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

Nicolas Arning, Helen R. Fryer, Daniel J. Wilson

University of Oxford
Big Data Institute, Oxford Population Health, and
Department for Continuing Education

December 2023

Abstract

Big data approaches to discovering non-genetic 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 miss important signals. Bayesian model averaging offers a solution, but classical statistics predominates, perhaps because of concern that the prior unduly influences results. Here we show that simultaneous Bayesian and frequentist discovery of direct risk factors is possible via a model-averaged hypothesis testing approach for large samples called 'Doublethink'. Doublethink produces interchangeable posterior odds and *p*-values that control the false discovery rate (FDR) and familywise error rate (FWER). We implement the Doublethink approach in R and apply it to discover direct risk factors for COVID-19 hospitalization in 2020 among 1,912 variables in UK Biobank. We find nine exposome-wide significant variables at 9% FDR and 0.05% FWER. These include several commonly reported risk factors (e.g. age, sex, obesity) but exclude others (e.g. diabetes, cardiovascular disease, hypertension) which might be mediated through variables measuring general comorbidity (e.g. numbers of medications). We identify significant direct effects among infrequently reported risk factors (psychiatric disorders, infection, dementia and aging), and show how testing groups of correlated variables is a useful alternative to pre-analysis variable selection. We discuss the potential for impact and limitations of joint Bayesian-frequentist inference, and the mutual insights afforded into the long-standing differences on statistical approaches to scientific discovery.

Keywords: UK Biobank, COVID-19, exposome, Bayesian, frequentist, closed testing procedure, multiple testing, epidemiology, posterior odds, *p*-values, confounding, Bayesian false discovery rate, false positive rate, strong-sense familywise error rate, hypothesis testing, risk factors, model averaging

Introduction

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 non-genetic determinants, termed risk factors, has not become commonplace. Instead, current epidemiology typically reports on candidate risk factors. Studies address the question: What is the total effect of a variable on the outcome? Is it non-zero? For instance, more than 100 published studies have analysed dozens of candidate risk factors for COVID-19 outcomes in UK Biobank (Table S1). Synthesizing these findings is difficult because: (i) Other, more important, risk factors that were not analysed may exist among the thousands measured. (ii) It is unclear how to appropriately limit false positives caused by multiple testing. (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, 4, 5).

A major complication for systematic studies of non-genetic risk factors, compared to GWAS, is the problem of mediation (6). Mediation occurs when the total effect of a variable (e.g. age) on an outcome (e.g. COVID-19 severity) is wholly or partially mediated 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 coinherited at conception; they cannot generally cause one another. So the question is effectively: What is the direct effect of a variable on the outcome? Is it non-zero? In GWAS, artefactual signals generated by confounding are instead the major concern. Controlling for other variables helps avoid confounding (7), but controlling for mediating variables alters the scientific question by shifting attention from total to direct effects. Unfortunately, direct effects can differ in direction and magnitude to total effects, a source of bias known as the Table 2 fallacy (8). Further pitfalls include reverse causation and collider bias (9).

Nevertheless, the demand for GWAS-inspired exposome-wide association studies (10) presents an opportunity, which has been partly filled by machine learning (11, 12). Machine learning offers a data-driven agnostic approach. A major advantage is the ability to analyse high dimensional data with minimal curation, even in the presence of collinearity and widespread correlation between variables. But the question is different: What is the contribution of a variable to predicting the outcome? Usually there is no formal test. More importantly, a variable can be valuable for prediction due to confounding (13). 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 (14).

Bayesian methodology offers a solution to the question of identifying direct effects in biobank-scale data while controlling for confounding (15). 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 non-zero? With a careful approach to feature engineering to mitigate issues around mediation, reverse causality and collider bias, and with independent validation akin to GWAS, 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 (Table S1). This might be explained by several issues, including lack of familiarity among researchers, high computational requirements, and difficulties specifying prior distributions (16). 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 (17).

Here we apply Bayesian model averaging to test for direct risk factors among individual variables and groups of variables. Moreover, we use our new method, Doublethink, that assumes a specific prior and large sample size, to produce a one-to-one correspondence between Bayesian model-averaged posterior odds and traditional *p*-values (18). This allows simultaneous control of not just the Bayesian false discovery rate (FDR), but also the frequentist familywise error rate (FWER). We apply Doublethink to investigate direct risk factors for COVID-19 hospitalization in UK Biobank, and we compare our results to the literature. This 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

The Doublethink method (18) considers a general regression setting in which there are n observed outcomes $Y_1 \ldots Y_n$ and v variables (features) with regression coefficients $\beta_1 \ldots \beta_v$. The aim is to identify which variables directly influence the outcome, i.e. which of the regression coefficients are non-zero. In total there are 2^v models which we identify via vector **s**, the jth element of which indicates whether variable j is included ($s_j = 1$; $\beta_j \neq 0$) or excluded ($s_j = 0$; $\beta_j = 0$) from the model.

We are interested in testing the null hypothesis that the variables indexed by a set ($\mathcal{V}$) all have regression coefficients $\beta_j = 0$ (for $j \in \mathcal{V}$). We can average over the inclusion or exclusion of all other variables to produce a set of models compatible with the null hypothesis,

$$\mathcal{O}_{\mathcal{V}} = \{s \in \mathcal{S} : s_j = 0 \text{ for all } j \in \mathcal{V}\}, \tag{1a}$$

and a complementary set of models compatible with the alternative hypothesis,

$$\mathcal{A}_{\mathcal{V}} = \mathcal{S} \setminus \mathcal{O}_{\mathcal{V}}, \tag{1b}$$

where $\mathcal{S}$ is the state space of **s**. In the Bayesian setting, we reject the null hypothesis if the posterior odds of $\mathcal{A}_{\mathcal{V}}$ versus $\mathcal{O}_{\mathcal{V}}$,

$$\mathrm{PO}_{\mathcal{A}_{\mathcal{V}}:\mathcal{O}_{\mathcal{V}}} = \frac{\sum_{s \in \mathcal{A}_{\mathcal{V}}} Pr(Y|s)\, Pr(s)}{\sum_{s \in \mathcal{O}_{\mathcal{V}}} Pr(Y|s)\, Pr(s)}, \tag{1c}$$

exceed some threshold τ. The Bayesian false discovery rate (FDR), both local and global (19), is then controlled at or below $1/(1+\tau)$, contingent on the prior.

Lemma 1 (18). The Bayesian procedure defined by Equation 1, which rejects the null hypothesis $\mathcal{O}_{\mathcal{V}}$ when $\mathrm{PO}_{\mathcal{A}_{\mathcal{V}}:\mathcal{O}_{\mathcal{V}}} > \tau$, is a closed testing procedure (20), and therefore controls the frequentist FWER in the strong sense.

To implement simultaneous Bayesian and frequentist inference, we assumed priors of the form

$$s_j \sim \text{Bernoulli}\left(\frac{\mu}{1+\mu}\right),\ j = 1 \ldots \nu \tag{2a}$$

$$\theta_s \sim \text{Normal}\left(0, h^{-1}\, \mathcal{I}_s{}^{-1}\right) \tag{2b}$$

where μ are the prior odds that $\beta_j \neq 0$, h is a prior scale factor, $\theta_{\mathbf{s}}$ are the unconstrained parameters in model **s** (the β_j for which $s_j = 1$, and any nuisance parameters), and $\mathcal{I}_s$ is the per-observation Fisher information matrix for model **s**, evaluated at $\theta_{\mathbf{s}} = \mathbf{0}$. Fisher's information matrix has been used widely in the definition of reference priors (e.g. 21, 22), and to generate concordance between Bayesian and frequentist point and interval estimates (see Table 1 of 18).

