nature

https://doi.org/10.1038/s41586-020-2521-4

Accelerated Article Preview

OpenSAFELY: factors associated with COVID-19 death in 17 million patients

Received: 15 May 2020

Accepted: 1 July 2020

Accelerated Article Preview Published online 8 July 2020

Cite this article as: Williamson, E. J. et al. OpenSAFELY: factors associated with COVID-19 death in 17 million patients. *Nature* https://doi.org/10.1038/s41586-020-2521-4 (2020).

Elizabeth J. Williamson, Alex J. Walker, Krishnan Bhaskaran, Seb Bacon, Chris Bates, Caroline E. Morton, Helen J. Curtis, Amir Mehrkar, David Evans, Peter Inglesby, Jonathan Cockburn, Helen I. McDonald, Brian MacKenna, Laurie Tomlinson, Ian J. Douglas, Christopher T. Rentsch, Rohini Mathur, Angel Y. S. Wong, Richard Grieve, David Harrison, Harriet Forbes, Anna Schultze, Richard Croker, John Parry, Frank Hester, Sam Harper, Rafael Perera, Stephen J. W. Evans, Liam Smeeth & Ben Goldacre

This is a PDF file of a peer-reviewed paper that has been accepted for publication. Although unedited, the content has been subjected to preliminary formatting. Nature is providing this early version of the typeset paper as a service to our authors and readers. The text and figures will undergo copyediting and a proof review before the paper is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

ACCELERATED ARTICLE PREVIEW

OpenSAFELY: factors associated with COVID-19 death in 17 million patients

https://doi.org/10.1038/s41586-020-2521-4

Received: 15 May 2020

Accepted: 1 July 2020

Published online: 8 July 2020

Elizabeth J. Williamson[2,6], Alex J. Walker[1,6], Krishnan Bhaskaran[2,6], Seb Bacon[1,6], Chris Bates[3,6], Caroline E. Morton[1], Helen J. Curtis[1], Amir Mehrkar[1], David Evans[1], Peter Inglesby[1], Jonathan Cockburn[3], Helen I. McDonald[2,5], Brian MacKenna[1], Laurie Tomlinson[2], Ian J. Douglas[2], Christopher T. Rentsch[2], Rohini Mathur[2], Angel Y. S. Wong[2], Richard Grieve[2], David Harrison[4], Harriet Forbes[2], Anna Schultze[2], Richard Croker[1], John Parry[3], Frank Hester[3], Sam Harper[3], Rafael Perera[1], Stephen J. W. Evans[2], Liam Smeeth[2,5,7] & Ben Goldacre[1,7] ✉

COVID-19 has rapidly affected mortality worldwide[1]. There is unprecedented urgency to understand who is most at risk of severe outcomes, requiring new approaches for timely analysis of large datasets. Working on behalf of NHS England, here we created OpenSAFELY: a secure health analytics platform covering 40% of all patients in England, holding patient data within the existing data centre of a major primary care electronic health records vendor. Primary care records of 17,278,392 adults were pseudonymously linked to 10,926 COVID-19-related deaths. COVID-19-related death was associated with: being male (hazard ratio (HR) 1.59, 95% confidence interval (CI) 1.53–1.65); older age and deprivation (both with a strong gradient); diabetes; severe asthma; and various other medical conditions. Compared with people with white ethnicity, Black and South Asian people were at higher risk even after adjustment for other factors (HR 1.48, 1.30–1.69 and 1.44, 1.32–1.58, respectively). We have quantified a range of clinical risk factors for COVID-19-related death in the largest cohort study conducted by any country to date. OpenSAFELY is rapidly adding further patients' records; we will update and extend results regularly.

On March 11th 2020, the World Health Organisation characterised COVID-19 as a pandemic after 118,000 cases and 4,291 deaths were reported in 114 countries.[2] As of 6 May (the date of latest data availability for this study), cases reached over 3.5 million globally, with more than 240,000 deaths attributed to the virus.[1] On the same day in the UK, there were 206,715 confirmed cases, with 30,615 deaths.[3]

Age and gender are well-established risk factors for severe COVID-19 outcomes, with over 90% of UK deaths being in people over 60, and 60% in men[4]. Various pre-existing conditions have also been associated with increased risk. For example, the Chinese center for disease control and prevention (44,672 patients, 1,023 deaths) reported cardiovascular disease, hypertension, diabetes, respiratory disease, and cancers as associated with increased risk of death[5], but correction for relationships with age was not possible. A UK cross-sectional survey describing 16,749 patients hospitalised with COVID-19 showed higher risk of death for patients with cardiac, pulmonary and kidney disease, as well as malignancy, dementia and obesity (hazard ratios 1.19-1.39 after age and sex correction).[6] Obesity was associated with treatment escalation in a French ITU cohort (n=124) and a New York hospital presentation cohort (n=3615).[7,8] Risks associated with smoking are unclear.[9,10,11] People from black and minority ethnic (BME) groups are at increased risk of bad outcomes from COVID-19, for reasons that are unclear.[12,13]

Patient care is typically managed through electronic health records (EHR) which are commonly used in research. However traditional approaches to EHR analysis rely on intermittent extracts of small samples of historic data. Evaluating a rapidly arising novel cause of death requires a new approach. We therefore set out to deliver a secure analytics platform inside the data centre of major electronic health records vendors, running across the full live linked pseudonymised electronic health records of a very large population of NHS patients, to determine factors associated with COVID-19 related death in England (referred to as "death" in text that follows).

Results

17,278,392 adults were included (Figure 1; cohort description in Table 1). 1,851,868 (11%) individuals had non-white ethnicities recorded. There were missing data for body mass index (3,751,769, 22%), smoking status (720,923, 4%), ethnicity (4,560,113, 26%), and blood pressure (1,715,095, 10%). 10,926 of the study population had COVID-19 related death recorded in linked death registration data.

