



# Identifying direct risk factors in UK Biobank via simultaneous Bayesian-frequentist model-averaged hypothesis testing using Doublethink

Nicolas Arning<sup>a</sup>, Helen R. Fryer<sup>a</sup> , and Daniel J. Wilson<sup>a,b,1</sup>

Edited by David Weitz, Harvard University, Cambridge, MA; received June 2, 2025; accepted November 9, 2025

Big data approaches to discovering nongenetic risk factors have lagged behind genome-wide association studies that routinely uncover novel genetic risk factors for diverse diseases. Instead, epidemiology typically focuses on candidate risk factors. Since modern biobanks contain thousands of potential risk factors, candidate approaches may introduce bias, inadequately control for multiple testing, and overlook important signals. Doublethink, a model-averaged hypothesis testing approach, offers a solution that simultaneously controls the Bayesian false discovery rate (FDR) and frequentist familywise error rate (FWER) while accounting for uncertainty in variable selection. Here, we investigate direct risk factors for COVID-19 hospitalization from among 1,912 variables in 201,917 UK Biobank participants by implementing a Doublethink-based exposome-wide association study using Markov Chain Monte Carlo. Focusing on the 2020 outbreak, we find nine individual variables and seven groups of variables exposome-wide significant at 9% FDR and 0.05% FWER. We identify significant direct effects among relatively overlooked risk factors including aging, dementia, and prior infection, which we evaluate in relation to studies of other populations. We detect significant direct effects among some commonly reported risk factors like age, sex, and obesity, but not others like cardiovascular disease. The effects of hypertension, depression, and diabetes appeared to be mediated via general comorbidity. Doublethink produces interchangeable posterior odds and *P*-values for individual variables and arbitrary groups, facilitating flexible and powerful post hoc hypothesis testing. We discuss the potential for impact and limitations of joint Bayesian-frequentist hypothesis testing, including the benefits of an agnostic exposome-wide approach to discovery.

UK Biobank | COVID-19 hospitalization | exposome-wide association studies | FDR | FWER

The big data era has seen the advent of biobank-scale scans for genetic determinants of diverse health outcomes in cohorts like UK Biobank (1, 2). But similar data-driven identification of nongenetic determinants, termed risk factors, has not become commonplace. Instead, current epidemiology typically reports on candidate risk factors. Studies aim to address the question: What is the total effect of a variable on the outcome? Is it nonzero? For instance, more than 100 published studies have analyzed dozens of candidate risk factors for COVID-19 outcomes in UK Biobank (Dataset S1 and SI Appendix). Synthesizing these findings is difficult because i) Other, more important, risk factors that were not analyzed may exist among the thousands measured; ii) It is unclear how to appropriately limit false positives caused by multiple testing; and iii) The processes of selecting candidate risk factors and deciding to publish are vulnerable to bias. The experience of candidate gene studies, largely superseded by genome-wide association studies (GWAS), raises further questions about strength of evidence and reproducibility in candidate risk factor studies (3–5).

In systematic studies of nongenetic risk factors, known as exposome-wide association studies (ExWAS; 6–10), mediation presents a major added complication (11). Mediation occurs when the total effect of a variable (e.g., age) on an outcome (e.g., COVID-19 severity) is wholly or partially caused through another variable (e.g., prior pneumonia). This conceptually divides the total effect into direct and indirect effects. Mediation is ignored in GWAS because genetic variables are co-inherited at conception; they cannot generally cause one another. So the question in GWAS is effectively: What is the direct effect of a variable on the outcome? Is it nonzero? Artifacts generated by confounding are instead the major concern. In confounding, a variable appears associated with an outcome it does not cause because of upstream variables that cause both. Adjusting for potential confounders avoids these associations (12), but in ExWAS, there is a danger of inadvertently adjusting for downstream mediating variables. This would distort estimates of total effects by excluding some indirect effects. Attention could be restricted to

## Significance

Understanding what causes disease is key to improving its treatment and prevention. Large health studies like UK Biobank measure thousands of possible causes of disease. Traditionally, scientists have tested possible causes (like smoking or exercise) one at a time, in depth. For greater perspective, variables could be tested altogether to find out which have any effect. We recently introduced Doublethink, which combines the advantages of two major statistical approaches to testing. Here, we use Doublethink to test 1,912 possible causes of COVID-19 hospitalization in UK Biobank. We found strong evidence for relatively overlooked causes: aging, dementia, and previous infections. Findings from other health studies support these causes, highlighting the need to reevaluate them and showing how our approach can reveal valuable insights.

Author affiliations: <sup>a</sup>Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, United Kingdom; and <sup>b</sup>Department for Continuing Education, University of Oxford, Oxford OX1 2JA, United Kingdom

Author contributions: D.J.W. designed research; N.A., H.R.F., and D.J.W. performed research; N.A., H.R.F., and D.J.W. contributed new reagents/analytic tools; N.A. and D.J.W. analyzed data; and N.A., H.R.F., and D.J.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2026 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).

<sup>1</sup>To whom correspondence may be addressed. Email: daniel.wilson@bdi.ox.ac.uk.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2514138122/-/DCSupplemental>.

Published January 2, 2026.

direct effects instead but, for nongenetic variables, direct effects can differ in direction and magnitude to total effects, a source of bias known as the [Table 2 fallacy](#) (13). Further pitfalls include reverse causation and collider bias (14).

Widespread correlation between the many variables measured by biobanks compounds these difficulties when applying familiar approaches, even under the modest aim of quantifying only direct effects. In univariable scans, spurious signals of association may arise through uncontrolled confounding or collider bias (14). Two-step approaches that use methods like stepwise regression, LASSO (15) or elastic net (16) to analyze the same data twice—first for model selection and then for parameter estimation or hypothesis testing—are susceptible to bias, uncalibrated confidence intervals, and inflated false positive rates (17), unless remedial steps are taken (18).

Nevertheless, the demand for GWAS-inspired ExWAS presents an opportunity, which has been partly filled by machine learning (19, 20). Machine learning offers a data-driven agnostic approach. A major advantage is its ability to analyze high-dimensional data with minimal intervention, even in the presence of collinearity and widespread correlation between variables. But the question in machine learning is different: What is the contribution of a variable to predicting the outcome? Usually there is no formal test. More importantly, a noncausal variable can be valuable for prediction due to confounding (21). Machine learning is therefore problematic for risk factor identification. Other concerns have been raised with artificial intelligence approaches in healthcare, particularly in terms of often difficult-to-achieve interpretability and equity (22).

Bayesian methodology offers a solution to the question of identifying direct effects in biobank-scale data while controlling for confounding (23). An important advantage is the ability to account for uncertainty in model choice by averaging over the inclusion or exclusion of other variables when estimating or testing the direct effect of each variable. This uncertainty can strongly influence conclusions. The question is therefore: What is the explanatory value of each variable, over and above all the other variables? Is it nonzero? With a careful approach to feature engineering to mitigate issues around mediation, reverse causality, and collider bias; with an assumption of no unmeasured confounders; and with independent replication of discoveries, Bayesian model averaging (BMA) offers a powerful approach. But Bayesian approaches are seldom used in current epidemiology: none of 127 published studies of risk factors for COVID-19 outcomes in UK Biobank was Bayesian ([Dataset S1](#)). This might be explained by several issues, including lack of familiarity among researchers, high computational requirements, and difficulties specifying prior distributions (24). Many practitioners worry about the role of the prior in Bayesian hypothesis testing, which can lead different researchers to different conclusions from the same data (25).

Doublethink (26) is a new approach that aims to address these concerns by facilitating joint Bayesian–frequentist model-averaged hypothesis testing while simultaneously controlling the Bayesian false discovery rate (FDR) and the frequentist familywise error rate (FWER). By implementing Doublethink using a Markov Chain Monte Carlo (MCMC) algorithm, we are able to test for direct risk factors among thousands of individual variables, and arbitrary groups of those variables, while accounting for uncertainty in variable selection. We apply Doublethink to investigate direct risk factors for COVID-19 hospitalization in UK Biobank among 1,912 variables in 201,917 participants, and we compare our results to the literature. Our framework provides a highly capable model-averaging approach that can be applied to the systematic evaluation of direct risk factors in biobank-scale resources.

## Theory

We consider a general regression setting in which there are  $n$  observed outcomes  $y_1 \dots y_n$  and  $\nu$  variables (features) with regression coefficients  $\beta_1 \dots \beta_\nu$ . The aim is to identify which variables directly influence the outcome, i.e. which of the regression coefficients are nonzero. In total, there are  $2^\nu$  hypotheses,  $\omega_\nu$ , which we index with vector  $\nu$ . The  $j$ th element of  $\nu$  indicates whether we are testing that variable  $j$  is zero ( $\nu_j = 0$ ;  $\beta_j = 0$ ) or not testing it ( $\nu_j = 1$ ;  $\beta_j = 0$  or  $\beta_j \neq 0$ ).

In parallel, we define  $2^\nu$  models with nonoverlapping parameter spaces  $\mathcal{O}_s$ , indexed by vector  $s$ , the  $j$ th element of which indicates whether variable  $j$  is zero ( $s_j = 0$ ;  $\beta_j = 0$ ) or nonzero ( $s_j = 1$ ;  $\beta_j \neq 0$ ). Each null hypothesis  $\omega_\nu$  is compatible with one or more models  $\mathcal{O}_s$ , indexed by the set

$$\mathcal{O}_\nu = \{s: s_j = 0 \text{ for all } \nu_j = 0\},$$

and incompatible with all the other models, indexed by the complementary set  $\mathcal{A}_\nu$ . In the Bayesian setting, we reject null hypothesis  $\omega_\nu$  if the posterior odds of models in  $\mathcal{A}_\nu$  versus  $\mathcal{O}_\nu$ ,

$$PO_{\mathcal{A}_\nu:\mathcal{O}_\nu} = \frac{\sum_{s \in \mathcal{A}_\nu} PO_s}{\sum_{s \in \mathcal{O}_\nu} PO_s}, \quad [1]$$

exceed some threshold  $\tau$ . Here,  $PO_s$  represents the posterior odds of model  $s$  versus the grand null model  $\mathbf{0}$ . The Bayesian FDR, both local and global (27), is then controlled at or below  $1/(1 + \tau)$ , contingent on the prior.

