BMJ Open

COVID-19 susceptibility and severity risks in a cross-sectional survey of over 500 000 US adults

Spencer C Knight,[1] Shannon R McCurdy,[1] Brooke Rhead,[1] Marie V Coignet,[1] Danny S Park,[1] Genevieve H L Roberts,[2] Nathan D Berkowitz,[1] Miao Zhang,[1] David Turissini,[1] Karen Delgado,[2] Milos Pavlovic,[2] AncestryDNA Science Team,[1,2] Asher K Haug Baltzell,[2] Harendra Guturu,[1] Kristin A Rand,[1] Ahna R Girshick,[1] Eurie L Hong,[1] Catherine A Ball[1]

To cite: Knight SC, McCurdy SR, Rhead B, *et al*. COVID-19 susceptibility and severity risks in a cross-sectional survey of over 500 000 US adults. *BMJ Open* 2022;**12**:e049657. doi:10.1136/bmjopen-2021-049657

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-049657).

SCK, SRM and BR contributed equally.

Received 30 January 2021
Accepted 27 July 2022

[1]Ancestry.com, San Francisco, California, USA
[2]Ancestry.com, Lehi, Utah, USA

Correspondence to
Dr Ahna R Girshick;
ahna.girshick@gmail.com

ABSTRACT

Objectives The enormous toll of the COVID-19 pandemic has heightened the urgency of collecting and analysing population-scale datasets in real time to monitor and better understand the evolving pandemic. The objectives of this study were to examine the relationship of risk factors to COVID-19 susceptibility and severity and to develop risk models to accurately predict COVID-19 outcomes using rapidly obtained self-reported data.

Design A cross-sectional study.

Setting AncestryDNA customers in the USA who consented to research.

Participants The AncestryDNA COVID-19 Study collected self-reported survey data on symptoms, outcomes, risk factors and exposures for over 563 000 adult individuals in the USA in just under 4 months, including over 4700 COVID-19 cases as measured by a self-reported positive test.

Results We replicated previously reported associations between several risk factors and COVID-19 susceptibility and severity outcomes, and additionally found that differences in known exposures accounted for many of the susceptibility associations. A notable exception was elevated susceptibility for men even after adjusting for known exposures and age (adjusted OR=1.36, 95% CI=1.19 to 1.55). We also demonstrated that self-reported data can be used to build accurate risk models to predict individualised COVID-19 susceptibility (area under the curve (AUC)=0.84) and severity outcomes including hospitalisation and critical illness (AUC=0.87 and 0.90, respectively). The risk models achieved robust discriminative performance across different age, sex and genetic ancestry groups within the study.

Conclusions The results highlight the value of self-reported epidemiological data to rapidly provide public health insights into the evolving COVID-19 pandemic.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ We performed association analyses for COVID-19 susceptibility and severity in a large, at-home survey and replicated much of the previous clinical literature.

⇒ We developed risk models and evaluated them across different age, sex and genetic ancestry cohorts, and showed robust performance across all cohorts in a holdout dataset.

⇒ The most severe cases, especially those resulting in mortality, were not sampled due to the self-reported nature of the data. As a result, many of the risk factor effect estimates may be underestimated for severe illness outcomes.

⇒ The AncestryDNA cohort is self-selected, slightly older, more European and more female than the broader US population.

⇒ Our results establish large-scale, self-reported surveys as a potential framework for investigating and monitoring rapidly evolving pandemics.

INTRODUCTION

The COVID-19 pandemic has resulted in over 346 million COVID-19 cases and over 5.5 million deaths worldwide,[1] including nearly 21 million cases and more than 870 000 deaths in the USA as of late January 2022.[2] The growing impact of the pandemic intensifies the need for real-time understanding of COVID-19 susceptibility and severity risk factors, not only for public health experts, but also for individuals seeking to assess their own personalised risk. Prior research has indicated that differences in COVID-19 *susceptibility*, defined in this study as a positive nasopharyngeal swab test result, are related to age,[3] sex-dependent immune responses[4] and genetics,[5 6] while heightened *severity* of COVID-19 illness, defined here as hospitalisation or progression to a critical case (intensive care unit (ICU) admittance, septic shock, organ failure or respiratory failure), is associated with risk factors such as age,[3 7–9] sex,[4 10–12] genetic factors[13] and underlying health conditions.[7 9 10 14–16] Self-reported survey data, which can easily be collected in the home, afford the opportunity to dynamically monitor the continually evolving pandemic and allow for real-time estimation of individual-level COVID-19 risk.[17–20] Furthermore, self-reported surveys allow for collection of information about known exposures, of which few epidemiological COVID-19 studies have explicitly accounted for in association analyses to date.[21]

In this paper, we aimed to replicate previous literature and to provide new insight into factors associated with susceptibility and severity of COVID-19 using a large survey cohort of 563 141 AncestryDNA customers who have consented to participate in the AncestryDNA COVID-19 Study.[5] We conducted the survey prior to widespread vaccine availability. We performed association tests of known or suspected COVID-19 risk factors with one susceptibility and two severity phenotypes and report unadjusted ORs and ORs adjusted for potential confounding factors. We additionally investigated associations of COVID-19 symptoms with susceptibility and severity.

We further demonstrate that this type of self-reported dataset can be used to build accurate predictive risk models for COVID-19 susceptibility and severity outcomes. For susceptibility, we designed two models and additionally applied two literature-based models[19] to predict COVID-19 cases among respondents reporting a test result. We also designed models to predict two different COVID-19 severity outcomes based on minimal information about demographics, health conditions and symptoms: hospitalisation due to COVID-19 infection and progression of an infection to a life-threatening critical case among those reporting a positive COVID-19 result.[14] To evaluate the potential for generalisability, we assessed performance of all of the risk models across different age, sex and genetic ancestry cohorts.

METHODS

Survey description

Survey responses were collected from AncestryDNA customers who consented to research in the USA between 22 April and 6 July 2020. The survey consisted of 50+ questions about COVID-19 test results, 15 symptoms among those who tested positive or who tested negative and had influenza-like symptoms, disease progression for positive testers, age, height, weight, known exposures to biological relatives, household members, patients or any other contacts with COVID-19, and 11 underlying health conditions (online supplemental tables 1 and 2). Collection of self-reported COVID-19 outcomes from US AncestryDNA customers who consented to research for the study and the survey design are described in more detail in a genome-wide association study on a very similar AncestryDNA dataset.[5] Here, participants reporting a negative test result were also assessed for symptoms and clinical outcomes.

Patient and public involvement

There was no patient or public involvement in the design, conduct, reporting or dissemination plans of this research.

Outcome definitions

The study assessed three outcomes: one for susceptibility and two for severity of COVID-19 infection. Cases for COVID-19 susceptibility were individuals who responded ‘Yes, and was positive’ to the question, ‘Have you been swab tested for COVID-19, commonly referred to as coronavirus?’ Responders who answered ‘Yes, and was negative’ were used as controls for the susceptibility analysis.

The hospitalisation outcome was defined among COVID-19-positive cases if a participant responded ‘Yes’ to a binary question about experiencing symptoms due to COVID-19 illness and ‘Yes’ to the hospitalisation question (‘Were you hospitalised due to these symptoms?’). Controls were defined by a response of ‘No’ to the symptoms question or a response of ‘No’ to the hospitalisation question in addition to reporting a self-reported positive COVID-19 test result.[5]

Critical cases of COVID-19 were defined via a response of ‘Yes’ to one or more questions about ICU admittance or, alternatively, self-reported septic shock, organ failure or respiratory failure resulting from a COVID-19 infection.[14] Controls were defined by a response of ‘No’ across all of these questions in addition to self-reporting a positive COVID-19 test result.