The overall cumulative incidence of death 90 days after study start was <0.01% in those aged 18-39 years, rising to 0.67% and 0.44% in men and women respectively aged ≥80 years (Figure 2).

Associations between patient-level factors and risk of death are shown in Table 2 and Figure 3. Increasing age was strongly associated with risk, with those ≥80 years having more than 20-fold increased risk than 50-59 year olds (fully adjusted HR 20.61; 95% CI 18.72-22.70).

[1]The DataLab, Nuffield Department of Primary Care Health Sciences, University of Oxford, OX26GG, Oxford, UK. [2]London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK. [3]TPP, TPP House, 129 Low Lane, Horsforth, Leeds, LS18 5PX, UK. [4]ICNARC, 24 High Holborn, Holborn, London, WC1V 6AZ, UK. [5]NIHR Health Protection Research Unit (HPRU) in Immunisation, London, UK. [6]These authors contributed equally: Elizabeth J. Williamson, Alex J. Walker, Krishnan Bhaskaran, Seb Bacon, Chris Bates. [7]These authors jointly supervised this work: Liam Smeeth, Ben Goldacre. ✉e-mail: ben.goldacre@phc.ox.ac.uk

ACCELERATED ARTICLE PREVIEW

With age fitted as a flexible spline, an approximately log-linear relationship was observed (Extended Data Figure 1). Men had higher risk than women (fully adjusted HR 1.59, 1.53-1.65). These findings are consistent with patterns observed in smaller studies worldwide and in the UK.[14]

All non-white ethnic groups had higher risk than those with white ethnicity: HRs adjusted for age and sex only ranged from 1.62-1.88 for Black, South Asian and mixed ethnicities compared to white; attenuated to 1.43-1.48 on adjustment for all included risk factors (results for more detailed categories are shown in Extended Data Table 1). Non-white ethnicity has previously been found to be associated with increased COVID-19 infection and poor outcomes.[12,13,15] Our findings show that only a small part of the excess risk is explained by higher prevalence of medical problems such as cardiovascular disease or diabetes among BME people, or higher deprivation.

We found a consistent pattern of increasing risk with greater deprivation, with the most deprived quintile having a HR of 1.80 compared to the least deprived, consistent with recent national statistics.[16] Again, very little of this increased risk was explained by pre-existing disease or clinical risk factors, suggesting that other social factors may have an important role.

Increasing risks were seen with increasing obesity (BMI >40 fully adjusted HR 1.92, 95% CI 1.72-2.13), and most comorbidities were associated with higher risk of death, including diabetes (with a greater HR for those with recent HbA1c >= 58 mmol/mol), severe asthma (defined as asthma with recent use of an oral corticosteroid), respiratory disease, chronic heart disease, liver disease, stroke/dementia, other neurological diseases, reduced kidney function (with greater HR for lower estimated glomerular filtration rate), autoimmune diseases (rheumatoid arthritis, lupus or psoriasis) and other immunosuppressive conditions, as per Table 2. Those with a recent (<5 years) history of haematological malignancy had a ≥2.5-fold increased risk, decreasing slightly after 5 years. For other cancers, increased HRs were smaller and mainly with recent diagnoses. History of dialysis or end-stage renal failure was associated with increased risk when added in a secondary analysis (HR 3.69, 3.10-4.39). These findings largely concurred with other data including the UK ISARIC study of hospitalised UK patients with COVID-19 that indicated increased risk of death with cardiac, pulmonary and kidney disease, malignancy, obesity and dementia,[6] and a large Chinese study which, though lacking age correction, suggested cardiovascular disease, hypertension, diabetes, respiratory disease, and cancers to be associated with increased mortality.[5] Our findings that severe asthma was associated with higher risk were notable since early data suggested underrepresentation of asthma in patients hospitalised or with severe COVID-19 outcomes[17,18]

Post-hoc analyses: smoking and hypertension

Both current and former smoking were associated with higher risk in models adjusted for age and sex only, but in the fully adjusted model current smoking was associated with a lower risk (fully adjusted HR 0.89, CI 0.82-0.97), concurring with lower than expected smoking prevalences in previous studies among hospitalised patients in China,[10] France[11] and the USA.[19] We further explored this post-hoc by adding covariates individually to the age, sex and smoking model, and found the change in HR to be largely driven by adjustment for chronic respiratory disease (HR 0.98, 0.90-1.06 after adjustment). This and other comorbidities could be consequences of smoking, highlighting that the fully adjusted smoking HR cannot be interpreted causally due to the inclusion of factors likely to mediate smoking effects. We therefore then fitted a model adjusted for demographic factors only (age, sex, deprivation, ethnicity), which showed a non-significant positive HR for current smoking (HR 1.07, 0.98-1.18). This does not support any postulated protective effect of nicotine[9,20] but suggests that any increased risk with current smoking is likely to be small, and will need to be clarified as the epidemic progresses and more data accumulate.

We similarly explored the change in the hypertension HR (from 1.09, 1.05-1.14 adjusted for age and sex to 0.89, 0.85-0.93 with all covariates included), and found diabetes and obesity to be principally responsible for this reduction (HR 0.97, 0.92-1.01 adjusted for age, sex, diabetes, obesity). Given the strong association between blood pressure and age we then examined an interaction between these variables; this revealed strong evidence of interaction (p<0.001) with hypertension associated with higher risk up to age 70 years and lower risk at older ages (adjusted HRs 3.11 [1.68-5.71], 2.75 [1.97-3.83], 2.07 [1.73-2.47], 1.32 [1.17-1.50], 0.94 [0.86-1.02], 0.73 [0.69-0.78] for ages 18-<40, 40-<50, 50-<60, 60-<70, 70-<80 and ≥80 respectively). The reasons for the inverse association between hypertension and mortality in older individuals are unclear and warrant further investigation including detailed examination by frailty, comorbidity and drug exposures in this age group.