Fryer, Arning, and Wilson (26) showed that Bayesian hypothesis tests, like the above, inherently control the FWER in the strong sense because they are closed testing procedures (28). The FWER can be quantified, subject to further assumptions. Johnson (29, 30) developed an approach in which the posterior odds of model  $s$  versus model  $\mathbf{0}$  can be approximated as

$$PO_s \approx \left( \mu \frac{b}{n+b} \right)^{|s|} R_s^{n/(n+b)}, \quad [2]$$

where  $R_s$  represents the ratio of maximized likelihoods under model  $s$  versus model  $\mathbf{0}$ ,  $|s|$  gives the difference in the number of parameters to be estimated under model  $s$  versus model  $\mathbf{0}$ ,  $\mu$  represents the prior odds that a regression coefficient is nonzero,  $n$  is the sample size, and  $b$  represents the precision of the prior on the nonzero regression coefficients. The approximation is based on assumptions including a large sample size, independent observations, and the following prior:

$$\theta_s \stackrel{d}{=} \text{Normal}(0, b^{-1} \mathbf{I}_s^{-1}), \quad [3]$$

$$s_j \stackrel{d}{=} \text{Bernoulli}\left(\frac{\mu}{1+\mu}, j = 1 \dots \nu\right). \quad [4]$$

Here,  $\theta_s$  represents the unconstrained parameters in model  $s$  (the  $\beta_j$  for which  $s_j = 1$ , and any nuisance parameters), and  $\mathbf{I}_s$  is the per-observation Fisher information matrix for model  $s$ , evaluated at  $\theta_s = \mathbf{0}$ . Fisher's information matrix has been used widely in the definition of reference priors (e.g., 31, 32), and to generate concordance between Bayesian and frequentist point and interval estimates (see [Table 1](#) of 26). Johnson's approach converges on the Bayesian information criterion (BIC), which has been shown to reasonably approximate a wide range of posterior odds when  $n$  is large (33, 34). The strength of Johnson's approach is the ability, for a pair of nested models, to interconvert posterior odds and  $P$ -values based on the likelihood ratio test (35).

Doublethink (26) extends this approach to the multiple testing setting, in which there is model uncertainty. Using the theory of heavy-tailed random variables (36–39) and asymptotic likelihood theory (e.g., 40), we showed that rejecting all null hypotheses  $\omega_\nu$  for which  $\text{PO}_{A_\nu, \mathcal{O}_\nu} > \tau$  controls the FWER in the strong sense at or below level

$$\alpha \sim \Pr \left( \chi_1^2 > 2 \log \frac{\tau}{\nu \mu \left( \frac{b}{n+b} \right)^{1/2}} \right) \text{ as } n \rightarrow \infty. \quad [5]$$

The Bayesian procedure is equivalent to rejecting the null hypothesis  $\omega_\nu$  when an asymptotic  $P$ -value, adjusted for multiple testing,

$$p_{A_\nu, \mathcal{O}_\nu}^* \sim \Pr \left( \chi_1^2 > 2 \log \frac{\text{PO}_{A_\nu, \mathcal{O}_\nu}}{\nu \mu \left( \frac{b}{n+b} \right)^{1/2}} \right) \text{ as } n \rightarrow \infty, \quad [6]$$

is smaller than threshold  $\alpha$ . In general, the convergence in these asymptotic results is pointwise.

An equivalent interpretation of the results is that the test statistic  $2 \log \text{PO}_{A_\nu, \mathcal{O}_\nu} / \left( \nu \mu \sqrt{b/(n+b)} \right)$ , which follows a chi-squared distribution with one degree of freedom when large, represents a model-averaged deviance. Doublethink  $P$ -values cannot be trivially rescaled by the prior parameters  $\mu$  and  $b$  because i) the null distribution of the model-averaged deviance does not depend on them and ii) the realized value depends on them only through weights. Therefore,  $\mu$  and  $b$  influence the power of the test, but not its large-sample theoretical distribution under the null hypothesis. This makes model-averaged hypothesis testing a workable frequentist method by facilitating a prior-agnostic approach to quantifying Bayesian significance thresholds in terms of frequentist FWER, for large samples.

In simulations based on real data with strong correlation structure, we compared Doublethink model averaging to two-step LASSO, elastic net, and stepwise regression approaches in which the same data were used first for model selection, and then reused for parameter estimation and hypothesis testing (26). Doublethink parameter estimates exhibited smaller variances and standard errors were better calibrated than other methods, with close-to-optimal performance across different values of  $\mu$  and  $b$ . All methods suffered inflated FWER due to pervasive correlation, which was attributed to tests that split up groups of highly correlated variables. Inflation was mitigated for all methods except stepwise regression by grouping correlated variables (see Methods below). In terms of power, Doublethink resembled LASSO and outperformed elastic net for tests of individual variables. Doublethink outperformed both for tests of grouped variables. We concluded that Doublethink performed best overall for inference, but computation took 50 times longer. For further details on theory, simulations, limitations of the approach such as inflation of the tests due to highly correlated variables, and mitigations such as testing groups of highly correlated variables, see (26).

## Methods

In this study, we developed a Monte Carlo Markov Chain algorithm (41) in R (42) and Python (43) that implements the Doublethink approach for thousands of variables, in order to identify direct risk factors for COVID-19 hospitalization in UK Biobank. We followed the COVID-19 Host Genetics Initiative definition of COVID-19 hospitalization, as applied to UK Biobank.

**Outcomes.** Cases were identified from Public Health England's Second Generation Surveillance System (SGSS), the National Health Service's Hospital Episode Statistics (HES) and the National Health Service's death registry between January and December 2020 as PCR positive for SARS-CoV-2 in SGSS, and hospitalized with International Classification of Diseases, Tenth Revision (ICD-10) diagnosis code U07.1 or U07.2 in HES. Participants not identified as cases were considered controls. We excluded participants who died before 2020, non-England residents determined by assessment center, and those who withdrew before the analysis. The total number of controls was down-sampled to 200,000 to speed computation. The total number of cases was 1,917.

**Variables.** We considered data fields approved for UK Biobank project 53100 'Microbiology, disease and genetics', across the categories Population characteristics, Assessment center, Biological samples, Online follow-up, Additional exposures, and Health-related outcomes. We excluded Compound, Date, Text and Time variables, and variables concerning genetics and sampling processes. For repeated measures, we took the first instance. We excluded factors exceeding 50 levels, except self-reported illnesses, and variables missing in more than 15% of participants. Special values as defined by UK Biobank, including negative factor levels, were treated as missing. The mean and interquartile range of missingness among continuous and integer covariates were 3.9% and (1.2%, 5.9%); we imputed them by taking the mean of nonmissing values. Missing factor levels were treated as a separate level and excluded. We created binary variables for all levels of every factor observed with frequency above 0.2%. We created a binary variable for every ICD-10 code with frequency above 0.2% recorded before 2020 in HES. Overall, we analyzed 184 covariates, binary variables encoding 865 levels across 193 factors, and 863 ICD-10 admission codes, a total of 1,912 variables (Dataset S2).

**Model.** We fitted the data via a logistic regression model implemented in R using the glm function, assuming an additive linear predictor with an intercept term. We assumed the prior odds of variable inclusion were  $\mu = 0.0053$ , independently for the  $\nu = 1,912$  variables, implying a prior expectation of 10 variables in the model. Assuming fixed prior odds of inclusion for all variables meant that we treated baseline risk factors like age, sex, and socioeconomic status, the same as modifiable exposures like environmental or behavioral factors, allowing the model to include or exclude any of them based on the data. We refer to all nonoutcome variables as 'exposures' (10). We assumed a unit information prior ( $h = 1$ ) for the regression coefficients (34). We disallowed the inclusion of collinear variables by defining a zero likelihood.

**Implementation.** In contrast to the Gibbs sampling approach of (44), we implemented a Metropolis-Hasting Markov Chain Monte Carlo (MCMC) sampler over the variable inclusion vector  $\mathbf{s}$ . We ran 100 chains in parallel with 25,000 iterations of burn-in and 75,000 iterations of sampling each. The average run-time per chain was 35 h. Chains were initialized using a furthest neighbor algorithm to avoid including correlated variables. For initialization, we clustered variables into 200 groups with the scikit-learn-extra KMedoids algorithm, using rank correlation distance. Each chain was initialized with the medoid of one group, before adding nine more variables iteratively from the next-least correlated variables. Three Metropolis Hastings moves were implemented that respectively added, removed, or swapped pairs of variables with relative proposal probabilities 9:9:2. Variables were swapped preferentially for those with high squared correlation  $r^2$ . We simulated regression coefficients directly from conditional Normal distributions by postprocessing the MCMC iterations. We calculated posterior odds and parameter estimates by combining chains, computing Monte Carlo standard errors across independent chains.

**Grouping Variables.** We were able to perform valid arbitrary variable grouping while controlling the FDR and FWER, which was useful since correlated variables reduce one-another's individual posterior inclusion probabilities. Moreover, tests of null hypotheses involving some but not all members of a cluster of highly correlated variables are liable to inflation (26). We grouped variables in two ways: predefined or post hoc. We computed the posterior odds of including one or more of the variables in each group. We constructed predefined groups hierarchically using UPGMA (45), with distances defined as one minus  $r^2$  between variables. Among these 1,911 nested groupings, we identified 27 broad-scale variable clusters, defined as groups containing at least 20 variables with mean  $r^2 \geq 0.02$ . When reporting the significance of predefined groups, we always report the smallest significant subgroup, because any group that contains a significant subgroup is significant by definition.

**Table 1. Doublethink allows the interconversion of model-averaged posterior odds and *P*-values for groups of variables, predefined here using variable correlation**