Subject to further assumptions, principally that (i) the outcomes are independent, given the model, (ii) n is large, and (iii) h is proportional to n (known as 'local alternatives'), Johnson (23, 24) showed that the posterior odds between model **t** and nested model **s** are

$$\mathrm{PO}_{t:s} \sim \mu^{|t|-|s|}\left(\frac{h}{n+h}\right)^{(|t|-|s|)/2} R_{t:s}, \tag{3}$$

i.e. proportional to the maximized likelihood ratio $R_{t:s}$. These assumptions enable interconversion between Bayesian and frequentist hypothesis tests. From the well-known result that the deviance ($2\,log\,R_{t:s}$) follows a chi-squared distribution with $|t| - |s|$ degrees of freedom, conditional on model **s**, one can transform the posterior odds into a *p*-value as

$$p_{\boldsymbol{t:s}} \sim Pr\left(\chi^2_{|\boldsymbol{t}|-|\boldsymbol{s}|} > 2\,log\,\frac{\mathrm{PO}_{\boldsymbol{t:s}}}{\mu^{|\boldsymbol{t}|-|\boldsymbol{s}|}\left(\frac{h}{n+h}\right)^{(|\boldsymbol{t}|-|\boldsymbol{s}|)/2}}\right), \tag{4}$$

which is a rearrangement of the familiar $p_{\boldsymbol{t:s}} = Pr\left(\chi^2_{|\boldsymbol{t}|-|\boldsymbol{s}|} > 2\,log\,R_{t:s}\right)$ (25). Note the *p*-value depends on $R_{t:s}$, but not the hyper-parameters μ and h.

Following Johnson (23, 24), we drop the assumption that h is proportional to n, and assume it is fixed. Although this contradicts a motivating assumption, it has the desirable effects of (i) removing the dependency of the prior on n, (ii) achieving statistical consistency and (iii) recapitulating the Bayesian information criterion (BIC) when $h = 1$, which has been shown to reasonably approximate a wide range of Bayes factors when n is large (26, 27). We further assume that $\mu < 1$, and each variable has one parameter, and one degree-of-freedom. We use the theory of regularly varying random variables (28, 29) to derive the following.

Theorem 1 (18). We find that, under these assumptions, the model-averaged posterior odds,

$$\mathrm{PO}_{\mathcal{A}_\mathcal{V}:\mathcal{O}_\mathcal{V}} \sim |\mathcal{V}|\,\mu\left(\frac{h}{n+h}\right)^{1/2} R_{\mathcal{A}_\mathcal{V}:\mathcal{O}_\mathcal{V}}, \tag{5}$$

can be transformed into an asymptotically valid *p*-value for large n. Here $R_{\mathcal{A}_\mathcal{V}:\mathcal{O}_\mathcal{V}}$ is a weighted mean of the nested maximized likelihood ratios in $\mathcal{O}_\mathcal{V}$ and $\mathcal{A}_\mathcal{V}$, and the *p*-value (unadjusted for multiple testing) is

$$p_{\mathcal{A}_\mathcal{V}:\mathcal{O}_\mathcal{V}} \sim Pr\left(\chi^2_1 > 2\,log\,\frac{\mathrm{PO}_{\mathcal{A}_\mathcal{V}:\mathcal{O}_\mathcal{V}}}{|\mathcal{V}|\,\mu\left(\frac{h}{n+h}\right)^{1/2}}\right)\ \text{as}\ n \to \infty. \tag{6}$$

Theorem 2 (18). The level at which this procedure controls the FWER in the strong sense is

$$\alpha \sim Pr\left(\chi^2_1 > 2\,log\,\frac{\tau}{\nu\,\mu\left(\frac{h}{n+h}\right)^{1/2}}\right)\ \text{as}\ n \to \infty. \tag{7}$$

As a corollary, the Bayesian procedure is equivalent to rejecting the null hypothesis $\mathcal{O}_\mathcal{V}$ when an adjusted *p*-value,

$$p^{\star}_{\mathcal{A}_\mathcal{V}:\mathcal{O}_\mathcal{V}} \sim Pr\left(\chi^2_1 > 2\,log\,\frac{\mathrm{PO}_{\mathcal{A}_\mathcal{V}:\mathcal{O}_\mathcal{V}}}{\nu\,\mu\left(\frac{h}{n+h}\right)^{1/2}}\right)\ \text{as}\ n \to \infty, \tag{8}$$

is smaller than threshold α.

An equivalent interpretation of these results is that the model-averaged deviance ($2\,log\,R_{\mathcal{A}_\mathcal{V}:\mathcal{O}_\mathcal{V}}$) follows a chi-squared distribution with one degree of freedom, when large. This means Doublethink *p*-values cannot be arbitrarily rescaled by the prior parameters μ and h 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 μ and h influence the power of the test, but not its theoretical distribution under the null hypothesis. This makes model-averaged hypothesis testing a viable frequentist procedure by facilitating a prior-agnostic approach to quantifying Bayesian significance thresholds in terms of FWER, for large samples.

Methods

We implemented the Doublethink approach as a Monte Carlo Markov Chain approach (30) in R (31) and Python (32) and applied it to identify risk factors for COVID-19 hospitalization in UK Biobank, following the COVID-19 Host Genetics Initiative definition 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 that died before 2020, non-England residents determined by assessment centre, and those that withdrew before the analysis. The total numbers of controls were 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 centre, 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, including negative factor levels, were treated as missing. We imputed missing continuous and integer covariates taking the mean of non-missing 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 analysed 184 covariates, binary variables encoding 865 levels across 193 factors, and 863 ICD-10 admission codes, a total of 1,912 variables (Supplementary Table S2).

Model. We fitted the data separately for each outcome 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 $v = 1{,}912$ variables, implying a prior expectation of 10 variables in the model. We assumed a unit information prior ($h = 1$) for the regression coefficients (27). We disallowed the inclusion of collinear variables by defining a zero likelihood.

Implementation. Like the implementation in (33), we employed a Markov Chain Monte Carlo (MCMC) sampler over the variable inclusion vector **s**. We ran 100 chains with 25,000 iterations of burn-in and 75,000 iterations of sampling. 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. We simulated regression coefficients directly from conditional Normal distributions by post-processing the MCMC iterations. We calculated posterior odds and parameter estimates by combining chains, computing standard errors across independent chains.

Grouping variables. We could perform valid post-hoc variable grouping while controlling the FDR and FWER, which was useful since correlated variables reduce one-another's individual posterior inclusion probabilities. We clustered variables in two ways: (i) *A posteriori*, using the scikit-learn OPTICS algorithm with distances defined by their posterior correlation in inclusion probabilities. This grouped the variables with the strongest negative correlations in inclusion probabilities. (ii) *A priori*, using the algorithm applied to one minus the squared correlation between variables. We computed the posterior odds of including any variable among each group.

***p*-value calculation.** We used the chi-squared distribution to compute adjusted *p*-values using Theorem 2. In the case of orthogonal variables with one degree-of-freedom, this is conservative for $p < 0.02$. Since the large n assumption implies interest in small significance thresholds, we report any *p*-value larger than 0.02 as n.s. (not significant) or '-'. This makes Doublethink incompatible with any threshold exceeding $\alpha = 0.02$. In practice, we did not explicitly set a threshold, instead reporting 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 on 19 September 2023. After excluding Review Articles and Editorial Material, this search returned 203 publications. We analysed a subset of 127 of these papers that quantified the effect of non-genetic 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 (Supplementary Table S1). We manually categorized the variables analysed by these 127 papers into groups (Supplementary Table 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 the UK Biobank, to help understand the underlying processes, by using model-averaged hypothesis tests to account for uncertainty in variable selection and deplete for potential confounders, subject to the limitations of the data in terms of unmeasured variables and measurement error. We intended 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 selection bias, for example access to testing, which may affect interpretation (9). We focus 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).