Model checking and sensitivity analyses

The average C-statistic was 0.77. Results were similar when missing data were handled using analysis of complete records only, or using multiple imputation (sensitivity analyses: Extended Data Table 2). Non-proportional hazards were detected in the primary model (p<0.001). A sensitivity analysis with earlier administrative censoring at 6th April 2020, before which mortality should not have been affected by UK social distancing policies introduced in late March, showed no evidence of non-proportional hazards (p=0.83). HRs were similar but somewhat larger in magnitude for some covariates, while the association with increasing deprivation appeared to be smaller (Extended Data Table 2).

Discussion

This secure analytics platform operating across over 23 million patient records for the COVID-19 emergency was used to identify, quantify, and explore risk factors for COVID-19 related death in the largest cohort study conducted by any country to date. Most comorbidities were associated with increased risk, including cardiovascular disease, diabetes, respiratory disease including severe asthma, obesity, history of haematological malignancy or recent other cancer, kidney, liver, neurological and autoimmune conditions. People from South Asian and black groups had a substantially higher risk of death, only partially attributable to co-morbidity, deprivation or other risk factors. A strong association between deprivation and risk was only partly attributable to co-morbidity or other risk factors.

These analyses provide a preliminary picture of how key demographic characteristics and a range of comorbidities, a priori selected as being of interest in COVID-19, are jointly associated with poor outcomes. These initial results may be used subsequently to inform the development of prognostic models. We caution against interpreting our estimates as causal effects. For example, the fully adjusted smoking hazard ratio does not capture the causal effect of smoking due to the inclusion of comorbidities which are likely to mediate any effect of smoking on COVID-19 death (e.g. COPD). Our study has highlighted a need for carefully designed causal analyses specifically focusing on the causal effect of smoking on COVID-19 death. Similarly, there is a need for analyses exploring the causal relationships underlying the associations observed between hypertension and COVID-19 death.

Strengths and weaknesses

The greatest strengths of this study were speed and size. By building a secure analytics platform across routinely collected live clinical data stored in situ we have produced timely results from the current records of approximately 40% of the English population. This scale allows more precision, on rarer exposures, on multiple risk factors, and rapid detection of important signals. Our platform will expand to provide updated

ACCELERATED ARTICLE PREVIEW

analyses over time. Another strength is our use of open methods: we pre-specified our analysis plan and shared our full analytic code and code lists for review and re-use. We ascertained demographics, medications and co-morbidities from full pseudonymised longitudinal primary care records, providing substantially more detail than data recorded on admission, and on the total population rather than the selected subset presenting at hospital. We censored deaths from other causes using ONS data. Analyses were stratified by area to account for known geographical differences in incidence of COVID-19.

We also identify important limitations. In our outcome definition, we included clinically suspected (non laboratory confirmed) COVID-19, because testing has not always been carried out, especially in older patients in care homes. However, this may have incorrectly identified some patients as having COVID-19. Some COVID-19 deaths may have been misclassified as non-COVID-19, particularly in the early stages of the pandemic, though this is likely to have reduced quickly as deaths accumulated, and a degree of outcome underascertainment, providing unrelated to patient characteristics, should not have biased our hazard ratios. Due to the rarity of the outcome, the associations observed will be driven primarily by the profile of risk factors in the included cases. Our findings reflect both an individual's risk of infection, and their risk of dying once infected. We will explore more detailed patient trajectories in future research within the OpenSAFELY platform.

Our large population may not be fully representative. We include only 17% of general practices in London, where many earlier COVID-19 cases occurred, due to the substantial geographic variation in choice of EHR system.. The user interface of electronic health records can affect prescribing of certain medicines[21–23] so it is possible that coding may vary between systems.

Primary care records, though detailed and longitudinal, can be incomplete for data on risk factors and other covariates. Ethnicity was missing for approximately 26%, but was broadly representative;[24] there were also missing data on obesity and smoking. Sensitivity analyses found our estimates were robust to our assumptions around missing data.

Non-proportional hazards could be due to very large numbers or unmeasured covariates. However, rapid changes in social behaviours (social distancing, shielding) and changes in the burden of infection may also have affected patient groups differentially. The larger hazard ratios seen for several covariates in a sensitivity analysis with earlier censoring (soon after social distancing and shielding policies were introduced) are consistent with more at-risk patients being more compliant with these policies. In contrast, the risk associated with deprivation may have increased over time. Subsequent analyses will further explore changes before and after national initiatives around COVID-19.

Policy Implications and Interpretation

The UK has a policy of recommending shielding (staying at home at all times and avoiding any face to face contact) for groups identified as being extremely vulnerable to COVID-19 on the basis of pre-existing medical conditions.[25] We were able to evaluate the association between most of these conditions and death from COVID-19, and confirmed increased mortality risks, supporting the targeted use of additional protection measures for people in these groups. We have demonstrated - for the first time - that only a small part of the substantially increased risks of COVID-19 related death among non-white groups and among people living in more deprived areas can be attributed to existing disease. Improved strategies to protect people in these groups are urgently needed.[26] These might include specific consideration of BME groups in shielding guidelines and work-place policies. Subsequent studies are needed to investigate the interplay of additional factors we were unable to explore, including employment, access to personal protective equipment and related risk of exposure to infection and household density.

The UK has an unusually large volume of very detailed longitudinal patient data, especially through primary care. We believe the UK has a responsibility to the global community to make good use of such data. OpenSAFELY demonstrates at an unprecedented scale that this can be done securely, transparently, and rapidly. We will enhance the OpenSAFELY platform to further inform the global response to the COVID-19 emergency.