Genetic sex and ancestry definitions

All individuals were genotyped, using previously described general genotyping and quality control procedures.[22] Both sex and genetic ancestry were defined for individuals based on their genotypes. Genetic ancestry was estimated using a proprietary algorithm to estimate continental admixture proportions.[23] All participants were assigned to one of four broad genetic ancestry groups: European ancestry, admixed African-European ancestry, admixed Amerindian ancestry or other ancestry combinations.

Data preparation

Only complete case analyses were performed. Multiple-choice categorical questions were one-hot (‘dummy’) encoded as binary risk factors. We considered several risk factors and outcomes questions in our association analyses and risk modelling efforts, some of which are summarised in online supplemental tables 1 and 2. Based on the dependency structure of the survey, we made the following inferences:

- Participants reporting ‘No’ to a binary question about symptoms arising from COVID-19 infection were designated as negatives for dependent questions about individual symptoms, hospitalisation due to symptoms and ICU admittance due to symptoms.
- Participants reporting ‘No’ to a binary question about hospitalisation were assigned to hospital duration of

0 days and designated as negative for ICU admittance due to symptoms.

For association analyses, individuals were asked to score each of their symptoms ('Between the beginning of February 2020 and now, have you had any of the following symptoms? fever; shortness of breath; dry cough; nasal congestion; runny nose; sore throat; feeling tired or fatigue; chills; body aches; headache; cough-producing phlegm; abdominal pain; nausea or vomiting; diarrhoea; change in taste or smell') as 'None', 'Very Mild', 'Moderate', 'Severe' or 'Very Severe'. Responses for each symptom were converted to a binary variable based on the following mapping: 0=None, Very Mild, Mild; 1=Moderate, Severe, Very Severe, for a total of 15 binary symptom variables.

Body mass index (BMI) was calculated from responses to questions about individual height ('How tall are you?') and weight ('How much do you weigh?') as BMI=(weight in kilograms)/(height in metres)^2. A BMI beyond six SDs of the appropriate sex-stratified mean was considered equivalent to a non-response for BMI. We used BMI categories reported by the Centers for Disease Control and Prevention (CDC) for these analyses: underweight (BMI <18.5), healthy (18.5≤BMI<25), overweight (25≤BMI<30), obese (BMI ≥30), along with the subcategories for obesity: obesity I (30≤BMI<35), obesity II (35≤BMI<40) and obesity III (BMI ≥40).[24]

Pre-existing health conditions considered in these analyses were gathered from the response to ('Do you currently have any of the following health conditions? Select all that apply.') Allowed responses to this question were: asthma; COPD (chronic obstructive pulmonary disease); other lung condition; cancer (treated in the past year); cardiovascular disease; chronic kidney disease (CKD); diabetes; hypertension; organ failure requiring a transplant (in the last year); blood disorder requiring haematopoietic stem cell/bone marrow transplant; other autoimmune disease; other immunodeficiency disorder; other; none; not sure. The 'pre-existing health conditions, any' variable was binarised from the survey responses as 'Y' for individuals selecting at least one of the listed conditions and/or 'Other', and 'N' for individuals selecting 'None'. Individuals selecting 'Not sure' were omitted from the analysis.

Association analysis

Analyses were performed either with the *statsmodels* package in Python V.3 or in base R with the *glm* function. For each susceptibility and severity outcome and risk factor of interest, a simple logistic regression (LR) model was fit using unpenalised maximum likelihood (online supplemental tables 3–11).[25] Multiple LR was used to adjust the ORs for known COVID-19 exposures and potentially confounding risk factors. The adjusted model included age, sex and four known exposures (Y/N if any) for susceptibility outcomes; and age, sex, obesity (binarised if BMI ≥30) and health conditions (binarised if any) for severity outcomes. Individual adjustment variables were omitted when analysing associations for risk factors within equivalent categories (eg, age was not included in adjusted models for age bin risk factors). Complete case analyses were performed for adjusted models. No interaction effects were considered.

For each risk factor, 95% CIs for the log OR were estimated under the normal approximation. The significance threshold was Bonferroni corrected for the 42 different risk factors examined (adjusted threshold of 0.05/42=0.0012).[25]

Risk factor selection and risk model training

Three risk models were constructed to predict one of three binary outcomes: a positive test result among those reporting a test result (susceptibility); a hospitalised COVID-19 case among those reporting a positive test result (hospitalisation) and a critical COVID-19 case among those reporting a positive test result (critical case). Prior to model training, the data were split with a fixed-random seed into training and holdout datasets. We chose risk factors based on a minimal subset of nominally significant ORs within our training data as well as literature guidance.[3 4 7 9 11 12 14–16] For the susceptibility models without symptoms, we included a subset of exposure-related questions, based on the training OR analyses, as well as two demographic variables (age and sex). For susceptibility models with symptoms, we additionally included the five symptoms most differentiated between symptomatic negative and positive testers from our training ORs. For the severity models, we included pre-existing conditions, based on the training OR analyses, predictive symptoms within our training dataset, severe obesity (obesity III, BMI ≥40), age and sex. See online supplemental table 12 for the final set of risk factors selected for each risk model.

Once final risk factors were selected, we trained LR models with fivefold cross-validated grid search on the training dataset to select an optimal lasso regularisation parameter lambda.[25] For the grid search, we scanned eight different values for lambda, equally partitioned geometrically across a four-log space. We then retrained on the entire training dataset with the optimal lambda and evaluated the final model on the holdout dataset.

Model thresholding

Phenotypes were predicted from the output of trained models based on a 50% probability threshold (ie, logistic model output >0.5). Sensitivity and specificity were then calculated based on the true versus predicted binary outcomes.

Estimation of performance error

To estimate error in model performances, we bootstrapped our holdout dataset 1000 times to generate a sampling distribution for each evaluation metric. We estimated the mean and 95% CIs for each metric based on the mean and SD of this sampling distribution.[25]

RESULTS

Survey response and study population

A total of 563 141 responses were collected, with 4726 individuals reporting a COVID-19 positive test result, 28 872 a negative test result, 71 761 no COVID-19 test but influenza-like symptoms, and 454 542 no COVID-19 test and no influenza-like symptoms. A total of 3240 reported pending test results and were excluded from further analyses. The survey completion rate was approximately 95%. In general, the COVID-19 positive test rate and self-reported clinical outcomes were consistent with those reported by the US CDC over a similar period (online supplemental note 1).[26] The majority of participants were female (67.5%) and of European ancestry (75.4%), with some individuals of admixed Amerindian (6.5%) or admixed African-European (3.0%) ancestries. The median age of the entire cohort was 56, and the median age of those reporting a positive test result was 49 (table 1 and online supplemental tables 13–15). Case definitions are summarised in figure 1 and table 2.