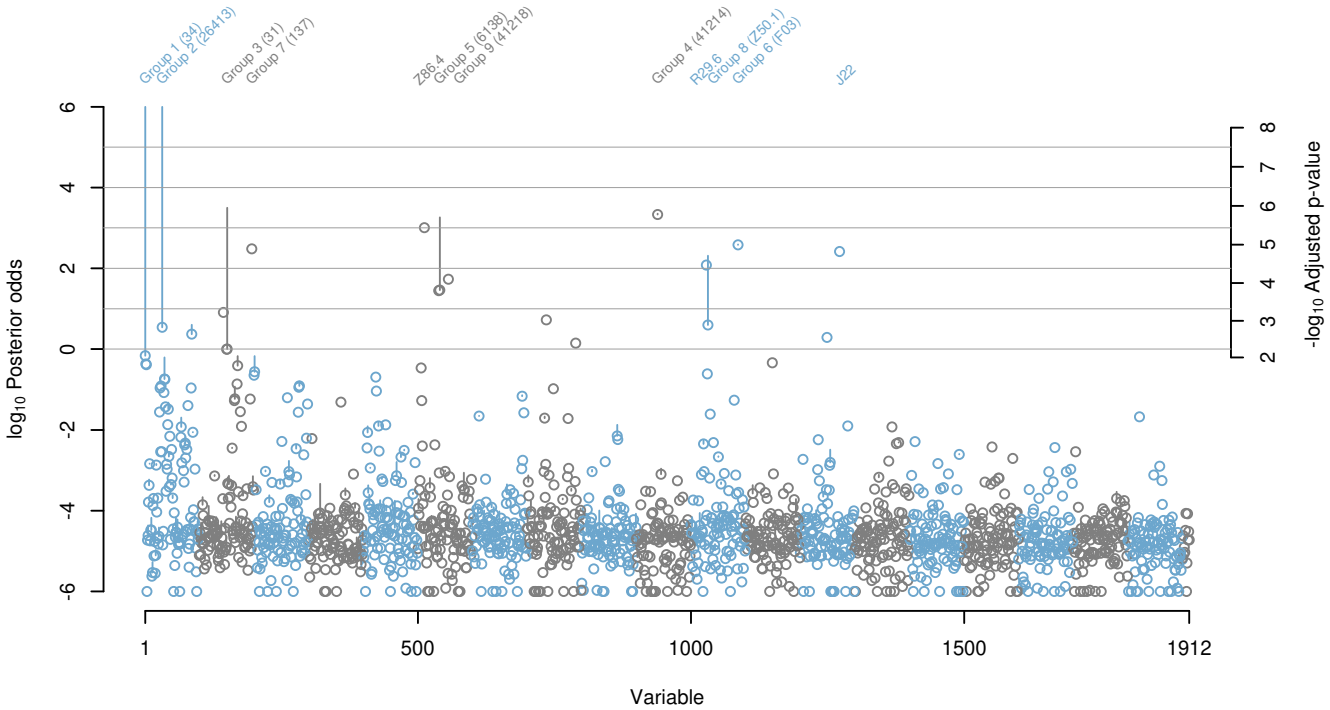

Figure 1 Individual variables (points) and pre-defined groups of variables (vertical lines) with the strongest evidence of direct effects on the risk of COVID-19 hospitalization in UK Biobank. Evidence was quantified simultaneously by $\log_{10}$ posterior odds and $-\log_{10}$ adjusted *p*-value using Doublethink. Groups were defined *a priori*. Points and lines are coloured for legibility. Variables were ordered horizontally using OPTICS. Vertical lines show the boost in significance (if any) from the most significant individual variable per group to the significance of the whole group. Significance was truncated to $\log_{10}$ posterior odds between -6 and 6. Individual variables and groups significant at $\log_{10}$ posterior odds ≥ 1 are labelled (with the most significant variable per group in parentheses). Individual variables are named by UK Biobank field ID or, when prefixed by a letter, ICD-10 code. Refer to Table 1 for full names.

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 Theorem 2 (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 FWER and – for the assumed prior – the FDR, under the asymptotic approximation. For example, a Bayesian threshold of $\tau = 10$ would control 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, partly 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. Significant variables, such as the ICD-10 codes `F03 Unspecified dementia`, `J22 Unspecified acute lower respiratory infection` and `R29.6 Tendency to fall, not elsewhere classified`, appeared to 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 was important here because no single model had high posterior probability.

Several significant variables were indicators or aggregates of presumptive underlying processes, such as `41214 Carer support indicators : 1 : Yes`, which indicates a hospital record of past carer support, `137 Number of treatments/medications taken`, which summarizes the recruitment interview, and `Z86.4 Personal history of psychoactive substance abuse`, which indicates a hospital record of past alcohol, tobacco or drug use. The 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 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, $PP = 49.9$; $p^\star = 10^{-2.23}$; where posterior odds = $PP/(1\text{-}PP)$) and `34 Year of birth (years)` ($PP = 40.8$; $p^\star = 10^{-2.05}$; 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, inevitably 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. However, correlation is pervasive in biobank-scale data. An alternative solution is to define groups of correlated variables and test whether one or more members of a group affect the outcome. Doublethink 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

Nine groups of variables defined *a priori* were significant at $\tau = 10$ and $\alpha = 10^{-3.3}$, often when the individual member variables were not. In Figure 1, vertical lines illustrate the boost in the significance of groups of variables compared to their most significant member. The groups are numbered for cross-reference with Table 1. For example, the well-established risk factors age (Group 1; $PP = 100$; $p^\star < 10^{-5.95}$), indices of multiple deprivation (Group 2; $PP = 100$; $p^\star < 10^{-5.95}$) and sex (Group 3; $PP = 100$; $p^\star = 10^{-5.95}$) 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.

Finding that a group of variables is significant means there is evidence that one or more of them influence the outcome, after controlling for all other measured variables. This controls confounding caused by variables outside the group, but combines signals within the group, increasing power. For example, Group 8 was strongly significant ($PP = 99.5$, $p^\star = 10^{-4.71}$) while containing variables that were individually less so: `Z50.1 Other physical therapy` ($PP = 79.9$, $p^\star = 10^{-2.89}$) and `Z50.7 Occupational therapy and vocational rehabilitation, not elsewhere classified` ($PP = 19.6$, $p^\star > 0.02$). One of these variables, or something they measure that is not captured by other variables, presumably influences the risk of COVID-19 hospitalization, even if we cannot attribute the effect specifically to either.

Testing groups is useful but defining them *a priori* is not the most effective method of discovering signals, because the groupings might not be relevant to the outcome under investigation. For example, Group 8 also included variable `Z50.5 Speech therapy`, which appeared to contribute nothing to the

group's overall significance ($PP = 0.0$, $p^\star > 0.02$). Conversely, failure to group relevant variables together can cause signals to be overlooked, as we will see.