Future Research

The underlying causes of higher risk of COVID-19 related death among those from non-white backgrounds, and deprived areas, require further exploration; we would suggest collecting data on occupational exposure and living conditions as first steps. The statistical power offered by our approach means that associations with less common risk factors can be robustly assessed in more detail, at the earliest possible date, as the pandemic progresses. We will therefore update our findings and address smaller risk groups as new cases arise over time. The open source reusable codebase on OpenSAFELY supports rapid, secure and collaborative development of new analyses: we are currently conducting expedited studies on the impact of various medical treatments and population interventions on the risk of COVID-19 infection, ITU admission, and death, alongside other observational analyses. OpenSAFELY is rapidly scalable for additional NHS patients' records, with new data sources progressing.

Conclusion

We generated early insights into risk factors for COVID-19 related death using an unprecedented scale of 17 million patients' detailed primary care records, maintaining privacy, in the context of a global health emergency.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2521-4.

1. COVID-19 situation reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.
2. WHO. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://web.archive.org/web/20200502133342/https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19–11-march-2020 (2020).
3. UK Government. Number of coronavirus (COVID-19) cases and risk in the UK. https://web.archive.org/web/20200501084711/https://www.gov.uk/guidance/coronavirus-covid-19-information-for-the-public (2020).
4. NHS England. COVID-19 Daily Deaths. https://web.archive.org/web/20200501094237/https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-daily-deaths/ (2020).
5. Deng, G., Yin, M., Chen, X. & Zeng, F. Clinical determinants for fatality of 44,672 patients with COVID-19. *Crit. Care* **24**, (2020).
6. Docherty, A. B. *et al*. Features of 16,749 hospitalised UK patients with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol. *medRxiv* (2020) https://doi.org/10.1101/2020.04.23.20076042.
7. Simonnet, A. *et al*. High prevalence of obesity in severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) requiring invasive mechanical ventilation. *Obesity* (2020) https://doi.org/10.1002/oby.22831.
8. Lighter, J. *et al*. Obesity in patients younger than 60 years is a risk factor for Covid-19 hospital admission. *Clin. Infect. Dis*. https://doi.org/10.1093/cid/ciaa415.
9. Simons, D., Shahab, L., Brown, J. & Perski, O. The association of smoking status with SARS-CoV-2 infection, hospitalisation and mortality from COVID-19: A living rapid evidence review. *Qeios* (2020) https://doi.org/10.32388/UJR2AW.2.
10. Guan, W.-J. *et al*. Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med*. (2020) https://doi.org/10.1056/NEJMoa2002032.
11. Miyara, M. *et al*. Low incidence of daily active tobacco smoking in patients with symptomatic COVID-19. *Qeios* (2020) https://doi.org/10.32388/WPP19W.3.
12. Khunti, K., Singh, A. K., Pareek, M. & Hanif, W. Is ethnicity linked to incidence or outcomes of covid-19? *BMJ* **369**, m1548 (2020).

ACCELERATED ARTICLE PREVIEW

13. The Institute for Fiscal Studies. Are some ethnic groups more vulnerable to COVID-19 than others? https://web.archive.org/web/20200502130148/https://www.ifs.org.uk/inequality/chapter/are-some-ethnic-groups-more-vulnerable-to-covid-19-than-others/ (2020).
14. Public Health England. *Disparities in the risk and outcomes from COVID-19*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/889195/disparities_review.pdf (2020).
15. ICNARC. COVID-19 Report. https://web.archive.org/web/20200425133758/https://www.icnarc.org/Our-Audit/Audits/Cmp/Reports (2020).
16. ONS. Deaths registered weekly in England and Wales, provisional: week ending 17 April 2020. https://web.archive.org/web/20200430191844/https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregisteredweeklyinenglandandwalesprovisional/weekending17april2020 (2020).
17. Halpin, D. M. G., Faner, R., Sibila, O., Badia, J. R. & Agusti, A. Do chronic respiratory diseases or their treatment affect the risk of SARS-CoV-2 infection? *Lancet Respir Med* (2020) https://doi.org/10.1016/S2213-2600(20)30167-3.
18. Boddington, N. L. *et al*. COVID-19 in Great Britain: epidemiological and clinical characteristics of the first few hundred (FF100) cases: a descriptive study. *{In preparation}* (2020).
19. Rentsch, C. T. *et al*. Covid-19 Testing, Hospital Admission, and Intensive Care Among 2,026,227 United States Veterans Aged 54-75 Years. *medRxiv* (2020) https://doi.org/10.1101/2020.04.09.20059964.
20. Farsalinos, K., Barbouni, A. & Niaura, R. Smoking, vaping and hospitalization for COVID-19. *Qeios* (2020) https://doi.org/10.32388/Z69O8A.13.
21. MacKenna, B. *et al*. Impact Of Electronic Health Record Interface Design On Unsafe Prescribing Of Ciclosporin, Tacrolimus and Diltiazem: A Cohort Study In English NHS Primary Care. *JMIR Preprints: Accepted for publication - in production* https://preprints.jmir.org/preprint/17003.
22. Opondo, D. *et al*. Quality of Co-Prescribing NSAID and Gastroprotective Medications for Elders in The Netherlands and Its Association with the Electronic Medical Record. *PLoS One* **10**, e0129515 (2015).
23. MacKenna, B. Ghost branded generics: Why does the cost of generic atorvastatin vary? *EBM DataLab* https://web.archive.org/web/20200502135915/https://ebmdatalab.net/ghost-branded-generics-why-does-the-cost-of-generic-atorvastatin-vary%ef%bb%bf/ (2018).
24. Mathur, R. *et al*. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *J. Public Health* **36**, 684–692 (2014).
25. Public Health England. Guidance on shielding and protecting people who are clinically extremely vulnerable from COVID-19. https://web.archive.org/web/20200501090127/https://www.gov.uk/government/publications/guidance-on-shielding-and-protecting-extremely-vulnerable-persons-from-covid-19/guidance-on-shielding-and-protecting-extremely-vulnerable-persons-from-covid-19 (2020).
26. Marmot, M. *et al*. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet* **372**, 1661–1669 (2008).