Susceptibility associations: replicated and novel

We replicated many previously reported literature associations for susceptibility. The strongest associations for a positive COVID-19 test result were known COVID-19 exposures, either through a household case (OR=26.03;

Table 1 Study population demographic information

	COVID-19 nasopharyngeal swab test positive	COVID-19 nasopharyngeal swab test negative	Full AncestryDNA COVID-19 cohort
	n=4726	n=28 872	n=563 141
No COVID-19 nasopharyngeal swab test			
With influenza-like symptoms			71 761
Without influenza-like symptoms			454 542
Nasopharyngeal swab test results pending			3240
Age			
Median, mean (SD)	49, 49.49 (15.43)	53, 52.68 (15.5)	56, 53.90 (16.13)
Bins (counts, fraction)			
18–30	600 (0.13)	2496 (0.09)	52 580 (0.09)
31–40	941 (0.20)	5001 (0.17)	87 261 (0.15)
41–50	926 (0.20)	5333 (0.18)	90 473 (0.16)
51–60	1057 (0.22)	6009 (0.21)	108 062 (0.19)
61–70	735 (0.16)	5961 (0.21)	128 200 (0.23)
71–90	467 (0.10)	4072 (0.14)	96 565 (0.17)
Genetic sex (counts, fraction)			
Female	3013 (0.64)	19 945 (0.69)	380 349 (0.67)
Male	1706 (0.36)	8867 (0.31)	183 153 (0.32)
Genetic ancestry continental groupings			
Admixed African-European ancestry	275 (0.06)	1244 (0.04)	17 019 (0.03)
Admixed Amerindian ancestry	520 (0.11)	2336 (0.08)	36 865 (0.07)
European ancestry	3026 (0.64)	20 269 (0.70)	424 328 (0.75)
Other ancestry	905 (0.19)	5023 (0.17)	84 929 (0.15)
Pre-existing health conditions (any)	1911 (0.46)	15 261 (0.55)	255 788 (0.47)
BMI			
Median, mean (SD)	28.59, 29.83 (7.04)	28.67, 29.92 (7.05)	28.29, 29.53 (6.87)
Bins (counts, fraction)			
Underweight (BMI<18.5)	<100 (0.03*)	204 (0.01)	4407 (0.01)
Healthy (18.5≤BMI<25)	971 (0.25*)	6442 (0.24)	139 565 (0.27)
Overweight (25≤BMI<30)	1225 (0.31*)	8503 (0.32)	172 941 (0.33)
Obese (BMI≥30)	1629 (0.42*)	11 264 (0.43)	203 325 (0.39)
Symptoms (tested positive)			
General, yes (counts, fraction)	3862 (0.82)		9237 (0.02)
Hospitalisation (counts, fraction)	453 (0.10)		1397 (0.00)
Duration, days (median, mean (SD))	5, 7.62 (9.08)		3, 4.78 (7.22)

*Approximate % for privacy reasons; 100 counts used in place of <100.
BMI, body mass index.

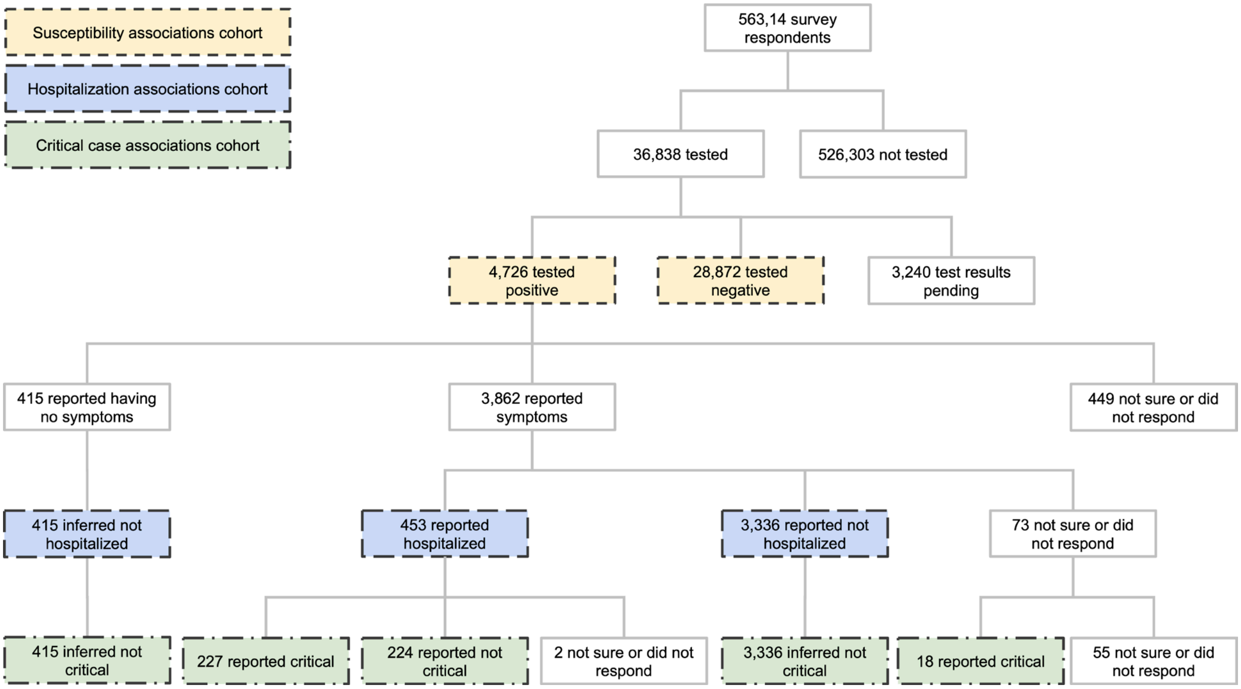
Susceptibility associations cohort
Hospitalization associations cohort
Critical case associations cohort
563,14 survey respondents
36,838 tested
526,303 not tested
4,726 tested positive
28,872 tested negative
3,240 test results pending
415 reported having no symptoms
3,862 reported symptoms
449 not sure or did not respond
415 inferred not hospitalized
453 reported hospitalized
3,336 reported not hospitalized
73 not sure or did not respond
415 inferred not critical
227 reported critical
224 reported not critical
2 not sure or did not respond
3,336 inferred not critical
18 reported critical
55 not sure or did not respond

Figure 1 Susceptibility and severity association cohort definitions. The susceptibility cohort for association analyses and risk models (short-dashed boxes) was comprised of a subset of the individuals who reported taking a nasopharyngeal swab test and receiving a positive or negative result. The severity cohort for the hospitalisation association analyses (long-dashed boxes) was comprised of those who reported receiving a positive test result. They were further subdivided into those who reported hospitalisation and those who did not (either directly or inferred, see the Methods section). The severity cohort for the critical case association analyses (dash-dotted boxes) was also comprised of those who reported receiving a positive test result. They were further subdivided into those who reported meeting the criteria for a critical case and those who did not (either directly or inferred, see the Methods section).

95% CI=22.26 to 30.43), biological relative (OR=5.77; 95% CI=4.99 to 6.68) or other source of 'direct' exposure (OR=6.94; 95% CI=6.02 to 7.99) (figure 2 and online supplemental table 3). In general, adjusting for known exposures, age and sex resulted in attenuation of the ORs, with many associations becoming insignificant after adjustment (figure 2 and online supplemental table 4).

One novel result was that the OR for men was not attenuated after adjustment, and men remained at elevated odds after adjusting for known exposures and age (adjusted OR (aOR)=1.36; 95% CI=1.19 to 1.55; figure 2 and online supplemental table 4). We also note that men and women reported comparable exposure burden, with men slightly more likely to report a household case of COVID-19 but less likely to report a case of COVID-19 among biological relatives (online supplemental tables 6 and 7).