Doublethink allows arbitrary groups to be tested

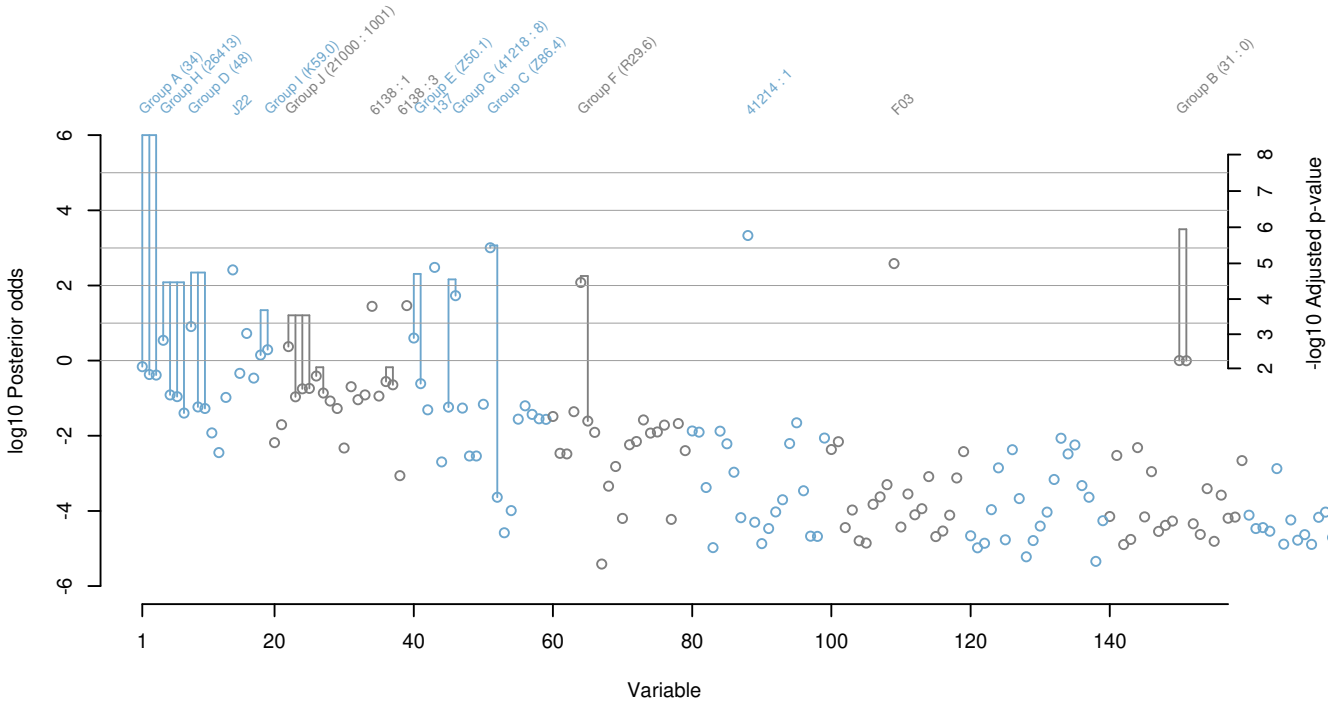

Figure 2 Individual variables (points) and post-hoc groups of variables (vertical lines) with the strongest evidence of direct effects on the risk of COVID-19 hospitalization in UK Biobank. The horizontal axis is truncated to show only significant variables and groups. Vertical lines show the boost in significance (if any) from individual member variables to the significance of the whole group. See Figure 1 legend for further details.

One of the advantages of the Doublethink approach is 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 pre-defined in the closed testing procedure. 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 (*PP*s) were negatively correlated, because this suggests they 'competed' for inclusion in the model.

Figure 2 and Table 2 show that post-hoc grouping revealed signals that were weaker in pre-defined groupings, such as Group D ($PP = 99.5$, $p^\star = 10^{-4.75}$), which captured aspects of obesity by combining the individual variables `48 Waist circumference (cm)` ($PP = 89.0$, $p^\star = 10^{-3.22}$), `21001 Body mass index (BMI) (Kg/m2)` ($PP = 5.5$, $p^\star > 0.02$) and `23104 Body mass index (BMI) (Kg/m2)` ($PP = 5.0$, $p^\star > 0.02$). In contrast, prior Group 10 ($PP = 89.0$, $p^\star = 10^{-3.22}$) had only combined `48 Waist circumference (cm)` with the non-significant variables `21002 Weight (Kg)` and `23098 Weight (Kg)` (Table 1).

Some post-hoc groups overlapped the pre-defined groups but dropped non-significant variables that did not contribute to the significance of the group. For example, Group H ($PP = 99.2$, $p^\star = 10^{-4.48}$), contained only the three most significant deprivation scores of the eight members of Group 2. Other groups revealed new connections between variables, such as Group I ($PP = 95.7$, $p^\star = 10^{-3.69}$), which combined the individually non-significant `K59.0 Constipation` ($PP = 66.1$, $p^\star = 10^{-2.55}$) and

`N39.0 Urinary tract infection, site not specified` ($PP = 58.4$, $p^\star = 10^{-2.40}$), a combination that might reflect underlying kidney disease.

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: $PP = 99.3$, $p^\star = 10^{-4.55}$). 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 risk of COVID-19 hospitalization (Table 2).

The ability to quantify the evidence for groups of variables offers an alternative to approaches such as pre-analysis 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 new possibilities for identifying significant groups, and the identification of these groups may help with the interpretation of the role in 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 focused on large established cohorts like UK Biobank that could rapidly link to data on SARS-CoV-2 testing (34), COVID-19 hospitalization (35) and mortality (36). Since then, many risk factors have been reported, including smoking (37, 38, 39, 40, 41, 42), diabetes (38, 43, 44, 45), asthma (46, 47) and vitamin D (48, 49) as predisposing to worse outcomes. We compared our results to the literature on COVID-19 in UK Biobank to identify any differences to standard approaches and find new 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.

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 $p^\star > 10^{-2.20}$) omitted, since the Bayesian interpretation is this represents evidence against those risk factors.

Age, Sex, Obesity, Ethnicity, Socioeconomic status (including deprivation indices) and Smoking were included in more than 66-90% of published analyses. These well-established risk factors were mentioned in 6-20% of abstracts. Our analysis supported direct effects of all these categories with $PP \geq 99.5\%$ and $p^\star \leq 10^{-4.75}$ except ethnicity (Supplementary Table S4, S5). Post-hoc Group J had strong support ($PP = 94.1$, $p^\star = 10^{-3.54}$) but combined self-reported ethnicity and country of birth with geographic measures of pollution. However, pre-defined Group 11, which did not contain pollution metrics, was not significant at $\tau = 10$ and $\alpha = 10^{-3.3}$, despite the evidence being suggestive ($PP = 80.0$, $p^\star = 10^{-2.89}$).

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 28-46% of published analyses and 4-9% of abstracts. Our analysis supported these categories with $PP \geq 95.7\%$ and $p^\star \leq 10^{-3.69}$.

Many categories of risk factors that appeared commonly in published analyses received no significant support for direct effects in our analysis. Diabetes, Cardiovascular disease and Hypertension were notable for inclusion in 54-63% of published analyses, and 10-12% of abstracts. No variables or groups of variables corresponding to these categories received support for direct effects in our analyses ($PP < 50\%$, $p^\star > 10^{-2.20}$). However, no evidence of a direct effect does not imply no evidence of an 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.

Several notable categories of risk factor that we found to have significant direct effects were rarely included in published analyses of COVID-19 outcomes in UK Biobank. Variables representing Psychiatric disorders, Infection, Dementia and Aging were included in 9-15% of published analyses, and 2-9% of abstracts, whereas we found strong evidence of direct effects of variables we assigned to these categories ($PP \geq 99.4\%$ and $p^\star \leq 10^{-4.65}$), including `41214 Carer support indicators : 1 : Yes` (which we categorised under Psychiatric disorders), `J22 Unspecified acute lower respiratory infection` (Infection), `F03 Unspecified dementia` (Dementia) and `R29.6 Tendency to fall, not elsewhere classified` (Aging). Therefore a model-averaging big data analysis that accounts for widespread correlations among variables and uncertainty in variable selection can bring new insight to our understanding of well-studied health outcomes like COVID-19 hospitalization in UK Biobank.

Discussion

Doublethink for discovery of non-genetic risk factors

Bayesian model averaging is an approach especially suited to accounting for model uncertainty, such as which variables directly affect an outcome. Here we showed that appropriate construction of the prior allows simultaneous control of the Bayesian FDR and the frequentist FWER, facilitating its use in fields, like epidemiology, where classical statistics predominates. This allowed us to screen 1,912 variables for direct effects on COVID-19 hospitalization and adjust for confounders via model averaging without the need to specify a candidate risk factor (primary exposure) nor select variables in a pre-analysis step such as a univariable scan, stepwise regression or machine learning. Instead of attempting to identify independent variables – a near-impossible task in biobank scale data where correlation is pervasive – we performed tests on groups of correlated variables, which can increase power and, in some cases, interpretability.