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

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

ACCELERATED ARTICLE PREVIEW

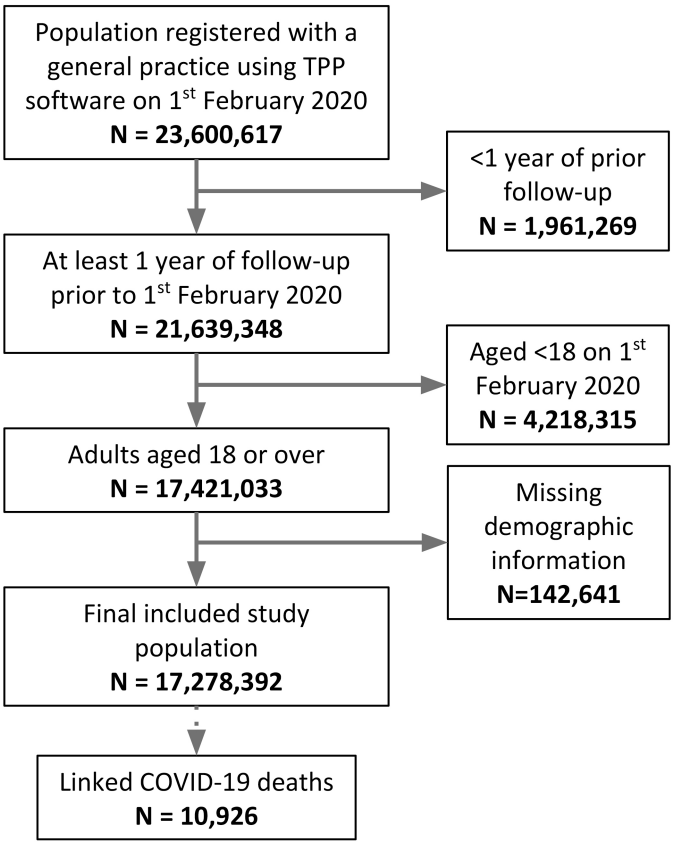


Fig. 1 | Flow diagram of cohort with numbers excluded at different stages and identification of cases for the main endpoints.

ACCELERATED ARTICLE PREVIEW

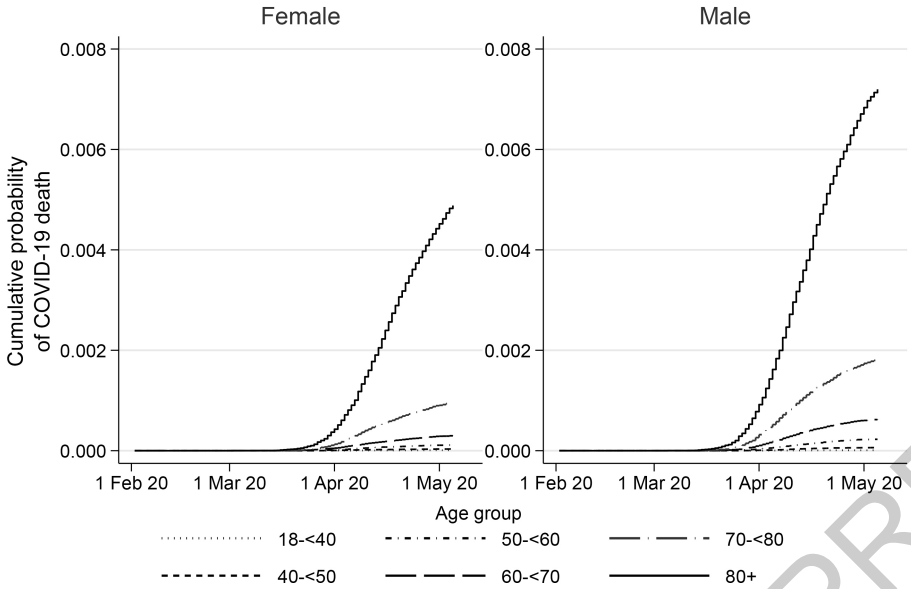


Fig. 2 | Kaplan-Meier plots for COVID-19 related death over time by age and sex.