	Group <i>PP</i> (%)	Group $-\log_{10} P^*$	Variable	<i>PP</i> (%)	$-\log_{10} P^*$	Direct effect when included	Standard error when included
1	<b>100.0</b>	<b>&gt;5.95</b>	34 Year of birth (years)	40.8	2.05	-0.03	0.00
			21003 Age when attended assessment center (years)	30.0	1.82	0.03	0.10
			21022 Age at recruitment (years)	29.2	1.80	0.03	0.10
2	<b>100.0</b>	<b>5.95</b>	31 Sex: 0: Female	50.1	2.24	-0.49	0.11
			31 Sex: 1: Male	49.9	2.23	0.49	0.11
-	<b>100.0</b>	<b>5.78</b>	41214 Carer support indicators: 1: Yes	<b>100.0</b>	<b>5.78</b>	0.56	0.08
3	<b>99.9</b>	<b>5.52</b>	48 Waist circumference (cm)	89.0	3.22	0.02	0.00
			21001 Body mass index (BMI) (Kg/m <sup>2</sup> )	5.5	-	0.04	0.01
			23104 Body mass index (BMI) (Kg/m <sup>2</sup> )	5.0	-	0.04	0.01
			23100 Whole body fat mass (Kg)	0.4	-	0.02	0.00
			21002 Weight (Kg)	0.0	-	0.01	0.00
			23120 Arm fat mass (right) (Kg)	0.0	-	0.06	0.06
			23124 Arm fat mass (left) (Kg)	0.0	-	0.03	0.05
			23128 Trunk fat mass (Kg)	0.0	-	0.01	0.01
			49 Hip circumference (cm)	0.0	-	0.00	0.00
			23098 Weight (Kg)	0.0	-	0.00	0.00
-	<b>99.9</b>	<b>5.44</b>	Z86.4 Personal history of psychoactive substance abuse	<b>99.9</b>	<b>5.44</b>	0.37	0.06
-	<b>99.7</b>	<b>5.00</b>	F03 Unspecified dementia	<b>99.7</b>	<b>5.00</b>	0.94	0.15
-	<b>99.7</b>	<b>4.89</b>	137 Number of treatments/medications taken	<b>99.7</b>	<b>4.89</b>	0.06	0.01
-	<b>99.6</b>	<b>4.82</b>	J22 Unspecified acute lower respiratory infection	<b>99.6</b>	<b>4.82</b>	0.51	0.09
4	<b>99.5</b>	<b>4.71</b>	Z50.1 Other physical therapy	79.9	2.89	0.53	0.09
			Z50.7 Occupational therapy and vocational rehabilitation, not elsewhere classified	19.6	-	0.61	0.11
-	<b>99.2</b>	<b>4.47</b>	R29.6 Tendency to fall, not elsewhere classified	<b>99.2</b>	<b>4.47</b>	0.63	0.11
5	<b>99.0</b>	<b>4.38</b>	26413 Health score (England)	77.7	2.83	0.26	0.03
			26412 Employment score (England)	10.9	-	2.79	0.37
			26410 Index of Multiple Deprivation (England)	9.9	-	0.01	0.00
			26414 Education score (England)	0.3	-	0.01	0.00
			26411 Income score (England)	0.3	-	1.70	0.34
-	<b>98.2</b>	<b>4.10</b>	41218 History of psychiatric care on admission: 8: Not applicable	<b>98.2</b>	<b>4.10</b>	0.49	0.09
-	<b>96.7</b>	<b>3.82</b>	6138 Qualifications: 3: O levels/GCSEs or equivalent	<b>96.7</b>	<b>3.82</b>	-0.29	0.05
-	<b>96.5</b>	<b>3.80</b>	6138 Qualifications: 1: College or University degree	<b>96.5</b>	<b>3.80</b>	-0.34	0.06
-	84.1	3.02	J18.1 Lobar pneumonia, unspecified	84.1	3.02	0.49	0.09
-	70.2	2.64	21000 Ethnic background: 1001: British	70.2	2.64	-0.40	0.08
-	66.1	2.55	K59.0 Constipation	66.1	2.55	0.40	0.08
-	58.4	2.40	N39.0 Urinary tract infection, site not specified	58.4	2.40	0.40	0.08
-	31.5	1.86	L97 Ulcer of lower limb, not elsewhere classified	31.5	1.86	0.69	0.14
-	28.1	1.77	3063 Forced expiratory volume in 1-s (FEV1) (liters)	28.1	1.77	-0.19	0.04
-	25.5	1.71	N18.9 Chronic renal failure, unspecified	25.5	1.71	0.48	0.10

The smallest groups significant at *PP* ≥ 91% are shown, alongside details of constituent variables. The most significant individual, ungrouped, variables are also shown. *PP*: posterior probability. *P*\*: adjusted *P*-value. Dashes (-) indicate *P*\* > 0.02.

(26). We used ChatGPT-4o (46) to help manually label these 27 variable clusters. We identified a dozen nonoverlapping post hoc groups using the scikit-learn OPTICS algorithm (47) with  $\xi = 0.05$  (the default) and distances defined by their posterior correlation in inclusion probabilities. This grouped the variables with the strongest negative correlations in posterior inclusion probabilities, allowing us to identify groups of more-or-less interchangeable variables.

**P-Value Calculation.** We used the chi-squared distribution to compute adjusted *P*-values using Eq. 6. In the case of orthogonal variables with one degree of freedom, this is conservative for *P* < 0.02; see (26). Since the large sample size assumption implies interest in small significance thresholds, we report any

adjusted *P*-value larger than 0.02 as n.s. (not significant) or “-”. Effectively, this makes Doublethink incompatible with any threshold exceeding  $\alpha = 0.02$ . We reported adjusted *P*-values alongside posterior odds.

**Literature Review.** We reviewed the variables included in published analyses of COVID-19 risk factors in UK Biobank using the query “UK Biobank” (Abstract) and “COVID” (Abstract) in [www.webofscience.com](http://www.webofscience.com) on 19 September 2023. After excluding Review Articles and Editorial Material, this search returned 203 publications. We analyzed a subset of 127 of these papers that quantified the effect of nongenetic risk factors on COVID-19 outcomes; this predominantly excluded papers reporting genetic risk factors, two-sample Mendelian randomization, and

COVID-19 as an exposure for other outcomes (Dataset S1). We manually categorized the variables analyzed by these 127 papers into groups (Dataset S3). We summarized the frequency with which each category of variable was included in the published analysis or abstract.

## Results

We aimed to identify risk factors that directly influenced COVID-19 hospitalization in UK Biobank participants to understand the underlying processes. We used model-averaged hypothesis tests to account for uncertainty in variable selection and deplete for potential confounders. We assumed the relevant variables were measured in UK Biobank. We aimed to limit the impact of collider bias by focusing on exposure variables measured before 2020, and by comparing cases to the rest of the biobank. This compounded the case definition with any selection biases in the sampling process, for example, access to testing, which may affect interpretation (14). We focused on risk factors for hospitalization with COVID-19, because there were more cases than critical illness, and less obvious selection bias than infection, since testing was more widely available in hospitals.

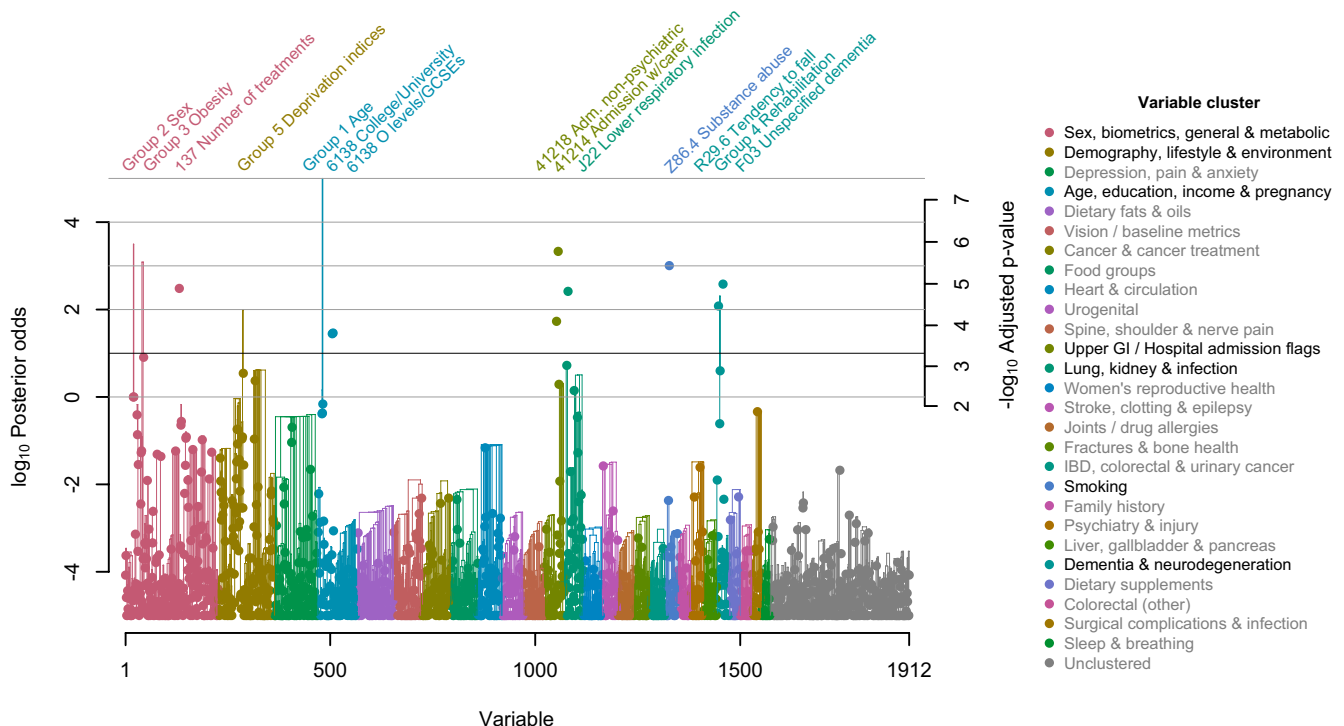
**Doublethink Facilitates Joint Bayesian-Frequentist Model-Averaged Hypothesis Tests.** Figure 1 shows a Manhattan plot displaying the evidence that each of the 1,912 individual variables (points) directly affected the risk of COVID-19 hospitalization in UK Biobank, averaged over uncertainty in the effect of all other variables. Points are plotted against both the  $\log_{10}$  posterior odds (left side) and the  $-\log_{10}$  adjusted  $P$ -value from Eq. 6 (Right side). This interconversion allows a Bayesian or frequentist approach to evaluating the strength of evidence.

Comparison of the two vertical axis scales shows that in the Doublethink model, the model-averaged posterior odds and adjusted  $P$ -values are approximately linearly related, for small

$P$ -values. Significant variables are identified by applying a threshold to either the posterior odds or the adjusted  $P$ -value; this simultaneously controls the FDR (subject to the assumed prior) and the FWER (subject to the asymptotic approximation). For example, a Bayesian threshold of  $\tau = 10$  controls the FDR at  $1/(1 + \tau) = 0.091$  and the FWER at  $\alpha = 10^{-3.3} = 0.00047$ . The latter is much smaller than the conventional threshold of 0.05 because of the large sample size.

At a significance threshold of  $\tau = 10$  and  $\alpha = 10^{-3.3}$ , nine variables were identified as individually exposome-wide significant (Table 1). The interpretation is that significant variables, such as the ICD-10 codes 'F03 Unspecified dementia' and 'J22 Unspecified acute lower respiratory infection', directly affect risk of COVID-19 hospitalization, even after controlling for the effects of all other measured variables. This differs from the common practice of testing the significance of a variable in the context of a single model that controls for a limited set of other variables. Model averaging is important in biobank-scale data where correlation between variables is pervasive, and no single model has high posterior probability.

For several significant variables, the interpretation that they directly affect risk of COVID-19 would seem too literal, such as '137 Number of treatments/medications taken', which is based on the recruitment interview, '41214 Carer support indicators: 1: Yes', which indicates a hospital record of past carer support, 'R29.6 Tendency to fall, not elsewhere classified', which indicates a history and future risk of falls, and 'Z86.4 Personal history of psychoactive substance abuse', which indicates a hospital record of past alcohol, tobacco or drug use. More plausibly, these variables represent or aggregate one or more (perhaps unmeasured) variables that directly affect risk of COVID-19, like aspects of general health or behavior. The estimated direct effect of these proxies was to increase the risk of COVID-19 hospitalization in all cases (Table 1). In contrast, significant measures of educational



**Fig. 1.** Predefined groups of variables (trees) and individual variables (points) with the strongest evidence of direct effects on the risk of COVID-19 hospitalization in UK Biobank. Evidence was quantified simultaneously by  $\log_{10}$  posterior odds (Left axis) and  $-\log_{10}$  adjusted  $P$ -value (Right axis) using Doublethink. Points and lines are colored by cluster (see key, Right). Tree branches show the boost in significance (if any) from testing groups of variables. Individual variables and groups significant at  $\log_{10}$  posterior odds  $\geq 1$  are labeled. Groups containing significant subgroups are always significant and therefore omitted for legibility. Significance was truncated to  $\log_{10}$  posterior odds between  $-5$  and  $5$ . Individual variables are named by abbreviated UK Biobank field ID or, when prefixed by a letter, ICD-10 code. See Table 1 for full names.

attainment, '6138 Qualifications: 3: O levels/GCSEs or equivalent', and '6138 Qualifications: 1: College or University degree', had protective direct effects on risk of COVID-19 hospitalization.