This work offers a new approach at a time when there are increasing calls for exposome-wide association studies (e.g. 10). With some exceptions (9, 12, 45, 50, 51, 52, 53), agnostic exposome-wide approaches to discovering new risk factors were absent from published studies of COVID-19 outcomes in UK Biobank (Table S1). Perhaps that is particularly surprising given that COVID-19 was a new disease in which all risk factors were initially unknown.

The approach of this paper had numerous limitations. Principally, we did not assess the total effect of a variable on the outcome, only the direct effect. In some 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 (8). Focusing on direct effects therefore limits the interpretation of our conclusions. In particular, we are unable to predict the effect on the outcome of a hypothetical intervention in the study population.

Importantly, we cannot conclude that no evidence of a direct effect implies no evidence of an effect. For example, hypertension was mentioned in 11% of abstracts and included in 54% of published analyses of COVID-19 outcomes in UK Biobank, but we did not find significant evidence of its direct effect on hospitalization. This does not rule out an indirect effect mediated through another variable, such as a general decline in health. Several fields captured general comorbidity, including `137 Number of treatments/medications taken`. Not only might they mediate indirect effects, but fields like `137` that pool, aggregate or summarize data from several other sources might be favoured for inclusion by the sparsity-inducing prior, which imposed a penalty μ on every additional parameter. This ability of the Bayesian prior to influence the final results is inevitable and exists despite the ability to control the frequentist FWER.

In practice, the main considerations for running Doublethink are (i) preparation of the outcome data and variables, (ii) choice of hyper-parameters and (iii) computational feasibility. (i) As demonstrated with the UK Biobank analysis, it is not necessary to filter variables *a priori* to find 'independent' sets, usually an impossible task. Instead, variables can be grouped *post hoc* to sidestep correlation and detect signals. Data quality control is still paramount. (ii) To choose the hyper-parameters, the main consideration is the average number of variables expected to be included in the model, which determines μ; we chose $v\mu/(1+\mu) = 10$. For many purposes, the unit information prior of $h = 1$ will suffice. For non-Bayesians, these hyper-parameters determine the performance envelope of the analysis, with performance optimal when the 'truth' resembles the prior. The aim of the paper is the method should still be useful, and the p-values theoretically well calibrated, at other times. (iii) Computation is the major limitation. The analysis of $v = 1{,}912$ variables in $n = 201{,}912$ UK Biobank participants required 100,000 iterations of 100 independent chains running for 35 hours each. This is substantially slower than many machine learning algorithms.

The Monte Carlo Markov Chain approach pursued here was computationally intensive, despite restricting our attention only to direct effects. Requiring 3500 CPU hours, its feasibility relied on efficiency gains stemming from (i) asymptotic approximations motivated by an assumption of large sample size and (ii) a convenient prior. The computational demands of the approach prevented us from investigating important phenomena like interactions between variables and non-linear effects such as time-since-exposure. With a less computationally expensive approach, we might have investigated other potential risk factors with large numbers of rare variables, such as occupation and use of specific medicines.

In an analysis of non-genetic direct risk factors, there is no possibility of a definitive approach, even for an agnostic scan. Partly, this is because direct effects are only defined relative to a fixed set of variables: conceptually, a direct effect could be mediated through one or more downstream variables that were not measured or included. Moreover, no method is free of data curation. This includes choice of the exposure variables to uphold quality control, avoid reverse causation and avoid collider bias (we restricted analysis to pre-2020 exposures), and choice of outcome (we restricted attention to 2020 given the time-varying dynamics and likely impact of vaccination status, which we did not know). Methods to impute missing values, handle repeat measures and encode factors can all impact the final results.

Doublethink and simultaneous Bayesian/frequentist hypothesis testing

Other theoretical considerations that may limit the applicability of Doublethink include assumptions of large sample sizes and a specific family of priors (or 'random effects') parameterized by μ and h. The prior covariance on the coefficients (β), based on Fisher information, is hard to justify except through its convenience for pursuing joint Bayesian/frequentist inference. This motivating aim is only achieved

theoretically as n becomes arbitrarily large. So strictly speaking that aim is not truly met, meaning the theoretical properties of the p-value may not hold precisely, leading to inflation or deflation, particularly when p is not small, and reducing robustness to poor choices of h and particularly μ. The theory contains a technical contradiction, because h is initially assumed to scale with n (the local alternatives assumption), allowing the posterior odds to be written in terms of the maximized likelihood ratio, but later h is assumed to be constant with respect to n, which affords simplifications in deriving a p-value for the model-averaged posterior odds (23, 24).

Beyond its practical utility, Doublethink has wider implications for bridging the gap between Bayesian and classical philosophies to scientific inference. First, Lemma 1 showed that the Bayesian approach to testing a collection of null hypotheses $\beta_j = 0, j \in \mathcal{V}$, in which the null hypothesis is rejected when the posterior odds exceed a fixed threshold τ, is a closed testing procedure (20), which therefore controls the frequentist FWER in the strong sense at or below some level α. This result is general and does not depend on the Doublethink model. Importantly, it disproves the idea that the FWER is a fundamentally non-Bayesian quantity, inherently more stringent than the FDR, which the Bayesian approach controls at or below level $1/(1+\tau)$.

Second, Theorem 2 gave an analytic expression for α, the level at or below which the FWER is controlled in the strong sense, asymptotically under the Doublethink model. From there, we could interconvert model-averaged posterior odds and adjusted p-values, and equivalently, we could interconvert FWER and FDR thresholds. This affords insights by allowing frequentist multiple testing thresholds to be understood in terms of Bayesian prior assumptions. According to Theorem 2, α is asymptotically proportional to $\nu\,\mu\,(h/n)^{1/2}/\,\tau$. That is, asymptotically proportional to (i) the number of variables, v, which underlies Bonferroni correction, (ii) the prior odds, μ, of the alternative hypothesis versus the null, and (iii) the square root of the prior precision h, higher values of which make the alternative hypothesis more similar to the null; and inversely proportional to (iv) the Bayesian threshold τ, and (v) the square root of the sample size n. The relationship $\alpha = \alpha_0\,\nu\,\mu\,(h/n)^{1/2}$, for some constant α_0, offers a resolution to the classical paradox about multiple testing (54): should I vary α_0 to correct for the number of tests in an analysis, the number of tests in the whole paper, the number of tests I perform in my career, or the number of tests in the scientific literature? The Bayesian response is to fix α_0 not α, that is to fix the FDR and allow the FWER to vary depending on v, μ, h and n. As n grows, this controls the FWER far more stringently than the FDR anyway.

Third, Theorem 2 revisits the Jeffreys-Lindley paradox (55) by emphasizing a principal difference between Bayesian and frequentist hypothesis tests, not in philosophical issues like the treatment of parameters as fixed or random, but in the practical choice of significance threshold. The common practice of fixing the FWER irrespective of n, e.g. at 0.05 or 0.005 (56), leads to tests that are inconsistent under the null, because there is a tangible probability, α, of wrongly rejecting the null hypothesis even for arbitrarily big data (57). This is solved by varying α with $n^{-1/2}$, an alternative starting point from which one could calculate either the FWER or the FDR.

The FDR considered here is related to, but distinct from, some FDR concepts common in the literature. First, we control the Bayesian FDR, rather than a frequentist FDR controlled by procedures like Benjamini and Hochberg's (58) or Storey's (59). Second, we control the local FDR (19), meaning we only call an individual variable significant when its posterior inclusion probability exceeds the threshold, $\tau/(1+\tau)$. Often, frequentist FDR procedures call as many individual variables significant as possible such that the mean FDR is controlled. On average this will reject more individual variables, but there are two caveats. First, it allows individual variables with lower posterior probability to be

called significant, meaning we reject the null hypothesis that their direct effects are zero. Second, it may still miss variables whose individual significance has been diluted by correlation with other variables. As we have seen, such signals can be recovered by testing groups of variables. When a group of variables is called significant, we reject the null hypothesis that all their direct effects are zero, without necessarily pinpointing which variables have a non-zero direct effect.