ACCELERATED ARTICLE PREVIEW

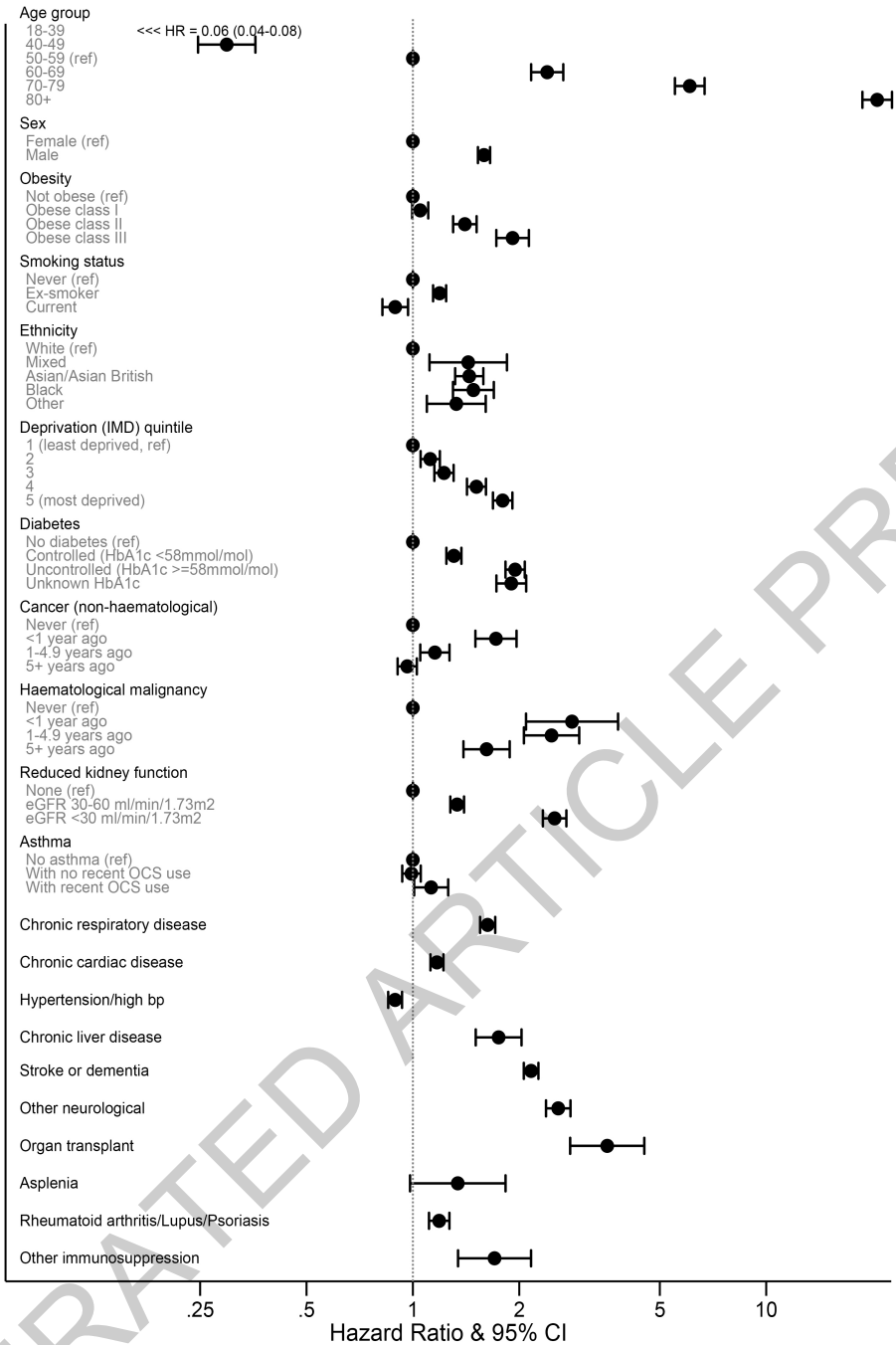


Fig. 3 | Estimated Hazard Ratios (shown on a log scale) for each potential risk factor from a multivariable Cox model. Error bars represent limits of the 95% confidence interval for the hazard ratio. Obese class I: 30-34.9kg/m^2, class II: 35-39.9kg/m^2, class III: >=40kg/m^2. OCS = oral corticosteroid. All HRs are adjusted for all other factors listed other than ethnicity. Ethnicity estimates are from a separate model among those with complete ethnicity data, and are fully adjusted for other covariates. Total n = 17,278,392 for non-ethnicity models, and 12,718,279 for ethnicity model.

ACCELERATED ARTICLE PREVIEW

Table 1 | Cohort description with number of COVID-19 deaths by potential risk factors

Characteristic	Category	N (column %)	Number of COVID-19 deaths (% within stratum)
Total		17,278,392 (100.0)	10,926 (0.06)
Age	18-<40	5,914,384 (34.2)	54 (0.00)
	40-<50	2,849,984 (16.5)	140 (0.00)
	50-<60	3,051,110 (17.7)	522 (0.02)
	60-<70	2,392,392 (13.8)	1,101 (0.05)
	70-<80	1,938,842 (11.2)	2,635 (0.14)
	80+	1,131,680 (6.5)	6,474 (0.57)
Sex	Female	8,647,989 (50.1)	4,764 (0.06)
	Male	8,630,403 (49.9)	6,162 (0.07)
BMI (kg/m2)	<18.5	310,721 (1.8)	522 (0.17)
	18.5-24.9	4,763,150 (27.6)	3,364 (0.07)
	25-29.9	4,682,906 (27.1)	3,068 (0.07)
	30-34.9 (Obese class I)	2,384,406 (13.8)	1,813 (0.08)
	35-39.9 (Obese class II)	922,398 (5.3)	762 (0.08)
	≥40 (Obese class III)	463,042 (2.7)	379 (0.08)
	Missing	3,751,769 (21.7)	1,018 (0.03)
Smoking	Never	7,924,739 (45.9)	3,598 (0.05)
	Former	5,690,966 (32.9)	6,531 (0.11)
	Current	2,941,764 (17.0)	708 (0.02)
	Missing	720,923 (4.2)	89 (0.01)
Ethnicity	White	10,866,411 (62.9)	7,119 (0.07)
	Mixed	169,697 (1.0)	62 (0.04)
	South Asian	1,022,130 (5.9)	608 (0.06)
	Black	339,909 (2.0)	250 (0.07)
	Other	320,132 (1.9)	110 (0.03)
	Missing	4,560,113 (26.4)	2,777 (0.06)
IMD quintile	1 (least deprived)	3,497,154 (20.2)	1,908 (0.05)
	2	3,476,668 (20.1)	2,030 (0.06)
	3	3,483,668 (20.2)	2,114 (0.06)
	4	3,480,459 (20.1)	2,388 (0.07)
	5 (most deprived)	3,340,443 (19.3)	2,486 (0.07)
Blood pressure	Normal	3,804,148 (22.0)	2,487 (0.07)
	Elevated	2,482,710 (14.4)	1,899 (0.08)
	High Stage 1	5,548,198 (32.1)	3,281 (0.06)
	High Stage 2	3,728,241 (21.6)	3,229 (0.09)
	Missing	1,715,095 (9.9)	30 (0.00)
High bp or diagnosed hypertension		5,925,492 (34.3)	8,049 (0.14)
Respiratory disease ex asthma		703,917 (4.1)	2,240 (0.32)
Asthma*	With no recent ocs use	2,454,403 (14.2)	1,211 (0.05)
	With recent ocs use	291,670 (1.7)	335 (0.11)
Chronic heart disease		1,167,455 (6.8)	3,811 (0.33)
Diabetes**	With HbA1c<58 mmol/mol	1,038,082 (6.0)	2,391 (0.23)
	With HbA1c>=58 mmol/mol	486,491 (2.8)	1,254 (0.26)
	With no recent HbA1c measure	193,993 (1.1)	444 (0.23)
Cancer (non-haematological)	Diagnosed < 1 year ago	79,964 (0.5)	220 (0.28)
	Diagnosed 1-4.9 years ago	234,186 (1.4)	449 (0.19)
	Diagnosed ≥5 years ago	542,320 (3.1)	1,125 (0.21)
Haematological malignancy	Diagnosed < 1 year ago	8,704 (0.1)	43 (0.49)
	Diagnosed 1-4.9 years ago	27,742 (0.2)	120 (0.43)
	Diagnosed ≥5 years ago	63,460 (0.4)	173 (0.27)
Reduced kidney function***	Estimated GFR 30-60	1,007,383 (5.8)	3,987 (0.40)