Consistent with previous reports,[27–31] younger individuals (ages 18–29 years; OR=1.51; 95% CI=1.26 to 1.81) were significantly more likely to test positive compared with older individuals (ages 50–64 years, the largest age group in this cohort), and individuals of admixed African-European (OR=1.48; 95% CI=1.18 to 1.85) or admixed Amerindian ancestry (OR=1.49; 95% CI=1.26 to 1.77) were more likely to test positive compared with those of European ancestry (figure 2 and online supplemental

Table 2 Case definitions

Outcome	Case definition	Total cases	Total controls
COVID-19 positive test result	Self-reported positive COVID-19 swab test result	4726	28 872
COVID-19 hospitalisation	Self-reported positive COVID-19 swab test result and hospitalisation due to COVID-19 symptoms	453	3751
COVID-19 critical case	Self-reported positive COVID-19 swab test result, hospitalisation due to COVID-19 symptoms, and one or more of: ICU admittance, ICU admittance with oxygen, respiratory failure, septic shock, or multiple organ dysfunction or failure OR Self-reported positive COVID-19 swab test result and one or more of: respiratory failure, septic shock, or multiple organ dysfunction or failure	245	3975

ICU, intensive care unit.

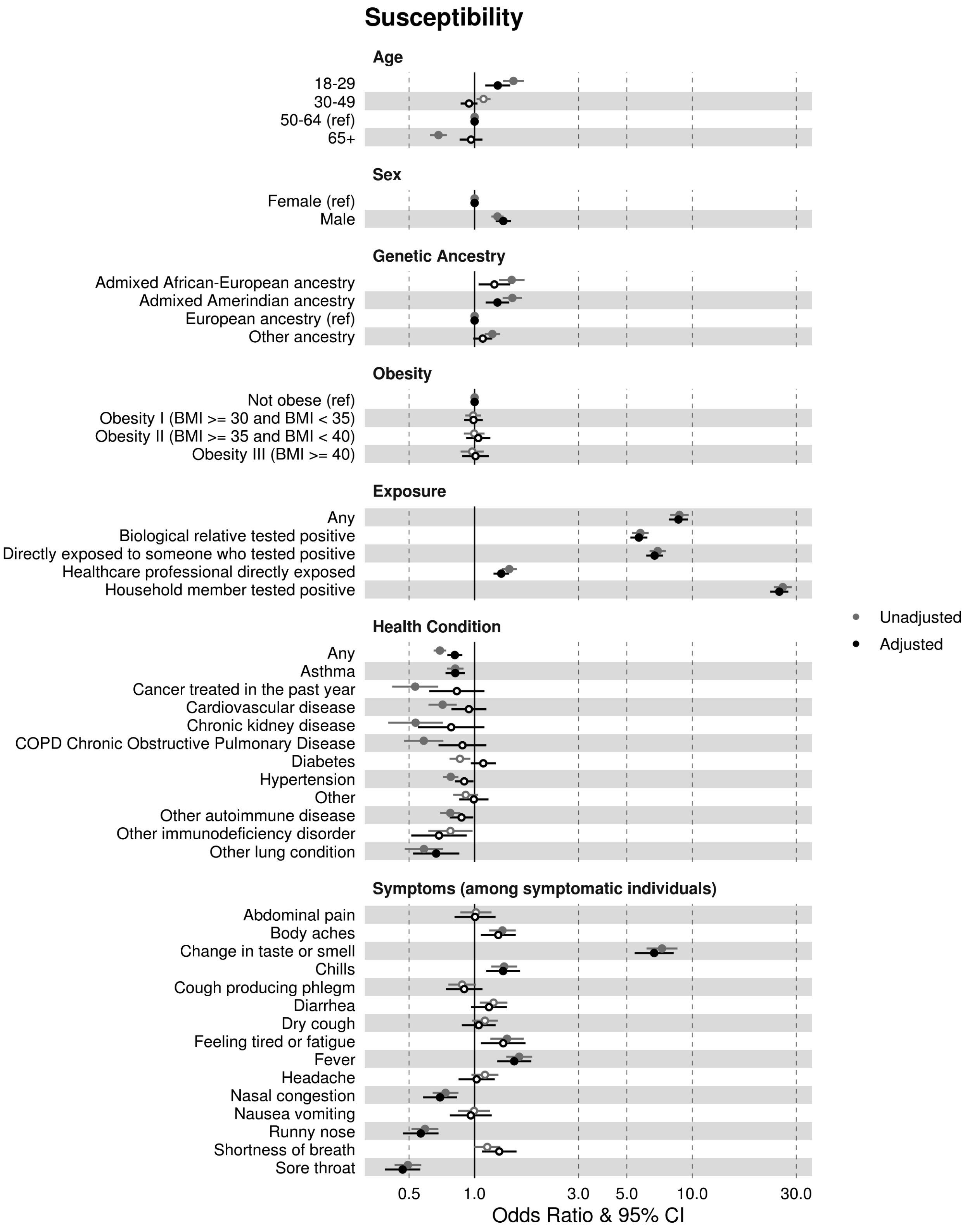

Figure 2 Susceptibility (positive test result) ORs and 95% CIs estimated from simple ('unadjusted models', grey) and multiple ('adjusted models', black) logistic regression with adjustment for other risk factors. Open circles indicate not significant (p>0.05) after accounting for multiple hypothesis tests using Bonferroni correction. Age, sex, genetic ancestry and obesity ORs were estimated in relation to the reference variables indicated. Exposure, health and symptom ORs were each estimated separately as binary variables. Symptom ORs were estimated as binary variables among symptomatic testers only (see the Methods section). Risk factor adjustments for susceptibility include: sex, age and at least one known COVID-19 exposure. Where applicable, individual adjustment variables were omitted to avoid duplicate adjustment (see the Methods section). BMI, body mass index.

table 3). Individuals in all three of these groups reported higher levels of COVID-19 cases within the household, cases among biological relatives, and/or other known 'direct' COVID-19 exposures (online supplemental tables 5–7). Adjusting for age (ancestry groups only), sex and known exposures attenuated the OR for all of these groups (younger aOR=1.28; 95% CI=1.03 to 1.59, African-European aOR=1.23; 95% CI=0.94 to 1.62, and Amerindian aOR=1.27; 95% CI=1.04 to 1.57; figure 2 and online supplemental table 4).

Individuals reporting pre-existing medical conditions (eg, cancer, cardiovascular disease, CKD, diabetes, hypertension) were less likely to test positive for COVID-19 (figure 2 and online supplemental table 3). We observed significantly decreased odds of a known 'direct' exposure to COVID-19, as well as significantly decreased odds of a household case of COVID-19, among such individuals relative to those without any health conditions (OR=0.71; 95% CI=0.65 to 0.78 and OR=0.74; 95% CI=0.65 to 0.84, respectively; online supplemental tables 5 and 6).