Further research is needed to determine the generalizability of some of our theoretical findings beyond the Doublethink model. The form of Theorem 2 depends on the model choice prior, in which μ is fixed. The impact of co-estimating μ requires attention. The prior on the coefficients also matters. On one hand, Doublethink may be general in that log Bayes factors for nested hypothesis tests converge asymptotically to the Schwarz criterion (27, 57), which Doublethink recapitulates when n is large. On the other hand, the derivation of Theorem 2 relies heavily on the theory of regular variation (28, 29, 60). In Doublethink, the posterior odds are regularly varying random variables, but slowly varying posterior odds arise in non-nested settings (61, 62). The interconversion of p-values and posterior odds (or equivalently FWER and FDR) should then behave quite differently.

Doublethink is closely related to recent developments in combined hypothesis tests that exploit heavy tailed distributions such as the Cauchy combination test (63) and the harmonic mean p-value (HMP; 64). The HMP provides a model-averaging approach, starting with p-values, whereas Doublethink pursues joint Bayesian/frequentist model-averaging beginning with nested maximized likelihood ratios. These are closely related, but a theoretical advance over the HMP is the ability of Doublethink to average over uncertainty in the null hypothesis, as well as the alternative hypothesis. An interesting result from Doublethink is that under the null hypothesis, the model-averaged deviance asymptotically follows a chi-squared distribution with one degree of freedom. This mirroring of the null distribution of the classical likelihood ratio test statistic emerges from the self-similarity or 'fractal' property of sums of heavy tailed random variables. Moreover, the model-averaged deviance could be interpreted instead of the posterior odds or Bayes factor, which are strongly influenced by the prior (17).

Doublethink, like the HMP, enables us to reconsider established positions concerning the philosophy and practice of hypothesis testing. In particular, the multilevel nature of these tests, in which all possible combinations of hypotheses are simultaneously controlled via pre-determined thresholds, like (65), supports refinements to concepts like fishing for significance, data dredging and p-hacking (66). For a fixed set of predetermined null hypotheses, Doublethink allows us to search in arbitrary ways for significant groups of variables, without impacting the FDR and, at least asymptotically, the strong-sense FWER. Since exhaustive searches are not generally practicable, the methods by which signals are sought through grouping variables become important. We examined just two possible methods of grouping variables, but the strong impact of the groupings on the relative prominence of signals in the data means more work is required in this area. From these insights and through new avenues of research, this work has the potential to help advance scientific discovery and bridge the differences between Bayesian and classical hypothesis testing.

Acknowledgements

The authors wish to thank Naomi Allen, Jeff Chen, Steven Lin, Gil McVean, Tim Peto and Sarah Walker for comments and advice, and the Mathematisches Forschungsinstitut Oberwolfach, organizers and participants of workshop 2308 *Design and Analysis of Infectious Disease Studies*.

Funding

This work was funded by the Robertson Foundation, the Wellcome Trust and the Royal Society (grant no. 101237/Z/13/B). This study was supported by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance (NIHR200915), a partnership between the UK Health Security Agency (UKHSA) and the University of Oxford. The views expressed are those of the author(s) and not necessarily those of the NIHR, UKHSA or the Department of Health and Social Care. The research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Supplementary Material

Table S1 Literature review of published papers analysing COVID-19 outcomes in UK Biobank, containing fields from Web of Science and curated with information on the type of analysis, exclusion criteria, and outcomes and exposures included in the abstract and analysis.

Table S2 All UK Biobank fields included in the analysis, annotated by field or ICD-10 code, UK Biobank or ICD-10 description, level (if a factor) and numeric encoding in R.

Table S3 Synonyms and categories of variables used in the interpretation of the literature review of published papers analysing COVID-19 outcomes in UK Biobank.

Table S4 Categories applied to results in Table 1, prior groupings, for comparison to the literature review

Table S5 Categories applied to results in Table 2, post hoc groupings, for comparison to the literature review

Literature cited

1. Sudlow, C., et al. (2015) 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.
2. Tan, V. Y., Timpson, N. J. (2022) 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.
3. Fachal, L., Dunning, A. M. (2015) From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. *Curr Opin Genet Dev* 30: 32-41.
4. Marigorta, U. M., et al. (2018) Replicability and prediction: lessons and challenges from GWAS. *Trends Genet* 34(7): 504–517.
5. Duncan, L. E., Ostacher, M., & Ballon, J. (2019) How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology* 44(9): 1518–1523.
6. Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.
7. Vandenbroucke, J. P. (2002). The history of confounding. *Sozial-und Präventivmedizin*, *47*, 216-224.
8. Westreich, D., & Greenland, S. (2013). The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American journal of epidemiology*, 177(4), 292-298.
9. Griffith, G. J., Morris, T. T., Tudball, M. J., Herbert, A., Mancano, G., Pike, L., ... & Hemani, G. (2020). Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature communications*, *11*(1), 5749.
10. Ding, E., Wang, Y., Liu, J., Tang, S., & Shi, X. (2022). A review on the application of the exposome paradigm to unveil the environmental determinants of age-related diseases. *Human Genomics*, *16*(1), 1-16.
11. Madakkatel, I., Zhou, A., McDonnell, M. D., & Hyppönen, E. (2021). Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study. *Scientific reports*, *11*(1), 22997.
12. Wan, T. K., Huang, R. X., Tulu, T. W., Liu, J. D., Vodencarevic, A., Wong, C. W., & Chan, K. H. K. (2022). Identifying predictors of COVID-19 mortality using machine learning. *Life*, *12*(4), 547.
13. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, *15*(11), e1002683.
14. House of Commons Science, Innovation and Technology Committee (2023) The governance of artificial intelligence: interim report. Ninth Report of Session 2022–23. HC 1769. https://committees.parliament.uk/publications/41130/documents/205611/default/
15. Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111-163.
16. Gelman, A. (2008). Objections to Bayesian statistics.
17. Gelman A, Shalizi CR (2012) Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology* 66:8-38.
18. Fryer, H. R., Arning, N., & Wilson, D. J. (2023). Doublethink: simultaneous Bayesian-frequentist model-averaged hypothesis testing. *arXiv* doi: 10.48550/arXiv.2312.17566
19. Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, *96*(456), 1151-1160.
20. Marcus, R., Eric, P., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, *63*(3), 655-660.