The significance of some variables was, at first glance, unexpectedly low, such as the well-established risk factors '31 Sex: 1: Male' [% posterior probability,  $PP = 49.9$ ;  $P^* = 10^{-2.23}$ ; where posterior odds =  $PP/(1-PP)$ ] and '34 Year of birth (years)' ( $PP = 40.8$ ;  $P^* = 10^{-2.05}$ ; Table 1). This is explained by the inclusion in the data of the other very highly correlated variables '31 Sex: 0: Female', '21003 Age when attended assessment centre (years)' and '21022 Age at recruitment (years)'. Including variables that are correlated, whether strongly or weakly, often dilutes the significance of individual variables when testing for the existence of a direct effect, over and above all other variables. For age and sex, an obvious solution would be to exclude these correlated variables—it may seem absurd not to have done so. However, it may not be obvious which variables to exclude because correlation is pervasive in biobank-scale data, and the impact of excluding variables on the results is hard to anticipate. An alternative solution is to define groups of correlated variables and test whether one or more members of a group directly affect the outcome. A major strength of Doublethink is that it allows arbitrary groups of variables to be tested in this way, while controlling the FDR and FWER.

**Testing the Significance of Groups of Variables Reveals More Signals.** Predefined groups were defined by hierarchically clustering variables based on pairwise squared correlation coefficients (SI Appendix, Fig. S1 and Dataset S4). Five predefined groups were significant at  $\tau = 10$  and  $\alpha = 10^{-3.3}$ . None of their member variables were individually significant. In Fig. 1, tree branches illustrate the boost in the significance of groups of variables compared to their individual member variables. The groups are numbered for cross-reference with Table 1. Reassuringly, the well-established risk factors age (Group 1;  $PP = 100$ ;  $P^* < 10^{-5.95}$ ), sex (Group 2;  $PP = 100$ ;  $P^* = 10^{-5.95}$ ), obesity (Group 3;  $PP = 99.9$ ;  $P^* = 10^{-5.52}$ ), and indices of multiple deprivation (Group 5;  $PP = 99.0$ ;  $P^* = 10^{-4.38}$ ) were significant despite containing no individually significant member variables. In these examples, testing groups of variables recovered signal that was diluted by the inclusion in the data of highly correlated variables.

Another group was significant, despite containing no individually significant variables, demonstrating the ability of combined tests to detect subtle signals. Group 4 ( $PP = 99.5$ ,  $P^* = 10^{-4.71}$ ) comprised 'Z50.1 Other physical therapy' ( $PP = 79.9$ ,  $P^* = 10^{-2.89}$ ) and 'Z50.7 Occupational therapy and vocational rehabilitation, not elsewhere classified' ( $PP = 19.6$ ,  $P^* > 0.02$ ). These indicators of rehabilitation might represent or aggregate aspects of convalescence less well captured by the other 1,912 variables analysed. The analysis suggests this history of convalescence directly increased the risk of COVID-19 hospitalization, after controlling for all other measured variables.

Testing groups is useful, but predefining them is not necessarily the most effective method of discovering signals, because the groupings might not be relevant to the outcome under investigation. For example, Group 3 (obesity) included six variables which did not contribute to the group's overall significance ( $PP = 0.0$ ,  $P^* > 0.02$ ). Conversely, failure to group relevant variables together can cause signals to be overlooked, as we see next.

**Doublethink Allows Arbitrary Groups to be Tested.** One of the advantages of the Doublethink approach is that it motivates the testing of arbitrary groups of variables without inflating the FWER or FDR through a multiple testing 'fishing expedition'. This is because the thresholds of all possible tests are predefined in the

closed testing procedure (26). Therefore, we were free to search for the most significant groups of variables. To this end, we grouped variables post hoc whose posterior inclusion probabilities ( $PPs$ ) were negatively correlated, because this suggests they 'competed' for inclusion in the model.

Fig. 2 and Table 2 show that post-hoc grouping can reveal significant groups of variables absent from pre-defined groupings. Group I ( $PP = 99.7$ ,  $P^* = 10^{-3.69}$ ) combined the individually non-significant 'K59.0 Constipation' ( $PP = 66.1$ ,  $P^* = 10^{-2.55}$ ) and 'N39.0 Urinary tract infection, site not specified' ( $PP = 58.4$ ,  $P^* = 10^{-2.40}$ ). These variables were assigned to distinct pre-defined variable clusters: Upper GI/Hospital admission flags and Lung, kidney and infection, respectively. However, their post-hoc grouping suggests they were interchangeable as direct risk factors for COVID-19 hospitalization because they appeared in the model together less often than expected. Constipation may increase the risk of urinary tract infection, but it is unclear how a history of these conditions could predispose to COVID-19 hospitalization, demonstrating how post-hoc grouping can surface unexpected signals.

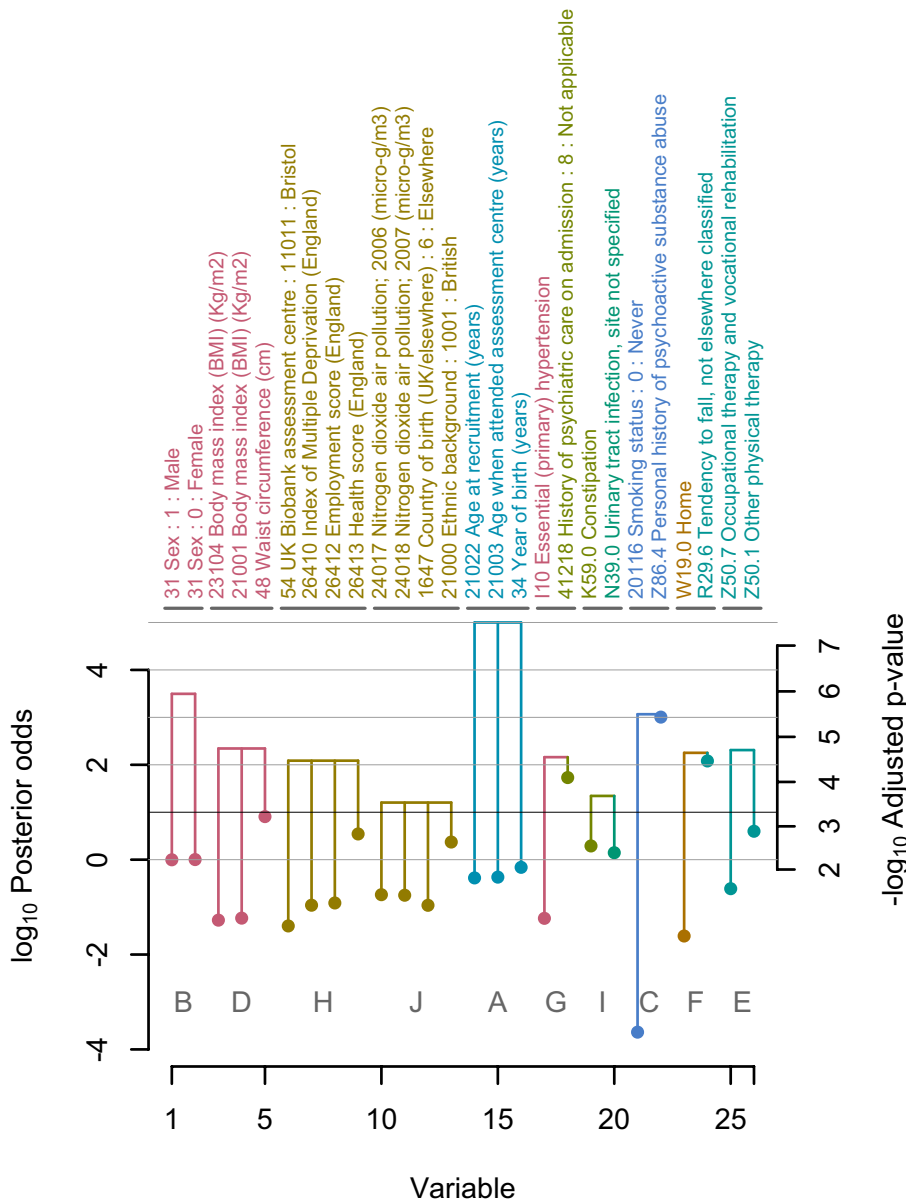
Group J ( $PP = 94.1$ ,  $P^* = 10^{-3.54}$ ) combined the individually non-significant '21000 Ethnic background: 1001: British' ( $PP = 70.2$ ,  $P^* = 10^{-2.64}$ ), with '1647 Country of birth (UK/elsewhere): 6: Elsewhere' ( $PP = 9.8$ ,  $P^* > 0.02$ ), '24017 Nitrogen dioxide air pollution; 2006 (micro-g/m<sup>3</sup>)' ( $PP = 15.5$ ,  $P^* > 0.02$ ) and '24018 Nitrogen dioxide air pollution; 2007 (micro-g/m<sup>3</sup>)' ( $PP = 15.1$ ,  $P^* > 0.02$ ). The nonobvious grouping of self-reported ethnicity and country of birth with air pollution reveals that these variables capture overlapping aspects of risk.

The post-hoc grouping of '41218 History of psychiatric care on admission: 8: Not applicable' with 'I10 Essential (primary) hypertension' was at first glance surprising from the field descriptions (Group G:  $PP = 99.3$ ,  $P^* = 10^{-4.55}$ ). However, the former variable indicates a history of non-psychiatric hospital care. This suggests it may act, in a manner interchangeable with I10, as a proxy for a history of underlying poor physical health. The direct effect of both variables was to increase the risk of COVID-19 hospitalization (Table 2).

Some post-hoc groups coincided with pre-defined groups but dropped non-significant variables that did not contribute to the overall significance of the group. For example, Group D ( $PP = 99.5$ ,  $P^* = 10^{-4.75}$ ) contained only the three most significant obesity metrics of the ten members of Group 3: '48 Waist circumference (cm)' ( $PP = 89.0$ ,  $P^* = 10^{-3.22}$ ), '21001 Body mass index (BMI) (Kg/m<sup>2</sup>)' ( $PP = 5.5$ ,  $P^* > 0.02$ ) and '23104 Body mass index (BMI) (Kg/m<sup>2</sup>)' ( $PP = 5.0$ ,  $P^* > 0.02$ ). Post-hoc grouping can therefore be more parsimonious than prior grouping.