Continued

ACCELERATED ARTICLE PREVIEW

Characteristic	Category	N (column %)	Number of COVID-19 deaths (% within stratum)
	Estimated GFR <30	78,093 (0.5)	864 (1.11)
Kidney dialysis		23,978 (0.1)	192 (0.80)
Liver disease		100,017 (0.6)	181 (0.18)
Stroke/dementia		390,002 (2.3)	2,423 (0.62)
Other neurological disease		170,448 (1.0)	665 (0.39)
Organ transplant		20,001 (0.1)	69 (0.34)
Asplenia		27,917 (0.2)	40 (0.14)
Rheumatoid/Lupus/ Psoriasis		878,475 (5.1)	962 (0.11)
Other immunosuppressive condition		278,948 (1.6)	69 (0.02)

*ocs= oral corticosteroid use, recent is <1 year before baseline, **classification by HbA1c based on measures within 15 months before baseline, ***GFR= glomerular filtration rate (ml/min/1.73 m^2), from most recent serum creatinine measure

ACCELERATED ARTICLE PREVIEW

Table 2 | Hazard Ratios (HRs) and 95% confidence intervals (CI) for COVID-19 death

Characteristic	Category	COVID-19 Death HR (95% CI)	
		Age-sex adj	Fully adj
Age	18-<40	0.05 (0.04-0.07)	0.06 (0.04-0.08)
	40-<50	0.28 (0.23-0.33)	0.30 (0.25-0.36)
	50-<60	1.00 (ref)	1.00 (ref)
	60-<70	2.79 (2.52-3.10)	2.40 (2.16-2.66)
	70-<80	8.62 (7.84-9.46)	6.08 (5.52-6.69)
	80+	38.29 (35.02-41.87)	20.61 (18.72-22.70)
Sex	Female	1.00 (ref)	1.00 (ref)
	Male	1.78 (1.71-1.85)	1.59 (1.53-1.65)
BMI	Not obese	1.00 (ref)	1.00 (ref)
	30-34.9kg/m2 (Obese class I)	1.23 (1.17-1.30)	1.05 (1.00-1.11)
	35-39.9kg/m2 (Obese class II)	1.81 (1.68-1.95)	1.40 (1.30-1.52)
	≥40 kg/m2 (Obese class III)	2.66 (2.39-2.95)	1.92 (1.72-2.13)
Smoking	Never	1.00 (ref)	1.00 (ref)
	Former	1.43 (1.37-1.49)	1.19 (1.14-1.24)
	Current	1.14 (1.05-1.23)	0.89 (0.82-0.97)
Ethnicity*	White	1.00 (ref)	1.00 (ref)
	Mixed	1.62 (1.26-2.08)	1.43 (1.11-1.85)
	South Asian	1.69 (1.54-1.84)	1.44 (1.32-1.58)
	Black	1.88 (1.65-2.14)	1.48 (1.30-1.69)
	Other	1.37 (1.13-1.65)	1.33 (1.10-1.61)
IMD quintile	1 (least deprived)	1.00 (ref)	1.00 (ref)
	2	1.16 (1.08-1.23)	1.12 (1.05-1.19)
	3	1.31 (1.23-1.40)	1.23 (1.15-1.30)
	4	1.69 (1.59-1.79)	1.51 (1.42-1.61)
	5 (most deprived)	2.11 (1.98-2.25)	1.80 (1.69-1.91)
Blood pressure	Normal	1.00 (ref)	1.00 (ref)
	High bp or diagnosed hypertension	1.09 (1.05-1.14)	0.89 (0.85-0.93)
Respiratory disease ex asthma		1.95 (1.86-2.04)	1.63 (1.55-1.71)
Asthma (vs none)**	With no recent OCS use	1.13 (1.07-1.20)	0.99 (0.93-1.05)
	With recent OCS use	1.55 (1.39-1.73)	1.13 (1.01-1.26)
Chronic heart disease		1.57 (1.51-1.64)	1.17 (1.12-1.22)
Diabetes (vs none)***	With HbA1c<58 mmol/mol	1.58 (1.51-1.66)	1.31 (1.24-1.37)
	With HbA1c>=58 mmol/mol	2.61 (2.46-2.77)	1.95 (1.83-2.07)
	With no recent HbA1c measure	2.27 (2.06-2.50)	1.90 (1.72-2.09)
Cancer (non-haematological, vs none)	Diagnosed < 1 year ago	1.81 (1.58-2.07)	1.72 (1.50-1.97)
	Diagnosed 1-4.9 years ago	1.20 (1.10-1.32)	1.15 (1.05-1.27)
	Diagnosed ≥5 years ago	0.99 (0.93-1.06)	0.96 (0.91-1.03)
Haematological malignancy (vs none)	Diagnosed < 1 year ago	3.02 (2.24-4.08)	2.82 (2.09-3.81)
	Diagnosed 1-4.9 years ago	2.56 (2.14-3.06)	2.47 (2.06-2.96)
	Diagnosed ≥5 years ago	1.70 (1.46-1.98)	1.62 (1.39-1.88)
Reduced kidney function (vs none)****	Estimated GFR 30-60	1.56 (1.49-1.63)	1.33 (1.28-1.40)
	Estimated GFR <30	3.48 (3.23-3.75)	2.52 (2.33-2.72)
Liver disease		2.39 (2.06-2.77)	1.75 (1.51-2.03)
Stroke/dementia		2.57 (2.46-2.70)	2.16 (2.06-2.27)
Other neurological disease		3.08 (2.85-3.33)	2.58 (2.38-2.79)
Organ transplant		6.00 (4.73-7.61)	3.55 (2.79-4.52)
Asplenia		1.62 (1.19-2.21)	1.34 (0.98-1.83)
Rheumatoid/Lupus/Psoriasis		1.30 (1.21-1.38)	1.19 (1.11-1.27)
Other immunosuppressive condition		2.06 (1.62-2.61)	1.70 (1.34-2.16)