21. Zellner A (1986) 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*, eds. Goel PK, Zellner A, Amsterdam: North-Holland/Elsevier, pp. 233–243.
22. Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of *g* priors for Bayesian variable selection. *J Am Stat Soc* 103:410-423.
23. Johnson VE (2005) Bayes factors based on test statistics. *J R Stat Soc B* 67:689-701.
24. Johnson VE (2008) Properties of Bayes factors based on test statistics. *Scand J Stat* 35:354-368.
25. Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann Math Stat* 9:60–62.
26. Schwarz G, et al. (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.
27. Kass RE, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J Am Stat Soc* 90:928-934.
28. Karamata J (1933) Sur un mode de croissance régulière. théorèmes fondamentaux. *Bull Soc Math France* 61:55–62.
29. Davis, R. A., & Resnick, S. I. (1996). Limit theory for bilinear processes with heavy-tailed noise. *The Annals of Applied Probability*, *6*(4), 1191-1210.
30. Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications.
31. R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
32. Anaconda Inc (2023). Python. https://www.python.org
33. Hu J, Johnson VE (2009) Bayesian model selection using test statistics. *J R Stat Soc B* 71:143-158.
34. Armstrong, J., Rudkin, J. K., Allen, N., Crook, D. W., Wilson, D. J., Wyllie, D. H., & O'Connell, A. M. (2020). Dynamic linkage of COVID-19 test results between Public Health England's second generation surveillance system and UK Biobank. *Microbial genomics*, *6*(7).
35. UK Biobank (2023) Hospital inpatient data. Version 4.0. https://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=138483
36. UK Biobank (2023) Mortality data: linkage to death registries. Version 3.0. https://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=115559
37. Lee, S. F., Nikšić, M., Rachet, B., Sanchez, M.-J., & Luque-Fernandez, M. A. (2021). Socioeconomic inequalities and ethnicity are associated with a positive COVID-19 test among cancer patients in the UK Biobank cohort. *Cancers*, *13*(7), 1514. doi:10.3390/cancers13071514
38. Elliott, J., Bodinier, B., Whitaker, M., Delpierre, C., Vermeulen, R., Tzoulaki, I., … Chadeau-Hyam, M. (2021). COVID-19 mortality in the UK Biobank cohort: revisiting and evaluating risk factors. *European Journal of Epidemiology*, *36*(3), 299–309. doi:10.1007/s10654-021-00722-y
39. Hamer, M., Kivimäki, M., Gale, C. R., & Batty, G. D. (2020). Lifestyle risk factors, inflammatory mechanisms, and COVID-19 hospitalization: A community-based cohort study of 387,109 adults in UK. *Brain, Behavior, and Immunity*, *87*, 184–187. doi:10.1016/j.bbi.2020.05.059
40. Prats-Uribe, A., Xie, J., Prieto-Alhambra, D., & Petersen, I. (2021). Smoking and COVID-19 infection and related mortality: A prospective cohort analysis of UK Biobank data. *Clinical Epidemiology*, *13*, 357–365. doi:10.2147/CLEP.S300597
41. Clift, A. K., von Ende, A., Tan, P. S., Sallis, H. M., Lindson, N., Coupland, C. A. C., … Hopewell, J. C. (2022). Smoking and COVID-19 outcomes: an observational and Mendelian

randomisation study using the UK Biobank cohort. *Thorax*, *77*(1), 65–73. doi:10.1136/thoraxjnl-2021-217080

42. Didikoglu, A., Maharani, A., Pendleton, N., Canal, M. M., & Payton, A. (2021). Early life factors and COVID-19 infection in England: A prospective analysis of UK Biobank participants. *Early Human Development*, *155*(105326), 105326. doi:10.1016/j.earlhumdev.2021.105326
43. Atkins, J. L., Masoli, J. A. H., Delgado, J., Pilling, L. C., Kuo, C.-L., Kuchel, G. A., & Melzer, D. (2020). Preexisting comorbidities predicting COVID-19 and mortality in the UK Biobank community cohort. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, *75*(11), 2224–2230. doi:10.1093/gerona/glaa183
44. Gao, M., Wang, Q., Piernas, C., Astbury, N. M., Jebb, S. A., Holmes, M. V., & Aveyard, P. (2022). Associations between body composition, fat distribution and metabolic consequences of excess adiposity with severe COVID-19 outcomes: observational study and Mendelian randomisation analysis. *International Journal of Obesity (2005)* , *46*(5), 943–950. doi:10.1038/s41366-021-01054-3
45. Wong, K. C.-Y., Xiang, Y., Yin, L., & So, H.-C. (2021). Uncovering clinical risk factors and predicting severe COVID-19 cases using UK Biobank data: Machine learning approach. *JMIR Public Health and Surveillance*, *7*(9), e29544. doi:10.2196/29544
46. Zhu, Z., Hasegawa, K., Ma, B., Fujiogi, M., Camargo, C. A., Jr, & Liang, L. (2020a). Association of asthma and its genetic predisposition with the risk of severe COVID-19. *The Journal of Allergy and Clinical Immunology*, 146(2), 327-329.e4. doi:10.1016/j.jaci.2020.06.001
47. Lodge, C. J., Doherty, A., Bui, D. S., Cassim, R., Lowe, A. J., Agusti, A., … Dharmage, S. C. (2021). Is asthma associated with COVID-19 infection? A UK Biobank analysis. *ERJ Open Research*, *7*(4), 00309–02021. doi:10.1183/23120541.00309-2021
48. Ma, H., Zhou, T., Heianza, Y., & Qi, L. (2021). Habitual use of vitamin D supplements and risk of coronavirus disease 2019 (COVID-19) infection: a prospective study in UK Biobank. *The American Journal of Clinical Nutrition*, *113*(5), 1275–1281. doi:10.1093/ajcn/nqaa381
49. Li, S., Cao, Z., Yang, H., Zhang, Y., Xu, F., & Wang, Y. (2021). Metabolic healthy obesity, vitamin D status, and risk of COVID-19. *Aging and Disease*, *12*(1), 61–71. doi:10.14336/AD.2020.1108
50. Dabbah, M. A., Reed, A. B., Booth, A. T. C., Yassaee, A., Despotovic, A., Klasmer, B., … Mohan, D. (2021). Machine learning approach to dynamic risk modeling of mortality in COVID-19: a UK Biobank study. *Scientific Reports*, *11*(1), 16936. doi:10.1038/s41598-021-95136-x
51. Xiang, Y., Wong, K. C.-Y., & So, H.-C. (2021). 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*(9), 1514. doi:10.3390/pharmaceutics13091514
52. Tangirala, S., Tierney, B. T., & Patel, C. J. (2023). Prioritization of COVID-19 risk factors in July 2020 and February 2021 in the UK. *Communications Medicine*, *3*(1), 45.
53. Córdova-Palomera, A., Siffel, C., DeBoever, C., Wong, E., Diogo, D., & Szalma, S. (2023). Assessing the potential of polygenic scores to strengthen medical risk prediction models of COVID-19. *Plos one*, *18*(5), e0285991.
54. Grafen, A., & Hails, R. (2002). *Modern statistics for the life sciences*. Oxford University Press.
55. Wagenmakers, E. J., & Ly, A. (2023). History and nature of the Jeffreys–Lindley paradox. *Archive for History of Exact Sciences*, *77*(1), 25-72.

56. Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Johnson, V. E. (2018). Redefine statistical significance. *Nature human behaviour*, *2*(1), 6-10.
57. O'Hagan A (1995) Fractional Bayes factors for model comparison. *J R Stat Soc B* 57:99-138.
58. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.
59. Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The annals of statistics*, *31*(6), 2013-2035.
60. Mikosch T (1999) *Regular Variation, Subexponentiality and Their Applications in Probability Theory* (Eindhoven University of Technology, Eindhoven, The Netherlands), Vol 99.
61. Held, L. (2019). On the Bayesian interpretation of the harmonic mean p-value. *Proceedings of the National Academy of Sciences*, *116*(13), 5855-5856.
62. Wilson, D. J. (2019). Reply to Held: When is a harmonic mean p-value a Bayes factor? *Proceedings of the National Academy of Sciences*, *116*(13), 5857-5858.
63. Liu, Y., & Xie, J. (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, *115*(529), 393-402.
64. Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, *116*(4), 1195-1200.
65. Goeman, J. J., & Solari, A. (2011). Multiple testing for exploratory research.
66. Andrade, C. (2021). HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *The Journal of clinical psychiatry*, *82*(1), 25941.

