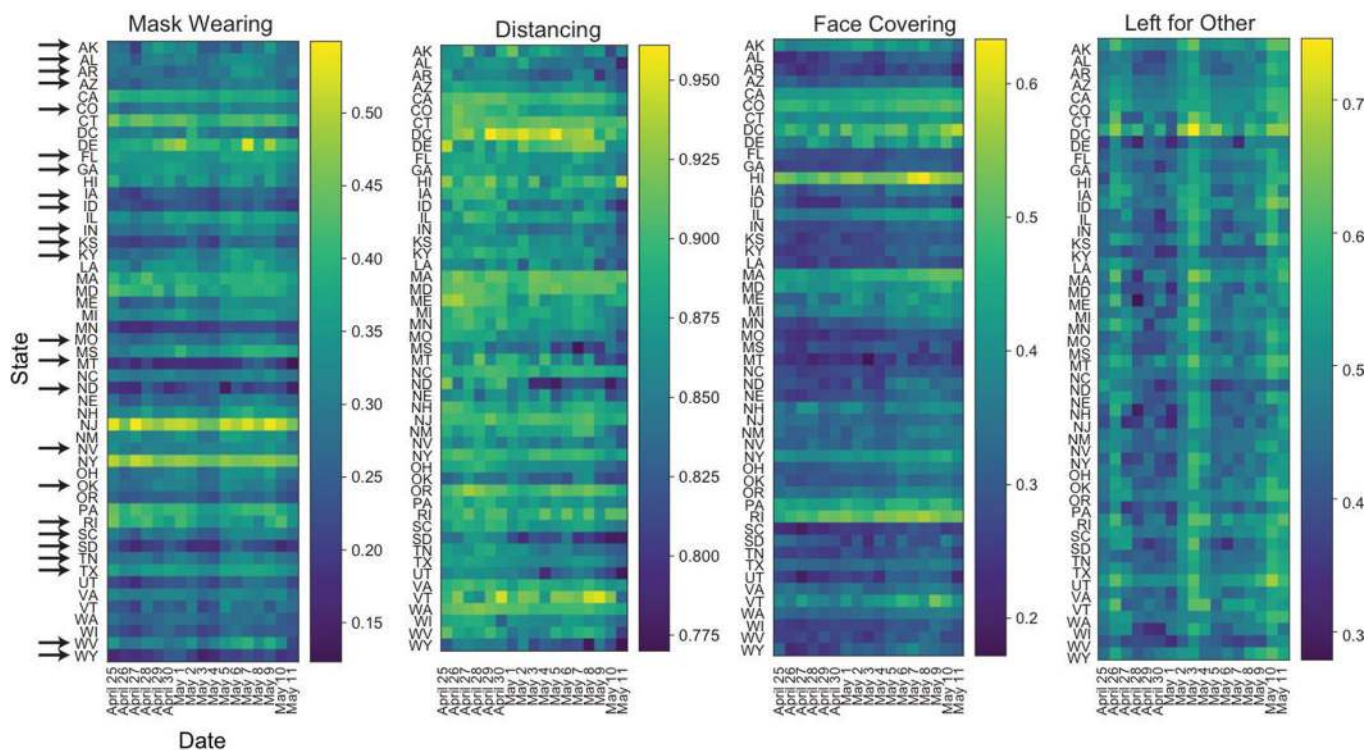
B



**Extended Data Fig. 7 | Results of Sensitivity Analyses for Biased Geographic Locations of Users and Demographics.** Comparison of testing outcome regression analysis between IPW correction alone and **a**, census based post-stratification + IPW correction and **b**, IPW correction on dataset with CT and CA users removed from the analysis. From left to right is 1) the comparison of the testing selection logistic regression model, 2) comparison of the predicted probability of getting tested using the testing selection logistic regression model, 3) comparison of the bootstrapped mean model coefficient from the testing outcome model, 4) comparison of the bootstrapped 95% confidence interval widths from the testing outcome model.