Replicated associations for COVID-19 severity

Consistent with previous reports,[7 9 12 14–16] we observed positive associations between certain health conditions and COVID-19 severity outcomes; many of these associations remained significant after adjustment for age, sex and obesity (BMI ≥30) (figure 3 and online supplemental tables 8–11). COVID-19 cases reporting at least one underlying health condition were significantly more likely to progress to a critical case (OR=2.85; 95% CI=1.78 to 4.57; figure 3, online supplemental figure 1 and online supplemental table 10). Specific underlying health conditions that were associated with hospitalisation and/or critical case progression included CKD, COPD, diabetes, cardiovascular disease and hypertension (figure 3, online supplemental figure 1 and online supplemental tables 9 and 11). Among individuals testing positive for COVID-19, the oldest (≥65 years) were significantly more likely to be hospitalised compared with those aged 50–64 years (OR=1.70; 95% CI=1.13 to 2.56; figure 3 and online supplemental table 8). Individuals of admixed African-European ancestry who tested positive were significantly more likely to report progression to a critical case, compared with those with European ancestry (OR=2.07; 95% CI=1.03 to 4.17; online supplemental figure 1 and online supplemental table 10). Among COVID-19 cases, men were significantly more likely than women to report progression to a critical case (OR=1.54, 95% CI=1.00 to 2.37; online supplemental figure 1 and online supplemental table 10); these findings are consistent with CDC reports of increased ICU admittance rates in men (3% vs 2%).[26]

Differential symptomatology between susceptibility and severity

We compared associations between susceptibility and severity to provide a more nuanced view of symptoms and other risk factors associated with susceptibility versus those associated with severity (figure 4 and online supplemental figure 2).[18 19 32] Among symptomatic people reporting a COVID-19 test result, those reporting change in taste or smell (OR=7.26; 95% CI=5.54 to 9.50), fever (OR=1.60; 95% CI=1.28 to 2.01), or feeling tired or fatigue (OR=1.41; 95% CI=1.05 to 1.89) were more likely to test positive (figure 4 and online supplemental table 3). Those reporting runny nose (OR=0.59; 95% CI=0.47 to 0.75) or sore throat (OR=0.49; 95% CI=0.39 to 0.62) were more likely to test negative, consistent with previous reports that these symptoms are more indicative of influenza or the common cold (figure 4 and online supplemental table 3).[18 19 32] Change in taste or smell, a hallmark symptom of COVID-19 infection, was not associated with hospitalisation (OR=0.77, 95% CI=0.55 to 1.07; figure 4 and online supplemental table 8). By contrast, dyspnoea (shortness of breath) was strongly associated with hospitalisation and critical case progression (OR=7.52; 95% CI=4.92 to 11.49 and OR=11.55; 95% CI=5.91 to 22.59, respectively),[33] but was not associated with susceptibility (OR=1.14; 95% CI=0.91 to 1.44; figure 4 and online supplemental tables 3, 8 and 10).

Predictive risk models

We further developed risk models that predict an individual's COVID-19 risk (susceptibility or severity, see the Methods section).[7 17–19 34 35] The susceptibility models were designed to predict a COVID-19 result (positive or negative) from risk factors among testers. We compared four models: our model based on demographics and exposures only ('Dem+Exp'); our model based on demographics, exposures and symptoms ('Dem+Exp+Symp'); and for benchmarking purposes, a replication of a previously published model called 'How We Feel' based on nearly identical self-reported symptoms ('HWF Symp'), and one which also included self-reported exposures ('HWF Exp+Symp') (online supplemental note 2 and online supplemental table 12).[19] The risk factors for the models 'Dem+Exp' and 'Dem+Exp+Symp' were selected from our training dataset (online supplemental table 16) and/or guidance from the literature (see the Methods section).

All four susceptibility models performed robustly; the three models that included one or more symptoms outperformed the model without symptoms (Dem+Exp), underscoring the value of self-reported symptoms for discriminating between cases and controls (figure 5, see online supplemental tables 17–20 for detailed model performance data). The model with demographics, exposures and symptoms (Dem+Exp+Symp) achieved the highest overall performance with an area under the curve (AUC) of 0.94±0.02, a sensitivity of 85% and a specificity of 91% (online supplemental note 3 and figures 3–4). Each of the models performed comparably across different age, sex and genetic ancestry cohorts (figure 5 and online supplemental tables 17–20). We observed no significant overfitting in any of the models as evidenced

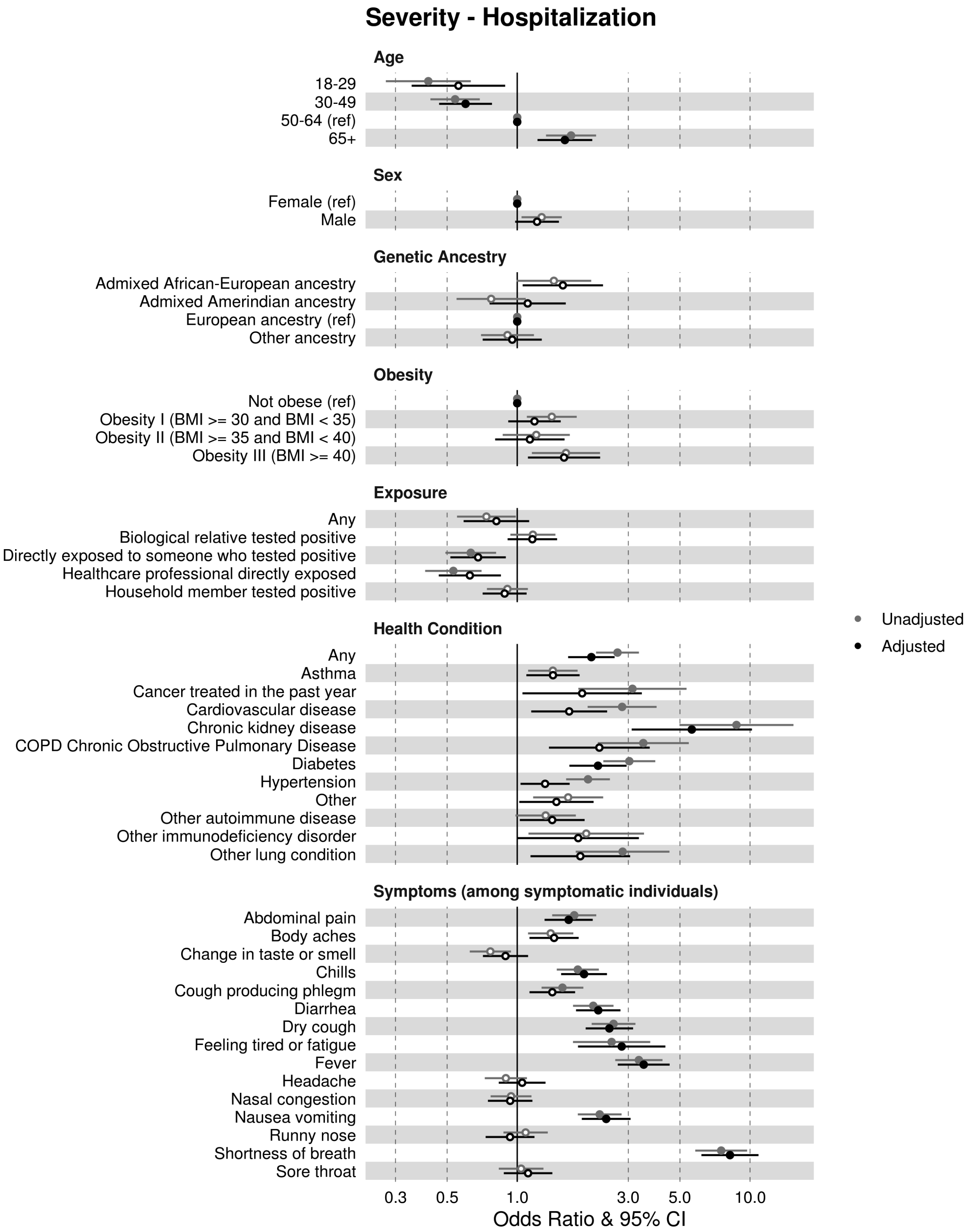

Figure 3 Severity (hospitalisation) ORs and 95% CIs estimated from simple ('unadjusted models', grey) and multiple ('adjusted models', black) logistic regression with adjustment for other risk factors. Open circles indicate not significant (p>0.05) after accounting for multiple hypothesis tests using Bonferroni correction. Age, sex, genetic ancestry and obesity ORs were estimated in relation to the reference variables indicated. Exposure, health and symptom ORs were each estimated separately as binary variables. Symptom ORs were estimated as binary variables among symptomatic testers only (see the Methods section). Risk factor adjustments for severity include: sex, age, obesity (binarised if BMI ≥30) and underlying health conditions (Y/N if any). Where applicable, individual adjustment variables were omitted to avoid duplicate adjustment (see the Methods section). See online supplemental figure 1 for critical case severity ORs. BMI, body mass index.