Tables

	Group *PP* (%)	Group -$\log_{10}$ p^*	Variable	*PP* (%)	-$\log_{10}$ p^*	Effect size when included	Standard error when included
1	100	>5.95	34 Year of birth (years)	40.8	2.05	-0.03	0.00
			21003 Age when attended assessment centre (years)	30.0	1.82	0.03	0.10
			21022 Age at recruitment (years)	29.2	1.80	0.03	0.10
2	100	>5.95	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
			189 Townsend deprivation index at recruitment	2.7	-	0.05	0.02
			26414 Education score (England)	0.3	-	0.01	0.00
			26411 Income score (England)	0.3	-	1.70	0.34
			26417 Living environment score (England)	0.0	-	0.00	0.00
			26416 Crime score (England)	0.0	-	0.01	0.04
3	100	5.95	31 Sex : 0 : Female	50.1	2.24	-0.49	0.11
			31 Sex : 1 : Male	49.9	2.23	0.49	0.11
4	100	5.80	41214 Carer support indicators : 1 : Yes	100	5.78	0.56	0.08
			54 UK Biobank assessment centre : 11006 : Stoke	0.0	-	0.24	0.21
5	99.9	5.70	6138 Qualifications : 3 : O levels/GCSEs or equivalent	96.7	3.82	-0.29	0.05
			6138 Qualifications : 1 : College or University degree	96.5	3.80	-0.34	0.06
			6138 Qualifications : 2 : A levels/AS levels or equivalent	0.1	-	-0.29	0.09
-	99.9	5.44	Z86.4 Personal history of psychoactive substance abuse	99.9	5.44	0.37	0.06
6	99.7	5.01	F03 Unspecified dementia	99.7	5.00	0.94	0.15
			G30.9 Alzheimer's disease, unspecified	0.0	-	0.57	0.23
			F00.9 Dementia in Alzheimer's disease, unspecified	0.0	-	0.59	0.26
7	99.7	4.89	137 Number of treatments/medications taken	99.7	4.89	0.06	0.01
			135 Number of self-reported non-cancer illnesses	0.0	-	0.02	0.01
-	99.6	4.82	J22 Unspecified acute lower respiratory infection	99.6	4.82	0.51	0.09
8	99.5	4.71	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
			Z50.5 Speech therapy	0.0	-	0.34	0.22
-	99.2	4.47	R29.6 Tendency to fall, not elsewhere classified	99.2	4.47	0.63	0.11
9	98.2	4.10	41218 History of psychiatric care on admission : 8 : Not applicable	98.2	4.10	0.49	0.09
			41214 Carer support indicators : 99 : Not known	0.0	-	0.03	0.19
10	89.0	3.22	48 Waist circumference (cm)	89.0	3.22	0.02	0.00
			21002 Weight (Kg)	0.0	-	0.01	0.00
			23098 Weight (Kg)	0.0	-	0.00	0.00
-	84.1	3.02	J18.1 Lobar pneumonia, unspecified	84.1	3.02	0.49	0.09
11	80.0	2.89	21000 Ethnic background : 1001 : British	70.2	2.64	-0.40	0.08
			1647 Country of birth (UK/elsewhere) : 6 : Elsewhere	9.8	-	0.42	0.09
			1647 Country of birth (UK/elsewhere) : 1 : England	0.0	-	-0.18	0.07
			21000 Ethnic background : 1003 : Any other white background	0.0	-	-0.21	0.14
-	66.1	2.55	K59.0 Constipation	66.1	2.55	0.40	0.08
12	58.4	2.40	N39.0 Urinary tract infection, site not specified	58.4	2.40	0.40	0.08
			B96.2 Escherichia coli [E. coli] as the cause of diseases classified to other chapters	0.0	-	0.32	0.12
13	40.1	2.04	3063 Forced expiratory volume in 1-second (FEV1) (litres)	28.1	1.77	-0.19	0.04
			3062 Forced vital capacity (FVC) (litres)	12.1	-	-0.15	0.03
			3064 Peak expiratory flow (PEF) (litres/min)	0.0	-	0.00	0.00
14	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
15	38.3	2.00	24017 Nitrogen dioxide air pollution; 2006 (micro-g/m3)	15.5	-	0.01	0.00
			24018 Nitrogen dioxide air pollution; 2007 (micro-g/m3)	15.1	-	0.01	0.00
			24016 Nitrogen dioxide air pollution; 2005 (micro-g/m3)	7.7	-	0.01	0.00
-	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

Table 1 Doublethink allows the interconversion of model-averaged posterior odds and p*-values for groups of variables, defined here* a priori *using variable correlation. The most significant groups are shown, alongside details of constituent variables. The most significant individual, ungrouped, variables are also shown.* PP*: posterior probability.* p**: adjusted* p*-value (only values below* $10^{-1.71}$ *are shown).*

	Group *PP* (%)	Group -$\log_{10}$ p^*	Variable	*PP* (%)	-$\log_{10}$ p^*	Effect size when included	Standard error when included
A	100	>5.95	34 Year of birth (years)	40.8	2.05	-0.03	0.00
			21003 Age when attended assessment centre (years)	30.0	1.82	0.03	0.10
			21022 Age at recruitment (years)	29.2	1.80	0.03	0.10
B	100	5.95	31 Sex : 0 : Female	50.1	2.24	-0.49	0.11
			31 Sex : 1 : Male	49.9	2.23	0.49	0.11
-	100	5.78	41214 Carer support indicators : 1 : Yes	100.0	5.78	0.56	0.08
C	99.9	5.50	Z86.4 Personal history of psychoactive substance abuse	99.9	5.44	0.37	0.06
			20116 Smoking status : 0 : Never	0.0	-	-0.17	0.07
-	99.7	5.00	F03 Unspecified dementia	99.7	5.00	0.94	0.15
-	99.7	4.89	137 Number of treatments/medications taken	99.7	4.89	0.06	0.01
-	99.6	4.82	J22 Unspecified acute lower respiratory infection	99.6	4.82	0.51	0.09
D	99.5	4.75	48 Waist circumference (cm)	89.0	3.22	0.02	0.00
			21001 Body mass index (BMI) (Kg/m2)	5.5	-	0.04	0.01
			23104 Body mass index (BMI) (Kg/m2)	5.0	-	0.04	0.01
E	99.5	4.71	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	99.4	4.65	R29.6 Tendency to fall, not elsewhere classified	99.2	4.47	0.63	0.11
			W19.0 Home	2.4	-	0.59	0.15
G	99.3	4.55	41218 History of psychiatric care on admission : 8 : Not applicable	98.2	4.10	0.49	0.09
			I10 Essential (primary) hypertension	5.5	-	0.24	0.06
H	99.2	4.48	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 centre : 11011 : Bristol	3.9	-	-0.47	0.14
-	96.7	3.82	6138 Qualifications : 3 : O levels/GCSEs or equivalent	96.7	3.82	-0.29	0.05
-	96.5	3.80	6138 Qualifications : 1 : College or University degree	96.5	3.80	-0.34	0.06
I	95.7	3.69	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	94.1	3.54	21000 Ethnic background : 1001 : British	70.2	2.64	-0.40	0.08
			24017 Nitrogen dioxide air pollution; 2006 (micro-g/m3)	15.5	-	0.01	0.00
			24018 Nitrogen dioxide air pollution; 2007 (micro-g/m3)	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) (litres)	28.1	1.77	-0.19	0.04
			3062 Forced vital capacity (FVC) (litres)	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

Table 2 Doublethink allows arbitrary groups of variables to be assessed for significance while simultaneously controlling the FWER and FDR. Here groups were defined a posteriori *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**: adjusted* p*-value (only values below* $10^{-1.71}$ *are shown).*

Category	% Papers	% Abstracts	*PP* (%)	-$\log_{10}$ *p**
Age	90	11	100	>5.95
Sex	84	14	100	5.95
Obesity	78	20	99.5	4.75
Ethnicity	78	16	80	2.89
Socioeconomic status	68	13	100	>5.95
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	5.5
General comorbidity	31	9	99.7	4.89
Cancer	29	1		
Kidney disease	28	6	95.7	3.69
Educational attainment	28	4	99.9	5.7
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	5.01
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	100	5.8
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		
Haematological disease	3	1		
Lifestyle	3	2		
Gastrointestinal disease	3	2		
Sex hormones	2	2		
Periodontal disease	2	2		

Table 3 Comparison of risk factors for COVID-19 outcomes in the UK Biobank versus this study. The number 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).*