**Extended Data Fig. 8 | Firth regression sensitivity analysis.** **a**, Comparison of regression coefficients (left), p-values (center) and standard errors (right) from Firth regression (y-axis) vs. logistic regression from Fig. 2c in the manuscript (x-axis) for the model predicting which users would be tested. The dotted line is the identity ( $y = x$ ) line. **b**, Comparison of regression coefficients (left), p-values (center) and standard errors (right) from Firth regression (y-axis) vs. unweighted logistic regression from Fig. 3a in the manuscript (x-axis) for the model predicting which users among the tested users would test positive. The dotted line is the identity ( $y = x$ ) line.



**Extended Data Fig. 9 | Timecourse of User Behavior in Different States.** Time course of fraction of users in each state reporting wearing masks, socially distancing, covering their faces when leaving home, as well as leaving home for other reasons or for work from April 25 through May 11. Arrows indicate states that reopened before May 10. The wide dark bands in “Left for Work” and “Left for Other” correspond to weekends. Users per state: AK 487, AL 2590, AR 1858, AZ 5302, CA 28860, CO 6373, CT 45295, DC 749, DE 752, FL 12621, GA 6803, HI 702, IA 2797, ID 1483, IL 9799, IN 4882, KS 2476, KY 2879, LA 1882, MA 7174, MD 4696, ME 1242, MI 8157, MN 5269, MO 4544, MS 1176, MT 784, NC 7314, ND 451, NE 1508, NH 1425, NJ 5758, NM 1667, NV 2057, NY 11072, OH 8244, OK 2608, OR 4371, PA 9804, RI 1051, SC 3298, SD 551, TN 4513, TX 17088, UT 3755, VA 7239, VT 587, WA 7560, WI 4711, WV 1153, WY 440.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- |                                     |                                     |  |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A description of all covariates tested   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated   |

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection We used custom software made with React Native and Google App Engine to implement the application.

Data analysis Data analysis was performed in Python and R using the following packages: Numpy, Matplotlib, Pandas, Scikit-learn, scipy, statsmodels, Tidyverse, Data.table. Custom software for analysis will be available at: <https://github.com/weallen/HWFPaper20>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data include potentially identifiable subject-level data and so are available by request with a signed Data Use Agreement and IRB approval.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences  Behavioural & social sciences  Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This was a longitudinal association study of qualitative and quantitative data, including symptoms, demographic information, and behavior.
Research sample	The research sample are de-identified, volunteer users of a web and mobile-phone based application. The app was freely available for download and users
Sampling strategy	Users could freely download the application and use it for as many days as they wanted.
Data collection	Data were collected using a web and mobile-phone based application that allowed individuals to report demographic information, symptoms, and other COVID-19-related information in a de-identified way.
Timing	Data were collected from April 2 2020 to May 12 2020.
Data exclusions	If a user logged in multiple times in a day, only the first was retained. We excluded any users who responded to a survey version on one day and then on a later day responded to an older survey version. We excluded any users who reported different genders on different days, and we excluded any observations with missing feeling, gender, or smoking history.  Prior to survey version 3, users responded only whether or not they received a COVID-19 test, and we assumed that they received a PCR test. In survey version 3, users reported the type of test they received, and we excluded antibodies tests from analyses.
Non-participation	Users used the app on average 7 times.
Randomization	There was no randomization.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	See above.
Recruitment	Users could freely download the application. They heard about it through word of mouth, online advertising, government press releases, and media coverage. Users who had tested COVID-19 positive or were in high risk groups may have been more likely to use the app, which would bias the results towards more severe
Ethics oversight	The How We Feel application was approved as exempt by the Ethical & Independent Review Services LLP IRB (Study ID: 20049 – 01). The analysis of HWF data was also approved as exempt by Harvard University Longwood Medical Area IRB

(Protocol #: IRB20-0514) and the Broad Institute of MIT and Harvard IRB (Protocol #: EX-1653). Informed consent was obtained from all users and the data were collected in de-identified form.

Note that full information on the approval of the study protocol must also be provided in the manuscript.