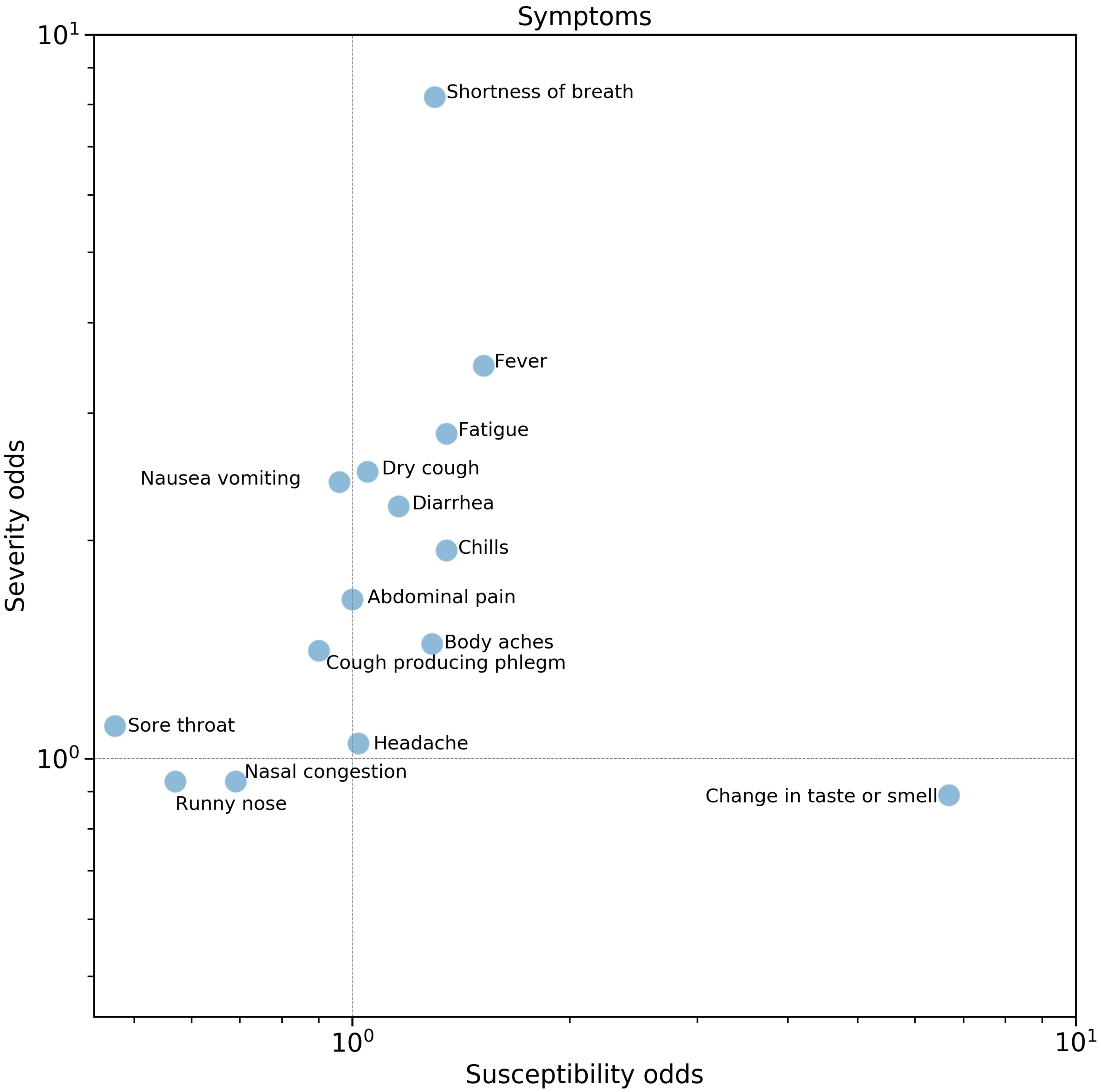

Figure 4 Comparison of susceptibility-adjusted ORs (horizontal axis) and severity-adjusted ORs (vertical axis) for symptoms in figures 2 and 3. Severity aORs are for hospitalisation. Note that aORs for susceptibility and severity are adjusted differently according to descriptions in figures 2 and 3 captions. The aORs are plotted on a log scale for visibility. Shortness of breath is the strongest indicator of increased severity, while change in taste or smell is the strongest indicator for testing positive for COVID-19 among symptomatic individuals (see the Methods section). Refer to online supplemental figure 2 for demographic, health condition and exposure aORs. aORs, adjusted ORs.

by comparable train–test performances (online supplemental table 21).

We trained two severity models, designed to predict either hospitalisation or progression to critical illness among COVID-19 cases. We included a number of risk factors and symptoms most associated with severe COVID-19 outcomes from the literature and/or our training dataset (figure 4 and online supplemental tables 22 and 23); these included age,[7–9 14] sex,[4 7 11 12 14] severe obesity (obesity III, BMI ≥40)[7 36] and health conditions,[7 9 12 14–16] as well as symptoms including shortness of breath,[33] fever, feeling tired or fatigue, dry cough and diarrhoea. Both models performed robustly on an independent holdout dataset (AUCs of 0.87±0.03 and 0.90±0.03 for the hospitalisation and critical models, respectively; figure 5). The severity models performed comparably when stratifying by age, sex and genetic ancestry (figure 5 and online supplemental tables 24 and 25), and there was no significant overfitting bias as evidenced by comparable train–test performances (online supplemental table 21).

DISCUSSION

The AncestryDNA COVID-19 Study provides a highly complete, self-reported dataset that contains information