The ability to quantify the evidence for groups of variables offers an alternative to approaches such as preanalysis selection of representative candidate variables among groups of correlated variables. Doublethink permits all and any groups of variables to be tested while controlling the FDR and FWER. This presents possibilities for identifying significant groups, and the identification of these groups may help with the interpretation of the role of the individual variables in the outcome.

**Comparison to the Literature on COVID-19 Outcomes in UK Biobank.** Since early in the COVID-19 pandemic, before the discovery of effective treatments, there were intense research efforts to understand susceptibility to infection, disease, and poor outcomes. Many studies focused on large established cohorts like UK Biobank that could rapidly link to data on SARS-CoV-2 testing (48), COVID-19 hospitalization (49), and mortality (50). Since then, many risk factors have been reported, including diabetes (51–54), asthma (55, 56), and vitamin D (57, 58) as predisposing to worse



**Fig. 2.** Post hoc groups of variables with significant evidence of direct effects on the risk of COVID-19 hospitalization in UK Biobank at  $\log_{10}$  posterior odds  $\geq 1$ . Trees show the boost in significance (if any) from individual member variables (points) to the significance of the whole group. Individual variables are colored by cluster, as in Fig. 1.

outcomes. We compared our results to the literature on COVID-19 in UK Biobank to identify any differences to standard approaches and find unique insights. At the time of analysis, we identified 127 comparable studies through Web of Science. We manually assigned the most common risk factors in published analyses of COVID-19 outcomes to larger categories for comparison to the variables and groups listed in Tables 1 and 2, which we assigned to the same list of categories (Datasets S5 and S6).

Table 3 shows the most common categories of risk factors included in published analyses of COVID-19 outcomes in the 127 UK Biobank studies. Two summaries are shown: the percentage of papers and the percentage of abstracts in which each category of risk factors appeared. Alongside, we show the evidence from our analysis, with values of  $PP < 50\%$  (corresponding to  $P^* > 10^{-2.20}$ ) omitted, since the Bayesian interpretation is that this represents evidence against a direct effect of those risk factors. An important caveat is that many studies focused on the total (direct and indirect) effects of candidate risk factors, which our method does not quantify. Therefore, the discovery of a risk factor absent from the literature is more interesting than the nondiscovery of (a possibly indirect) effect reported in the literature.

Age, Sex, Obesity, Ethnicity, Socioeconomic status (including deprivation indices), and Smoking were included in 66 to 90% of published analyses, but mentioned in just 6-20% of abstracts. Our analysis strongly supported direct effects of all these categories with  $PP \geq 99.2\%$  and  $P \leq 10^{-4.48}$  except ethnicity, which was only significant as part of post hoc Group J, in combination with country of birth and geographic measures of pollution ( $PP = 94.1$ ,  $P^* = 10^{-3.54}$ ). Other reasonably common categories of risk factor for which our analysis found evidence of direct effects included Lung disease, Alcohol intake, General comorbidity, Kidney disease, and Educational attainment. Risk factors in these categories featured in 28 to 46% of published analyses and 4 to 9% of abstracts. Our analysis supported these categories with  $PP \geq 95.7\%$  and  $P^* \leq 10^{-3.69}$ .

Many categories of risk factors that appeared commonly in published analyses received no significant support for direct effects in our analysis. Diabetes and Cardiovascular disease were notable for inclusion in 59 to 63% of published analyses, and 10 to 12% of abstracts. No variables or groups of variables corresponding to these categories received support for direct effects in our analyses ( $PP < 50\%$ ,  $P^* > 10^{-2.20}$ ). Hypertension was included in 54% of published analyses and 11% of abstracts, but evidence for direct effects was

**Table 2. Doublethink allows arbitrary groups of variables to be assessed for significance while simultaneously controlling the FWER and FDR**

	Group <i>PP</i> (%)	Group $-\log_{10} P^*$	Variable	<i>PP</i> (%)	$-\log_{10} P^*$	Direct effect when included	Standard error when included
A	<b>100.0</b>	<b>&gt;5.95</b>	34 Year of birth (years)	40.8	2.05	-0.03	0.00
			21003 Age when attended assessment center (years)	30.0	1.82	0.03	0.10
			21022 Age at recruitment (years)	29.2	1.80	0.03	0.10
B	<b>100.0</b>	<b>5.95</b>	31 Sex: 0: Female	50.1	2.24	-0.49	0.11
			31 Sex: 1: Male	49.9	2.23	0.49	0.11
-	<b>100.0</b>	<b>5.78</b>	41214 Carer support indicators: 1: Yes	<b>100.0</b>	<b>5.78</b>	0.56	0.08
C	<b>99.9</b>	<b>5.50</b>	Z86.4 Personal history of psychoactive substance abuse	<b>99.9</b>	<b>5.44</b>	0.37	0.06
			20116 Smoking status: 0: Never	0.0	-	-0.17	0.07
-	<b>99.7</b>	<b>5.00</b>	F03 Unspecified dementia	<b>99.7</b>	<b>5.00</b>	0.94	0.15
-	<b>99.7</b>	<b>4.89</b>	137 Number of treatments/medications taken	<b>99.7</b>	<b>4.89</b>	0.06	0.01
-	<b>99.6</b>	<b>4.82</b>	J22 Unspecified acute lower respiratory infection	<b>99.6</b>	<b>4.82</b>	0.51	0.09
D	<b>99.5</b>	<b>4.75</b>	48 Waist circumference (cm)	89.0	3.22	0.02	0.00
			21001 Body mass index (BMI) (Kg/m <sup>2</sup> )	5.5	-	0.04	0.01
			23104 Body mass index (BMI) (Kg/m <sup>2</sup> )	5.0	-	0.04	0.01
E	<b>99.5</b>	<b>4.71</b>	Z50.1 Other physical therapy	79.9	2.89	0.53	0.09
			Z50.7 Occupational therapy and vocational rehabilitation, not elsewhere classified	19.6	-	0.61	0.11
F	<b>99.4</b>	<b>4.65</b>	R29.6 Tendency to fall, not elsewhere classified	<b>99.2</b>	<b>4.47</b>	0.63	0.11
			W19.0 Home	2.4	-	0.59	0.15
G	<b>99.3</b>	<b>4.55</b>	41218 History of psychiatric care on admission: 8: Not applicable	<b>98.2</b>	<b>4.10</b>	0.49	0.09
			I10 Essential (primary) hypertension	5.5	-	0.24	0.06
H	<b>99.2</b>	<b>4.48</b>	26413 Health score (England)	77.7	2.83	0.26	0.03
			26412 Employment score (England)	10.9	-	2.79	0.37
			26410 Index of Multiple Deprivation (England)	9.9	-	0.01	0.00
			54 UK Biobank assessment center: 11011: Bristol	3.9	-	-0.47	0.14
-	<b>96.7</b>	<b>3.82</b>	6138 Qualifications: 3: O levels/GCSEs or equivalent	<b>96.7</b>	<b>3.82</b>	-0.29	0.05
-	<b>96.5</b>	<b>3.80</b>	6138 Qualifications: 1: College or University degree	<b>96.5</b>	<b>3.80</b>	-0.34	0.06
I	<b>95.7</b>	<b>3.69</b>	K59.0 Constipation	66.1	2.55	0.40	0.08
			N39.0 Urinary tract infection, site not specified	58.4	2.40	0.40	0.08
J	<b>94.1</b>	<b>3.54</b>	21000 Ethnic background: 1001: British	70.2	2.64	-0.40	0.08
			24017 Nitrogen dioxide air pollution; 2006 (micro-g/m <sup>3</sup> )	15.5	-	0.01	0.00
			24018 Nitrogen dioxide air pollution; 2007 (micro-g/m <sup>3</sup> )	15.1	-	0.01	0.00
			1647 Country of birth (UK/elsewhere): 6: Elsewhere	9.8	-	0.42	0.09
-	84.1	3.02	J18.1 Lobar pneumonia, unspecified	84.1	3.02	0.49	0.09
K	40.1	2.04	3063 Forced expiratory volume in 1-second (FEV1) (liters)	28.1	1.77	-0.19	0.04
			3062 Forced vital capacity (FVC) (liters)	12.1	-	-0.15	0.03
L	40.1	2.04	2188 Long-standing illness, disability or infirmity: 0: No	21.6	-	-0.25	0.06
			2188 Long-standing illness, disability or infirmity: 1: Yes	18.5	-	0.25	0.06
-	31.5	1.86	L97 Ulcer of lower limb, not elsewhere classified	31.5	1.86	0.69	0.14
-	25.5	1.71	N18.9 Chronic renal failure, unspecified	25.5	1.71	0.48	0.10

Here groups were defined post hoc by identifying variables whose *PP*s were negatively correlated. The most significant groups are shown, alongside details of constituent variables. The most significant individual, ungrouped, variables are also shown. *PP*: posterior probability. *P*<sup>\*</sup>: adjusted *P*-value. Dashes (-) indicate *P*<sup>\*</sup> > 0.02.

lacking, and while it contributed to the significance of Group G, we interpreted that group as capturing general poor health. However, no evidence of a direct effect does not imply no evidence of a total effect. These common diseases contribute to a general decline of health, and it is possible that their effects were mediated through pathways better represented by variables or groups we categorised under General comorbidity, such as ‘137 Number of treatments/medications taken’ and Group G. Mediation is not the only explanation; the sparsity-favouring prior may have penalized the inclusion of direct effects of several variables in favour of an aggregate variable like ‘137 Number of treatments/medications taken’ that captured the signal more parsimoniously.