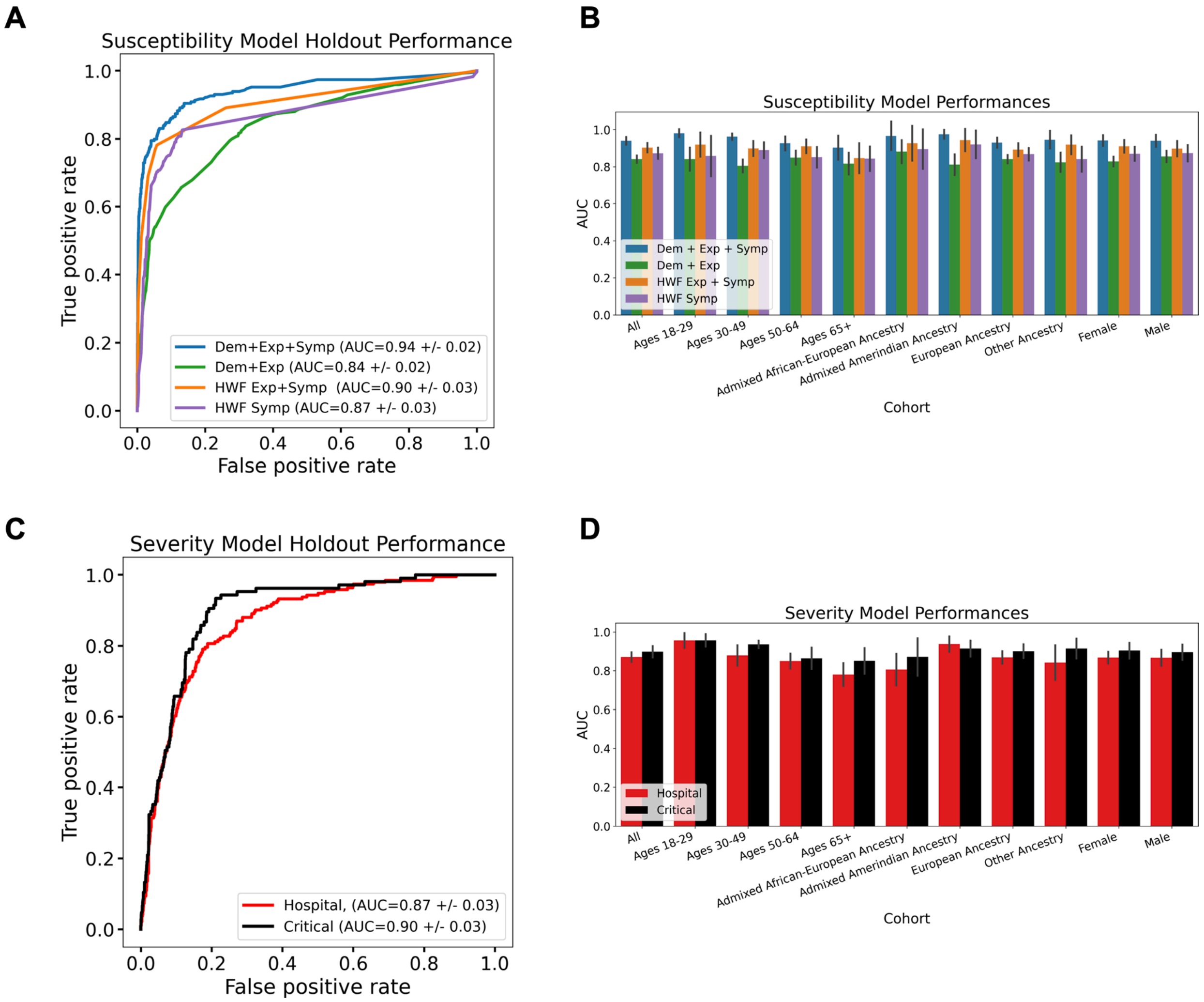

Figure 5 Performance of risk models on independent holdout data. (A) Receiver operating characteristic (ROC) curves for susceptibility models to predict COVID-19 cases among testers reporting a result (positive or negative). (B) Area under the curve (AUC) for the four susceptibility models in (A), stratified by cohort. 'All' represents everyone in (A). (C) ROC curves for severity models to predict either hospitalisation (red) or critical illness progression (black) among COVID-19 cases. (D) Area under the curve (AUC) for the two severity models in (B), stratified by cohort. 'All' represents everyone in (C). Refer to the Methods section as well as online supplemental figure 3 and online supplemental tables 12, 17–21, 24 and 25 for additional model performance data and model risk factor information. Dem+Exp, model based on demographics and exposures only; Dem+Exp+Symp, model based on demographics, exposures and symptoms; HWF Exp+Symp, model called 'How We Feel' based on nearly identical self-reported symptoms and self-reported exposures; HWF Symp, model called 'How We Feel' based on nearly identical self-reported symptoms.

about a plethora of risk factors in the context of COVID-19 susceptibility and severity outcomes. The self-report framework provides fast, low-cost, population-scale data that are particularly valuable in a pandemic, where knowledge is both limited and evolving rapidly based on changing circumstances. Additionally, the broad collection mechanism enables data gathering from many more participants than typically seen in a medical setting, including those with mild or no symptoms, and participants can safely provide data from their homes.

The study highlights exposure burden as the primary risk factor for COVID-19 susceptibility, and the importance of accounting for known exposures when assessing differences in susceptibility to COVID-19. Few studies have measured and explicitly adjusted for known COVID-19 exposures at this scale.[21] Importantly, we found elevated susceptibility risk in men after adjusting for age and known exposures, and unlike most of the risk factors we evaluated, the adjusted odds were not attenuated compared with the unadjusted odds. This finding is distinct from previous findings on elevated severity risk in men.[4 7 11] This result could be due to differences between men and women in behaviours, unknown exposures, biology, genetics,[4–6] or other risk factors not measured within this dataset and should be investigated in future studies.

Another major contribution of this study is the use of self-reported data for the development of novel risk models for predicting an individual's COVID-19 susceptibility and severity risk. The risk models presented here perform comparably or better than similar and more complex models reported previously.[17–19 34 35] Although some previously reported risk models have been assessed in different age or sex cohorts,[17–19] we are not aware of any that have been assessed across genetic ancestry cohorts.[7 17–19 34 35] To ensure model fairness, it is important to assess risk model performance parity (or lack thereof) on known subgroups in the cohort. The parity in performance across genetic ancestry cohorts highlights the potential utility and generalisability of the models to broader populations.[18 19 32]

Limitations

We note that there are some inherent limitations of self-reported data for studying COVID-19 risk factors. The most severe cases, especially those resulting in mortality, were not sampled. As a result, many of the risk factor effect estimates may be underestimated. Additionally, the AncestryDNA cohort is self-selected, slightly older, more European and more female than the broader US population. Another potential issue is that those who reported a negative test may have underestimated their exposures and symptoms relative to those who tested positive, leading to upwardly biased exposure effect estimates. Finally, misclassification of COVID-19 positive status is likely given the uneven availability of tests over the time period surveyed, potentially leading to susceptibility effect estimates that are biased toward the null. However, the fact that most of the associations observed in this study were similar to those previously reported in the literature and the fact that risk model performance remained high when data were stratified by age, sex and genetic ancestry lend confidence to our findings in spite of limitations.

CONCLUSION

The COVID-19 pandemic has exacted a historical toll on healthcare systems and global economies and continues to evolve based on changes in human behaviour, public health guidelines and societal factors. This study demonstrates the power of self-reported data in a large cohort to rapidly elucidate more details about COVID-19 risk factors and help point the way to minimising disease burden.

Acknowledgements We thank our AncestryDNA customers who made this study possible by contributing information about their experience with COVID-19 through our survey. Without them, this work would not be possible. We would like to thank Zach Bass, Robert Dowling, Disha Akarte, Swapnil Sneham, Sean Enright and the entire Cyborg team for their tireless work in the release and continued support of the COVID-19 survey.

Collaborators AncestryDNA Science Team: Yambazi Banda, Ke Bi, Robert Burton, Marjan Champine, Ross Curtis, Abby Drokhlyansky, Ashley Elrick, Cat Foo, Michael Gaddis, Jialiang Gu, Shannon Hateley, Heather Harris, Shea King, Christine Maldonado, Evan McCartney-Melstad, Alexandra McFarland, Patty Miller, Luong Nguyen, Keith Noto, Jingwen Pei, Jenna Petersen, Scott Pew, Chodon Sass, Josh Schraiber, Alisa Sedghifar, Andrey Smelter, Sarah South, Barry Starr, Cecily Vaughn, Yong Wang.

Contributors SCK, SRM and BR contributed equally to the manuscript and wrote the first draft of the paper. ARG provided direct project guidance and led the COVID-19 research teams. SCK, SRM and BR performed the association analyses. SCK developed and assessed the risk models. MVC and KAR designed the COVID-19 survey questionnaire. SCK and DSP supported the dataset creation. GHLR supported the phenotype definitions. MVC and NDB built the demographic tables. SCK, SRM, BR, MVC, GHLR, ARG and KAR helped with additional analyses and interpretation. MZ, DSP, DT, KD, MP, HG and AKHB helped with the EGA dataset. The AncestryDNA Science Team contributed to additional work, allowing for the completion of the COVID-19 research and manuscript. KAR, ELH and CAB provided additional project guidance. KAR is the guarantor. All authors contributed to the final manuscript.

Funding All work was supported and funded by Ancestry.com, a privately owned corporation.

Competing interests Authors affiliated with AncestryDNA may have equity in Ancestry.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval All data for this research project were from subjects who have provided informed consent to participate in AncestryDNA's Human Diversity Project, as reviewed and approved by our external institutional review board (IRB), Advarra (formerly Quorum, IRB approval number: Pro00034516). Advarra operates under ethical principles underlying the involvement of human subjects in research, including the Declaration of Helsinki. All data were de-identified prior to use.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Data are available upon reasonable request. A dataset (EGAC00001001762) is available to qualified scientists through the European Genome-phenome Archive (EGA). The EGA dataset includes the risk factors and outcomes studied here. The EGA dataset is de-identified and comprises ~15 000 individuals who tested for COVID-19, including more than 3000 individuals who tested positive, many of whom are in this study. The EGA cohort is sufficient to nominally replicate the vast majority of susceptibility and severity associations from this study. Risk models trained within the EGA cohort achieve comparable discriminative performance to the models presented here when evaluated in an independent holdout dataset (online supplemental figure 5).

ORCID iD

Kristin A Rand http://orcid.org/0000-0002-1739-3520

REFERENCES

1 World Health Organization. Coronavirus Disease (COVID-19)– Weekly Epidemiological Update, 2022. Available: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19-25-january-2022 [Accessed 27 Jan 2022].

2 U.S. centers for disease control and prevention (CDC). coronavirus disease 2019 (COVID-19): CDC COVID data Tracker, 2021. Available: https://covid.cdc.gov/covid-data-tracker/#cases_casesper100klast7days [Accessed 27 Jan 2022].

3 Davies NG, Klepac P, Liu Y, *et al*. Age-Dependent effects in the transmission and control of COVID-19 epidemics. *Nat Med* 2020;26:1205–11.

4 Takahashi T, Ellingson MK, Wong P, *et al*. Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature* 2020;588:315–20.

5 Roberts GHL, Partha R, Rhead B, *et al*. Expanded COVID-19 phenotype definitions reveal distinct patterns of genetic association and protective effects. *Nat Genet* 2022;54:374–81.

6 Zhao Y, Zhao Z, Wang Y, *et al*. Single-Cell RNA expression profiling of ACE2, the receptor of SARS-CoV-2. *Am J Respir Crit Care Med* 2020;202:756–9.
7 Williamson EJ, Walker AJ, Bhaskaran K, *et al*. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020;584:430–6.
8 Liu Y, Mao B, Liang S, *et al*. Association between age and clinical characteristics and outcomes of COVID-19. *Eur Respir J* 2020;55:2001112.
9 Li X, Xu S, Yu M, *et al*. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *J Allergy Clin Immunol* 2020;146:110–8.
10 Clark A, Jit M, Warren-Gash C, *et al*. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. *Lancet Glob Health* 2020;8:e1003–17.
11 Gebhard C, Regitz-Zagrosek V, Neuhauser HK, *et al*. Impact of sex and gender on COVID-19 outcomes in Europe. *Biol Sex Differ* 2020;11:29.
12 Grasselli G, Zangrillo A, Zanella A, *et al*. Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy region, Italy. *JAMA* 2020;323:1574.
13 , Ellinghaus D, Degenhardt F, *et al*, Severe Covid-19 GWAS Group. Genomewide association study of severe Covid-19 with respiratory failure. *N Engl J Med* 2020;383:NEJMoa2020283.
14 Richardson S, Hirsch JS, Narasimhan M, *et al*. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the new York City area. *JAMA* 2020;323:323.
15 Guan W-J, Liang W-H, Zhao Y, *et al*. Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis. *Eur Respir J* 2020;55:2000547.
16 Zhou F, Yu T, Du R, *et al*. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020;395:1054–62.
17 Wynants L, Van Calster B, Collins GS, *et al*. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020;369:m1328.
18 Menni C, Valdes AM, Freidin MB, *et al*. Real-Time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med* 2020;26:1037–40.
19 Allen WE, Altae-Tran H, Briggs J, *et al*. Population-Scale longitudinal mapping of COVID-19 symptoms, behaviour and testing. *Nat Hum Behav* 2020;4:972–82.
20 Martin LM, Leff M, Calonge N, Garrett C, *et al*. Validation of self-reported chronic conditions and health services in a managed care population. *Am J Prev Med* 2000;18:215–8.
21 Nguyen LH, Drew DA, Graham MS, *et al*. Risk of COVID-19 among front-line health-care workers and the general community: a prospective cohort study. *Lancet Public Health* 2020;5:e475–83.
22 Han E, Carbonetto P, Curtis RE, *et al*. Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nat Commun* 2017;8:14238.
23 Ball CA. Ethnicity estimate 2019 white paper, 2019. Available: https://www.ancestrycdn.com/dna/static/pdf/whitepapers/EV2019_white_paper_2.pdf [Accessed 1 Aug 2020].
24 Centers for Disease Control and Prevention. Defining adult overweight and obesity. Available: https://www.cdc.gov/obesity/adult/defining.html [Accessed 1 Aug 2020].
25 Hastie T. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York:Springer, 2009.
26 Stokes EK, Zambrano LD, Anderson KN, *et al*. Coronavirus Disease 2019 Case Surveillance - United States, January 22-May 30, 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:759–65.
27 Rosenberg A, Keene DE, Schlesinger P, *et al*. COVID-19 and hidden housing vulnerabilities: implications for health equity, new Haven, Connecticut. *AIDS Behav* 2020;24:2007–8.
28 Price-Haywood EG, Burton J, Fort D, *et al*. Hospitalization and mortality among black patients and white patients with Covid-19. *N Engl J Med* 2020;382:2534–43.
29 Millett GA, Jones AT, Benkeser D, *et al*. Assessing differential impacts of COVID-19 on black communities. *Ann Epidemiol* 2020;47:37–44.
30 Webb Hooper M, Nápoles AM, Pérez-Stable EJ. COVID-19 and racial/ethnic disparities. *JAMA* 2020;323:2466.
31 Bosman J, Mervosch S. As virus surges, younger people account for 'disturbing' number of cases. N. Y. Times, 2020. Available: https://www.nytimes.com/2020/06/25/us/coronavirus-cases-young-people.html [Accessed 1 August 2020].
32 Chow EJ, Schwartz NG, Tobolowsky FA, *et al*. Symptom screening at illness onset of health care personnel with SARS-CoV-2 infection in King County, Washington. *JAMA* 2020;323:323.
33 Shi L, Wang Y, Wang Y, *et al*. Dyspnea rather than fever is a risk factor for predicting mortality in patients with COVID-19. *J Infect* 2020;81:647–79.
34 Jehi L, Ji X, Milinovich A, *et al*. Individualizing risk prediction for positive coronavirus disease 2019 testing: results from 11,672 patients. *Chest* 2020;158:1364-1375.
35 Jehi L, Ji X, Milinovich A, *et al*. Development and validation of a model for individualized prediction of hospitalization risk in 4,536 patients with COVID-19. *PLoS One* 2020;15:e0237419.
36 Dietz W, Santos-Burgoa C. Obesity and its implications for COVID-19 mortality. *Obesity* 2020;28:1005.