Several notable categories of risk factor that we found to have significant direct effects were included infrequently in published analyses of COVID-19 outcomes in UK Biobank. Variables representing Dementia, Aging (over and above Age) and Infection were included in 10-15% of published analyses, and 2 to 9% of abstracts, whereas we found strong evidence of direct effects of variables we assigned to these categories (*PP* ≥ 99.4% and *P*<sup>\*</sup> ≤ 10<sup>-4.65</sup>), including ‘F03 Unspecified dementia’ (Dementia), ‘R29.6 Tendency to fall, not elsewhere classified’ (Aging) and ‘J22 Unspecified acute lower respiratory infection’ (Infection). These variables were significant even after accounting for all others, such as age and number of treatments/medications. Therefore a

**Table 3. Comparison of risk factors for COVID-19 outcomes in previous UK Biobank studies versus this study**

Category	% Papers	% Abstracts	PP (%)	$-\log_{10} P^*$
Age	90	11	100.0	>5.95
Sex	84	14	100.0	5.95
Obesity	78	20	99.9	5.78
Ethnicity	78	16	94.1	3.69
Socioeconomic status	68	13	99.2	4.48
Smoking	66	6	99.9	5.5
Diabetes	63	10		
Cardiovascular disease	59	12		
Hypertension	54	11		
Lung disease	46	6	99.6	4.82
Alcohol intake	35	3	99.9	5.50
General comorbidity	31	9	100.0	5.78
Cancer	29	1		
Kidney disease	28	6	95.7	3.69
Educational attainment	28	4	96.7	3.82
Asthma	26	3		
Physical activity	24	6		
Neurological disease	21	3		
Liver disease	19	2		
Inflammatory disease	17	2		
Geographic region	17	0		
Aging	15	9	99.4	4.65
Dementia	15	2	99.7	5.00
Employment	13	3		
Immune disease	12	2		
Diet	11	6		
Depression	11	4		
Infection	10	3	99.6	4.82
Arthritis	10	3		
Other	9	2		
Sleep disturbance	9	6		
Psychiatric disorders	9	5		
Mental health	8	4		
Vitamin D	8	4		
Lipid disorders	7	2		
Pollution	5	2		
Covid-19 related	4	3		
Vaccination	4	3		
Allergy	4	1		
Hematological disease	3	1		
Lifestyle	3	2		
Gastrointestinal disease	3	2		
Sex hormones	2	2		
Periodontal disease	2	2		

The percentage of papers, out of 127, are shown. Categories were assigned manually from a literature review, and from Tables 1 and 2. When there were multiple matches in Tables 1 and 2, the maximum significance is given. PP: posterior probability (only values above 50% are shown).  $P^*$ : adjusted  $P$ -value (only values below  $10^{-2.2}$  are shown).

model-averaging big data approach that accounts for widespread correlations among variables and uncertainty in variable selection can bring useful perspective on our understanding of well-studied health outcomes like COVID-19 hospitalization in UK Biobank.

**Disaggregating General Comorbidity.** To test whether the effects of variables like diabetes, cardiovascular disease and hypertension were mediated through or aggregated by general comorbidity, we repeated the Doublethink analysis, removing variables correlated at  $r^2 \geq 0.001$  with any variable we had categorized under General

comorbidities. We removed 86 variables after reinstating those that captured specific comorbidities, age, pollution and blood biomarkers (Datasets S7 and S8). Four variables increased in posterior probability by 25% or more: 'I10 Essential (primary) hypertension' ( $PP = 5.5$  before vs 100.0 after), 'F32.9 Depressive episode, unspecified' ( $PP = 16.9$  vs 92.0), 'E11.9 Type 2 diabetes without complications' ( $PP = 5.9$  vs 54.1) and 'D64.9 Anaemia, unspecified' ( $PP = 5.2$  vs 30.4). The latter two variables formed a post-hoc group with 'N18.9 Chronic renal failure, unspecified', though this group fell short of significance ( $PP = 85.7$ ,

$P^* = 10^{-3.08}$ ). In summary, the evidence that variables capturing general comorbidity mediated or aggregated the effects of other variables was strong for hypertension and depression, suggestive for diabetes and specific complications (chronic renal failure and anemia), and lacking for cardiovascular disease.

## Discussion

All exposome-wide significant direct effects we found had received some attention among the 127 UK Biobank studies in our literature review. However, we found several strongly significant signals in categories of variable overlooked by 85% of studies or more. Since our analysis investigated the same cohort, we drew on studies from other populations to assess the plausibility of these signals.

Dementia was investigated by nineteen out of 127 UK Biobank studies (15%), reporting increased risk of infection, hospitalization, and mortality, with Alzheimer's disease showing the highest risk of COVID-19 diagnosis and mortality (52, 59–76). We found exposome-wide significant evidence that prior hospitalization with unspecified dementia directly increased the subsequent risk of COVID-19 hospitalization. In the United States, analyses of electronic health records found that dementia increased the risk of COVID-19 diagnosis two-fold, with the strongest effect (3.2-fold) for vascular dementia (77), and increased the risk of mortality with COVID-19 by 1.3-fold (78). Proposed mechanisms for a causal effect of dementia on COVID-19 outcomes include compromised social distancing, particularly in care home settings, challenges maintaining personal hygiene, physical frailty, dementia-associated inflammation and immune dysregulation, and direct aggravating effects of viral infection on cardiovascular and respiratory brain function. Interestingly, imaging of 785 UK Biobank participants revealed brain tissue damage following SARS-CoV-2 infection (79), and the ACE2 receptor, by which SARS-CoV-2 invades human cells, is reportedly upregulated in the brains of Alzheimer's patients (80).

Whereas age was investigated by 114/127 UK Biobank studies (90%), molecular and physical signs of aging were investigated by just 19/127 studies (15%). Phenotypic age acceleration [which estimates excess aging via blood biomarkers (81)], shorter leukocyte telomere length, physical frailty including falls, and slower walking pace were associated with worse COVID-19 outcomes (64, 70, 82–98). We found exposome-wide significant evidence that a prior hospital diagnosis of a tendency to fall directly increased the risk of COVID-19 hospitalization. Falls have been identified as an atypical presenting symptom of COVID-19 in some patients, and serve as a marker of underlying frailty (99). Frailty, frequently measured through a subjective clinical frailty score (100), has been reported as a risk factor for COVID-19 mortality in multiple countries, with three meta-analyses supporting the conclusion that frailty increases risk even after accounting for age (101–103).

Lung disease, including prior pneumonia, was commonly investigated (58/127 UK Biobank studies; 46%), but other markers of infection were investigated infrequently (13/127 studies; 10%). In those studies, elevated immune cell counts, infection-related biomarkers, and virus antibody titers were associated with increased COVID-19 infection and severity (61, 64, 68, 74, 89, 104–110). We found exposome-wide significant evidence for an increased risk of COVID-19 hospitalization directly associated with prior hospital episodes of i) unspecified acute lower respiratory infection and ii) constipation or urinary tract infection. In other populations, prior pneumonia, lung disease, and genetic susceptibility to lung disease have been identified as risk factors for COVID-19 (111–116). It is unclear why prior hospital episodes of constipation or urinary tract

infection should increase the risk of subsequent COVID-19 hospitalization (117), although perhaps interestingly, ACE2 is reportedly expressed in the kidney, bladder, and intestine (118, 119). Preexisting pathologies in these tissues might be exacerbated directly by the cell tropism of SARS-CoV-2 infection.

We did not find exposome-wide significant evidence for direct effects of some previously reported risk factors, like diabetes (51–54), asthma (55, 56), and vitamin D (57, 58). This does not rule out indirect effects, but it highlights an important contrast with variables for which we did find significant evidence of direct effects, like the self-reported number of treatments/medications taken, prior hospital diagnosis of acute lower respiratory infection, and a history of medically relevant psychoactive substance abuse (including alcohol and tobacco). We found evidence that the effects of hypertension, dementia, and diabetes were mediated or aggregated by general comorbidity, after we removed 86 variables correlated with general comorbidity above  $r^2 \geq 0.001$ .

There are several reasons we may not have detected some direct risk factors. i) The exposome-wide approach demands stringent multiple testing, reducing power. ii) We employed a nominal FWER of  $\alpha = 10^{-3.3}$ , 100-fold more stringent than the conventional threshold 0.05. Our FWER threshold was interconverted from an FDR of 0.09 using Doublethink (26), but stringent FWER thresholds such as this improve replicability (120) and scaling the FWER with  $1/\sqrt{n}$  ensures the consistency of hypothesis testing in large samples (121). iii) The Doublethink framework incorporates a prior distribution with hyperparameters  $\mu$  and  $b$  that affect power. iv) The Bayesian prior penalizes complex models, which may favor aggregate variables (like number of treatments/medications) over a set of individual risk factors (like diabetes and hypertension).

Our approach has other important limitations. In most settings, the existence of unmeasured mediators means that a direct effect can be defined only relative to the measured variables. Data preprocessing steps that selected the variables for analysis therefore affected the interpretation of direct effects. We curated variables to uphold quality control, avoid reverse causation and avoid collider bias (e.g., by restricting analysis to pre-2020 exposures), and to fix outcome definitions (restricting attention to 2020 due to time-varying vaccine effects). We restricted analysis to 1,912 variables measured in 85% or more of the cohort. The methods chosen to impute missing values, handle repeat measures, and encode factors further shaped our results. If unmeasured confounders remained, this would undermine causal interpretation. We assumed no unmeasured confounders, a strong assumption, but perhaps more defensible in biobank-scale data with thousands of measured variables capturing many possible causal pathways.

Our work offers a new approach at a time when there are increasing calls for exposome-wide association studies (e.g., 9). Agnostic exposome-wide approaches to discovering new risk factors were rare among published studies of COVID-19 outcomes in UK Biobank (Dataset S1). A few studies applied machine learning and univariable scans (14, 20, 54, 69, 117–123), but lacked a principled framework to control false positives rates. In contrast, Bayesian model averaging simultaneously controls the Bayesian FDR and the frequentist FWER, at a level that can be quantified using Doublethink (26). It naturally incorporates uncertainty in which variables to include, which in high-correlation biobank settings can strongly influence the evidence of direct effects. In this study it allowed us to test null hypotheses concerning arbitrary groups of variables, which brought multiple advantages by i) avoiding the need to manually remove correlated variables preanalysis; ii) improving power by combining signals and reducing the stringency of significance thresholds; and iii) conferring flexibility to pursue significant signals

post hoc (124), thereby challenging existing notions of fishing for significance, data dredging, and *p*-hacking (125).

Despite its strengths, our approach has limitations. We tested only for direct effects, not total effects. This is an important distinction: In many applications, it is necessary to estimate the total effect to understand the likely impact of an intervention on the outcome. The direct effect can differ in magnitude and direction to the total effect, and confusing the two is a pitfall known as the Table 2 fallacy (13). Our Monte Carlo Markov Chain approach was computationally expensive, requiring 3500 CPU hours. Its feasibility depended on asymptotic approximations and a computationally expedient prior (26). These demands limited our ability to model important phenomena like interactions between variables, nonlinear effects such as time-since-exposure, or variables with sparse representation like medication use and occupation. Like other methods, Doublethink is subject to *P*-value inflation when testing subsets of highly correlated variables; this can be mitigated by testing groups of correlated variables together (26).

By advancing a more powerful and agnostic approach to identifying direct nongenetic risk factors, our approach has the potential to help advance scientific discovery and bring together the advantages of Bayesian and classical hypothesis testing in biobank-scale settings.

1. C. Sudlow *et al.*, UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
2. V. Y. Tan, N. J. Timpson, The UK Biobank: A shining example of genome-wide association study science with the power to detect the murky complications of real-world epidemiology. *Annu. Rev. Genomics Hum. Genet.* **23**, 569–589 (2022).
3. L. Fachal, A. M. Dunning, From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. *Curr. Opin. Genet. Dev.* **30**, 32–41 (2015).
4. U. M. Marigorta *et al.*, Replicability and prediction: Lessons and challenges from GWAS. *Trends Genet.* **34**, 504–517 (2018).
5. L. E. Duncan, M. Ostacher, J. Ballon, How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology* **44**, 1518–1523 (2019).
6. C. J. Patel, J. Bhattacharya, A. J. Butte, An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS One* **5**, e10746 (2010).
7. C. J. Patel, A. K. Manrai, Development of exposome correlation globes to map out environment-wide associations. *Pac. Symp. Biocomput.* **20**, 231–242 (2015).
8. R. Vermeulen, E. L. Schymanski, A. L. Barabási, G. W. Miller, The exposome and health: Where chemistry meets biology. *Science* **367**, 392–396 (2020).
9. E. Ding, Y. Wang, J. Liu, S. Tang, X. Shi, A review on the application of the exposome paradigm to unveil the environmental determinants of age-related diseases. *Hum. Genomics* **16**, 1–16 (2022).
10. M. K. Chung *et al.*, Decoding the exposome: Data science methodologies and implications in exposome-wide association studies (ExWAS). *Exposome* **4**, osae001 (2024).
11. J. Pearl, M. Glymour, N. P. Jewell, *Causal Inference in Statistics: A Primer* (Wiley, Hoboken, NJ, 2016).
12. J. P. Vandenberg, The history of confounding. *Soz. Präventivmed.* **47**, 216–224 (2002).
13. D. Westreich, S. Greenland, The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *Am. J. Epidemiol.* **177**, 292–298 (2013).
14. G. J. Griffith *et al.*, Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat. Commun.* **11**, 5749 (2020).
15. R. Tibshirani, Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B* **58**, 267–288 (1996).
16. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B* **67**, 301–320 (2005).
17. H. Leeb, B. M. Pötscher, Model selection and inference: Facts and fiction. *Econom. Theory* **21**, 21–59 (2005).
18. R. J. Tibshirani, J. Taylor, R. Lockhart, R. Tibshirani, Exact post-selection inference for sequential regression procedures. *J. Am. Stat. Assoc.* **111**, 600–620 (2016).
19. I. Madakkat, A. Zhou, M. D. McDonnell, E. Hyppönen, Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study. *Sci. Rep.* **11**, 22997 (2021).
20. T. K. Wan *et al.*, Identifying predictors of COVID-19 mortality using machine learning. *Life* **12**, 547 (2022).
21. J. R. Zech *et al.*, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
22. House of Commons Science, Innovation and Technology Committee, The governance of artificial intelligence: interim report. Ninth Report of Session 2022–23. HC 1769 (2023), <https://committees.parliament.uk/publications/41130/documents/205611/default/>.
23. A. E. Raftery, Bayesian model selection in social research. *Sociol. Methodol.* **25**, 111–163 (1995).
24. A. Gelman, Objections to bayesian statistics. *Bayesian Anal.* **3**, 445–450 (2008).
25. A. Gelman, C. R. Shalizi, Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* **66**, 8–38 (2012).
26. H. R. Fryer, N. Arning, D. J. Wilson, Doublethink: simultaneous Bayesian-frequentist model-averaged hypothesis testing. *arXiv [Preprint]* (2025), <https://doi.org/10.48550/arXiv.2312.17566>.
27. B. Efron, R. Tibshirani, J. D. Storey, V. Tusher, Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160 (2001).
28. R. Marcus, P. Eric, K. R. Gabriel, On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660 (1976).
29. V. E. Johnson, Bayes factors based on test statistics. *J. R. Stat. Soc. B.* **67**, 689–701 (2005).
30. V. E. Johnson, Properties of Bayes factors based on test statistics. *Scand. J. Stat.* **35**, 354–368 (2008).
31. A. Zellner, “On assessing prior distributions and Bayesian regression analysis with g-Prior distributions” in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, P. K. Goel, A. Zellner, Eds. (North-Holland, Amsterdam, 1986), pp. 233–243.
32. F. Liang, R. Paulo, G. Molina, M. A. Clyde, J. O. Berger, Mixtures of g priors for Bayesian variable selection. *J. Am. Stat. Assoc.* **103**, 410–423 (2008).
33. G. Schwarz *et al.*, Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
34. R. E. Kass, L. Wasserman, A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.* **90**, 928–934 (1995).
35. S. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**, 60–62 (1938).
36. J. Karamata, Sur un mode de croissance régulière. théorèmes fondamentaux. *Bull. Soc. Math. France* **61**, 55–62 (1933).
37. S. V. Nagaev, Some limit theorems for large deviations. *Theory Probab. Appl.* **10**, 219–233 (1965).
38. R. A. Davis, S. I. Resnick, Limit theory for bilinear processes with heavy-tailed noise. *Ann. Appl. Probab.* **6**, 1191–1210 (1996).
39. D. J. Wilson, The harmonic mean *p*-value for combining dependent tests. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 1195–1200 (2019).
40. D. R. Cox, D. V. Hinkley, *Theoretical Statistics* (Chapman and Hall, London, 1974).
41. W. K. Hastings, Monte carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
42. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2023).
43. Anaconda Inc, Python (2023), <https://www.python.org>.
44. J. Hu, V. E. Johnson, Bayesian model selection using test statistics. *J. R. Stat. Soc. B.* **71**, 143–158 (2009).
45. R. R. Sokal, C. D. Michener, A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* **38**, 1409–1438 (1958).
46. OpenAI, ChatGPT-4o [large multimodal language model]. <https://chat.openai.com/> (2024).
47. M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander, OPTICS: Ordering points to identify the clustering structure. *ACM SIGMOD Rec.* **28**, 49–60 (1999).
48. J. Armstrong *et al.*, Dynamic linkage of COVID-19 test results between Public Health England’s second generation surveillance system and UK Biobank. *Microb. Genomics* **6**, mgen000397 (2020).
49. UK Biobank, Hospital inpatient data. Version 4.0 (2023), <https://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=138483>.
50. UK Biobank, Mortality data: linkage to death registries. Version 3.0 (2023), <https://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=115559>.
51. J. Elliott *et al.*, COVID-19 mortality in the UK Biobank cohort: Revisiting and evaluating risk factors. *Eur. J. Epidemiol.* **36**, 299–309 (2021).
52. J. L. Atkins *et al.*, Preexisting comorbidities predicting COVID-19 and mortality in the UK Biobank community cohort. *J. Gerontol. A Biol. Sci. Med. Sci.* **75**, 2224–2230 (2020).
53. M. Gao *et al.*, Associations between body composition, fat distribution and metabolic consequences of excess adiposity with severe COVID-19 outcomes: Observational study and Mendelian randomisation analysis. *Int. J. Obes.* **46**, 943–950 (2022).
54. K. C. Y. Wong, Y. Xiang, L. Yin, H. C. So, Uncovering clinical risk factors and predicting severe COVID-19 cases using UK Biobank data: Machine learning approach. *JMIR Public Health Surveill.* **7**, e29544 (2021).
55. Z. Zhu *et al.*, Association of asthma and its genetic predisposition with the risk of severe COVID-19. *J. Allergy Clin. Immunol.* **146**, 327–329.e4 (2020).

56. C. J. Lodge *et al.*, Is asthma associated with COVID-19 infection? A UK Biobank analysis. *ERJ Open Res.* **7**, 00309–2021 (2021).
57. H. Ma, T. Zhou, Y. Heianza, L. Qi, Habitual use of vitamin D supplements and risk of coronavirus disease 2019 (COVID-19) infection: A prospective study in UK Biobank. *Am. J. Clin. Nutr.* **113**, 1275–1281 (2021).
58. S. Li *et al.*, Metabolic healthy obesity, vitamin D status, and risk of COVID-19. *Aging Dis.* **12**, 61–71 (2021).
59. Y. Wang *et al.*, Preexisting mental disorders increase the risk of COVID-19 infection and associated mortality. *Front. Public Health* **9**, 684112 (2021).
60. Y. Ma *et al.*, Effect of psychotropics on the risk of COVID-19 in middle-aged and older adults. *Eur. Neuropsychopharmacol.* **66**, 67–77 (2023).
61. X. Fan *et al.*, The effects of obesity and metabolic abnormalities on severe COVID-19-related outcomes after vaccination: A population-based study. *Cell Metab.* **35**, 585–600.e5 (2023).
62. N. Aung, M. Y. Khanji, P. B. Munroe, S. E. Petersen, Causal inference for genetic obesity, cardiometabolic profile and COVID-19 susceptibility: A Mendelian randomization study. *Front. Genet.* **11**, 586308 (2020).
63. J. R. Hilder *et al.*, Association of serum HDL-cholesterol and apolipoprotein A1 levels with risk of severe SARS-CoV-2 infection. *J. Lipid Res.* **62**, 100061 (2021).
64. C. L. Kuo *et al.*, Biological aging predicts vulnerability to COVID-19 severity in UK Biobank participants. *J. Gerontol. A Biol. Sci. Med. Sci.* **76**, e133–e141 (2021).
65. R. K. Topless *et al.*, Gout, rheumatoid arthritis, and the risk of death related to coronavirus disease 2019: An analysis of the UK Biobank. *ACR Open Rheumatol.* **3**, 333–340 (2021).
66. A. C. Tahira, S. Verjovski-Almeida, S. T. Ferreira, Dementia is an age-independent risk factor for severity and death in COVID-19 inpatients. *Alzheimers Dement.* **17**, 1818–1831 (2021).
67. H. Julkunen, A. Cichoska, P. E. Slagboom, P. Würtz, Metabolic biomarker profiling for identification of susceptibility to severe pneumonia and COVID-19 in the general population. *eLife* **10**, e63033 (2021).
68. X. Fan *et al.*, Alcohol consumption is associated with poor prognosis in obese patients with COVID-19: A Mendelian randomization study using UK Biobank. *Nutrients* **13**, 1383 (2021).
69. Y. Xiang, K. C. Wong, H. C. So, Exploring drugs and vaccines associated with altered risks and severity of COVID-19: A UK Biobank cohort study of all ATC level-4 drug categories reveals repositioning opportunities. *Pharmaceutics* **13**, 1337 (2021).
70. J. Ren *et al.*, Impact of allergic rhinitis and asthma on COVID-19 infection, hospitalization, and mortality. *J. Allergy Clin. Immunol. Pract.* **10**, 124–133 (2022).
71. Y. Hu *et al.*, COVID-19 related outcomes among individuals with neurodegenerative diseases: A cohort analysis in the UK Biobank. *BMC Neurol.* **22**, 15 (2022).
72. Z. Liu *et al.*, Associations of sleep and circadian phenotypes with COVID-19 susceptibility and hospitalization: An observational cohort study based on the UK Biobank and a two-sample Mendelian randomization study. *Sleep* **45**, zsc005 (2022).
73. R. K. Topless *et al.*, Gout and the risk of COVID-19 diagnosis and death in the UK Biobank: A population-based study. *Lancet Rheumatol.* **4**, e274–e281 (2022).
74. J. Xie *et al.*, Comparative effectiveness of the BNT162b2 and ChAdOx1 vaccines against COVID-19 in people over 50. *Nat. Commun.* **13**, 1519 (2022).
75. L. Zhou *et al.*, Impact of ultra-processed food intake on the risk of COVID-19: A prospective cohort study. *Eur. J. Nutr.* **62**, 275–287 (2023).
76. A. P. Patel, M. D. Paranjpe, N. P. Kathiresan, M. A. Rivas, A. V. Khera, Race, socioeconomic deprivation, and hospitalization for COVID-19 in English participants of a national biobank. *Int. J. Equity Health* **19**, 114 (2020).
77. Q. Wang, P. B. Davis, M. E. Gurney, R. Xu, COVID-19 and dementia: Analyses of risk, disparity, and outcomes from electronic health records in the US. *Alzheimers Dement.* **17**, 1297–1306 (2021).
78. S. L. Harrison, E. Fazio-Eynullayeva, D. A. Lane, P. Underhill, G. Y. H. Lip, Comorbidities associated with mortality in 31, 461 adults with COVID-19 in the United States: A federated electronic medical record analysis. *PLoS Med.* **17**, e1003321 (2020).
79. G. Douaud *et al.*, SARS-CoV-2 is associated with changes in brain structure in UK Biobank. *Nature* **604**, 697–707 (2022).
80. Q. Ding, N. V. Shults, S. G. Gychka, B. T. Harris, Y. J. Suzuki, Protein expression of angiotensin-converting enzyme 2 (ACE2) is upregulated in brains with Alzheimer's disease. *Int. J. Mol. Sci.* **22**, 1687 (2021).
81. Z. Liu *et al.*, A new aging measure captures morbidity and mortality risk across diverse subpopulations from NHANES IV: A cohort study. *PLoS Med.* **15**, e1002718 (2018).
82. C. Lassale, B. Gaye, M. Hamer, C. R. Gale, G. D. Batty, Ethnic disparities in hospitalisation for COVID-19 in England: The role of socioeconomic factors, mental health, and inflammatory and pro-inflammatory factors in a community-based cohort study. *Brain Behav. Immun.* **88**, 44–49 (2020).
83. G. D. Batty *et al.*, Psychosocial factors and hospitalisations for COVID-19: Prospective cohort study based on a community sample. *Brain Behav. Immun.* **89**, 569–578 (2020).
84. J. L. Atkins *et al.*, Preexisting comorbidities predicting COVID-19 and mortality in the UK Biobank community cohort. *J. Gerontol. A Biol. Sci. Med. Sci.* **75**, 2224–2230 (2020).
85. S. J. Woolford *et al.*, COVID-19 and associations with frailty and multimorbidity: A prospective analysis of UK Biobank participants. *Aging Clin. Exp. Res.* **32**, 1897–1905 (2020).
86. C. Toh, J. P. Brody, Evaluation of a genetic risk score for severity of COVID-19 using human chromosomal-scale length variation. *Hum. Genomics* **14**, 36 (2020).
87. F. K. Ho *et al.*, Is older age associated with COVID-19 mortality in the absence of other risk factors? General population cohort study of 470, 034 participants. *PLoS One* **15**, e0241824 (2020).
88. F. Petermann-Rocha *et al.*, Comparison of two different frailty measurements and risk of hospitalisation or death from COVID-19: Findings from UK Biobank. *BMC Med.* **18**, 355 (2020).
89. F. K. Ho *et al.*, Modifiable and non-modifiable risk factors for COVID-19, and comparison to risk factors for influenza and pneumonia: Results from a UK Biobank prospective cohort study. *BMJ Open* **10**, e040402 (2020).
90. T. Yates *et al.*, Obesity, walking pace and risk of severe COVID-19 and mortality: Analysis of UK Biobank. *Int. J. Obes. (Lond.)* **45**, 1155–1159 (2021).
91. G. D. Batty, I. J. Deary, C. R. Gale, Pre-pandemic cognitive function and COVID-19 mortality: Prospective cohort study. *Eur. J. Epidemiol.* **36**, 559–564 (2021).
92. N. Veronese *et al.*, Neurological diseases and COVID-19: Prospective analyses using the UK Biobank. *Acta Neurol. Belg.* **121**, 1295–1303 (2021).
93. Q. Wang *et al.*, Shorter leukocyte telomere length is associated with adverse COVID-19 outcomes: A cohort study in UK Biobank. *EBioMedicine* **70**, 103485 (2021).
94. Y. V. Chudasama *et al.*, Patterns of multimorbidity and risk of severe SARS-CoV-2 infection: An observational study in the U.K. *BMC Infect. Dis.* **21**, 908 (2021).
95. C. V. Schneider *et al.*, Association of telomere length with risk of disease and mortality. *JAMA Intern. Med.* **182**, 291–300 (2022).
96. J. Reeves, J. S. Koener, W. Zhang, Accelerated ageing is associated with increased COVID-19 severity and differences across ethnic groups may exist. *Front. Public Health* **10**, 1034227 (2022).
97. M. Hamrouni, M. J. Roberts, N. C. Bishop, High grip strength attenuates risk of severe COVID-19 in males but not females with obesity: A short communication of prospective findings from UK Biobank. *Obes. Res. Clin. Pract.* **17**, 82–85 (2023).
98. W. S. Harris, N. L. Tintle, S. P. Sathyanarayanan, J. Westra, Association between blood N-3 fatty acid levels and the risk of coronavirus disease 2019 in the UK Biobank. *Am. J. Clin. Nutr.* **117**, 357–363 (2023).
99. T. Chen *et al.*, Syncope, near syncope, or nonmechanical falls as a presenting feature of COVID-19. *Ann. Emerg. Med.* **76**, 115 (2020).
100. K. Rockwood *et al.*, A global clinical measure of fitness and frailty in elderly people. *CMAJ* **173**, 489–495 (2005).
101. X. M. Zhang *et al.*, Frailty as a predictor of mortality among patients with COVID-19: A systematic review and meta-analysis. *BMC Geriatr.* **21**, 1–11 (2021).
102. R. Pranata *et al.*, Clinical frailty scale and mortality in COVID-19: A systematic review and dose-response meta-analysis. *Arch. Gerontol. Geriatr.* **93**, 104324 (2021).
103. M. Rottler *et al.*, Clinical frailty scale (CFS) indicated frailty is associated with increased in-hospital and 30-day mortality in COVID-19 patients: A systematic review and meta-analysis. *Ann. Intensive Care* **12**, 17 (2022).
104. G. S. Dite, N. M. Murphy, R. Allman, An integrated clinical and genetic model for predicting risk of severe COVID-19: A population-based case-control study. *PLoS One* **16**, e0247205 (2021).
105. R. A. G. Christensen, J. Arneja, K. St Cyr, S. L. Sturrock, J. D. Brooks, The association of estimated cardiorespiratory fitness with COVID-19 incidence and mortality: A cohort study. *PLoS One* **16**, e0250508 (2021).
106. G. S. Dite, N. M. Murphy, R. Allman, Development and validation of a clinical and genetic model for predicting risk of severe COVID-19. *Epidemiol. Infect.* **149**, e162 (2021).
107. B. H. Huang *et al.*, Alcohol intake and mortality risk of COVID-19, pneumonia, and other infectious diseases: An analysis of 437191 UK Biobank participants. *Prev. Med. Rep.* **26**, 101751 (2022).
108. A. A. Willette *et al.*, Using machine learning to predict COVID-19 infection and severity risk among 4510 aged adults: A UK Biobank cohort study. *Sci. Rep.* **12**, 7736 (2022).
109. Y. Lin *et al.*, Impaired kidney function biomarkers and risk of severe COVID-19: Analysis of population-based cohort data. *Mol. Genet. Genomic Med.* **10**, e2047 (2022).
110. E. L. Yanik, B. A. Evanoff, A. M. Dale, Y. Ma, K. E. Walker-Bone, Occupational characteristics associated with SARS-CoV-2 infection in the UK Biobank during August–November 2020: A cohort study. *BMC Public Health* **22**, 1884 (2022).
111. H. Estiri *et al.*, Predicting COVID-19 mortality with electronic medical records. *NPJ Digit. Med.* **4**, 15 (2021).
112. P. Aveyard *et al.*, Association between pre-existing respiratory disease and its treatment, and severe COVID-19: A population cohort study. *Lancet Respir. Med.* **9**, 909–923 (2021).
113. J. Signes-Costa *et al.*, Prevalence and 30-day mortality in hospitalized patients with covid-19 and prior lung diseases. *Arch. Bronconeumol.* **57**, 13–20 (2021).
114. A. Gülsen, I. R. König, U. Jappe, D. Drömann, Effect of comorbid pulmonary disease on the severity of COVID-19: A systematic review and meta-analysis. *Respirology* **26**, 552–565 (2021).
115. Y. Wang, Y. Hao, M. Hu, Y. Wang, H. Yang, Interstitial lung disease independently associated with higher risk for COVID-19 severity and mortality: A meta-analysis of adjusted effect estimates. *Int. Immunopharmacol.* **111**, 109088 (2022).
116. M. Kanai *et al.*, A second update on mapping the human genetic architecture of COVID-19. *Nature* **621**, E7–E26 (2023).
117. M. A. Dabbah *et al.*, Machine learning approach to dynamic risk modeling of mortality in COVID-19: A UK Biobank study. *Sci. Rep.* **11**, 16936 (2021).
118. F. Hikmet *et al.*, The protein expression profile of ACE2 in human tissues. *Mol. Syst. Biol.* **16**, e9610 (2020).
119. X. Zou *et al.*, Single-cell RNA-seq data analysis on the receptor ACE2 expression reveals the potential risk of different human organs vulnerable to 2019-nCoV infection. *Front. Med.* **14**, 185–192 (2020).
120. D. J. Benjamin *et al.*, Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).
121. E.-J. Wagenmakers, A. Ly, History and nature of the Jeffreys-Lindley paradox. *Arch. Hist. Exact Sci.* **77**, 25–72 (2023).
122. S. Tangirala, B. T. Tierney, C. J. Patel, Prioritization of COVID-19 risk factors in July 2020 and February 2021 in the UK. *Commun. Med.* **3**, 45 (2023).
123. A. Córdova-Palamera *et al.*, Assessing the potential of polygenic scores to strengthen medical risk prediction models of COVID-19. *PLoS One* **18**, e0285991 (2023).
124. J. J. Goeman, A. Solari, Multiple testing for exploratory research. *Stat. Sci.* **26**, 584–597 (2011).
125. C. Andrade, Harking, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *J. Clin. Psychiatry* **82**, 25941 (2021).
126. C. Sudlow *et al.*, Data from "UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age." UK Biobank. <https://www.ukbiobank.ac.uk/use-our-data/apply-for-access/>. Accessed 15 October 2021.
127. D. J. Wilson, doublethink-mcmc. Github. <https://github.com/danny-wilson/doublethink-mcmc>. Deposited 2 June 2025.