Models adjusted for age using a 4-knot cubic spline age spline, except for estimation of age group hazard ratios. *Ethnicity hazard ratios estimated from a model restricted to those with recorded ethnicity. **OCS = oral corticosteroids. Recent OCS use defined as in the year before baseline. ***HbA1c classification based on latest measure within 15 months before baseline. ****GFR = glomerular filtration rate in ml/min/1.73m2, based on most recent serum creatinine measure

ACCELERATED ARTICLE PREVIEW

Methods

Study design

We conducted a cohort study using national primary care electronic health record data linked to COVID-19 death data (see Data Source). The cohort study began on 1st February 2020, chosen as a date several weeks prior to the first reported COVID-19 deaths and the day after the second laboratory confirmed case;[27] and ended on 6th May 2020. The cohort explores risk among the general population rather than in a population infected with SARS-COV-2. Therefore, all patients were included irrespective of any SARS-COV-2 test results.

Data Source

We used patient data from general practice (GP) records managed by the GP software provider The Phoenix Partnership (TPP), linked to Office for National Statistics (ONS) death data. ONS data includes information on all deaths, including COVID-19 related death, defined as a COVID-19 ICD-10 code mentioned anywhere on the death certificate and non-COVID-19 death, which was used for censoring.

The data were accessed, linked and analysed using OpenSAFELY, a new data analytics platform created to address urgent questions relating to the epidemiology and treatment of COVID-19 in England. OpenSAFELY provides a secure software interface that allows detailed pseudonymised primary care patient records to be analysed in near real-time where they already reside, hosted within the EHR vendor's highly secure data centre, to minimise the re-identification risks when data are transported off-site; other smaller datasets are linked to these data within the same environment using a matching pseudonym derived from the NHS number. More information can be found on https://opensafely.org/.

The dataset analysed with OpenSAFELY is based on 24 million currently registered patients (approximately 40% of the English population) from GP surgeries using the TPP SystmOne electronic health record system. SystmOne is a secure centralised EHR used in English clinical practice since 1998; it records data entered (in real time) by GPs and practice staff during routine primary care. The system is accredited under the NHS approved systems framework for General Practice.[28,29] Data extracted from TPP SystmOne have previously been used in medical research, as part of the ResearchOne dataset.[30,31] From this EHR a pseudonymised dataset was created for OpenSAFELY consisting of 20 billion rows of structured data including for example pseudonymised patients' diagnoses, medications, physiological parameters, and prior investigations [Extended Data Figure 2, Level 1]. All OpenSAFELY data processing took place on TPP's servers; external data providers securely transferred pseudoymised data (such as COVID-19 related death from ONS) for linkage to OpenSAFELY [Extended Data Figure 2, Level 2]; study definitions developed in Python on GitHub were pulled into the OpenSAFELY infrastructure, and used to create a study dataset of one row per patient [Extended Data Figure 2, Level 3]. Statistical code was developed using synthetic data and used to analyse the study dataset; this included code to check data ranges, to check consistency of data columns, and to produce descriptive statistics for comparison with expected disease prevalences to ensure validity, as well as code to fit our analysis models. Only two authors (KB/AJW) accessed OpenSAFELY to run code; no pseudonymised patient-level data were ever removed from TPP infrastructure; only aggregated, anonymous, manually checked study results were released for publication [Extended Data Figure 2, Level 4], All code for data management and analysis is archived online (see Code Availability, below).

Study Population and Observation Period

Our study population consisted of all adults (males and females 18 years and above) currently registered as active patients in a TPP general practice in England on 1st February 2020. To be included in the study, participants were required to have at least 1 year of prior follow-up in the GP practice to ensure that baseline patient characteristics could be adequately captured, and to have recorded sex, age, and deprivation (see covariates, below).[32] Patients were observed from the 1st of February 2020 and were followed until the first of either their death date (whether COVID-19 related or due to other causes) or the study end date, 6th May 2020. For this analysis, ONS death data were available to 11th May 2020, but we used an earlier censor date to allow for delays in reporting in the last few days of available data.

Outcomes

The outcome was death among people with COVID-19, ascertained from ONS death certificate data, where the COVID related ICD-10 codes U071 or U072 were present in the record.

Covariates

Potential risk factors included: health conditions listed in UK guidance on "higher risk" groups;[33] other common conditions which may cause immunodeficiency inherently or through medication (cancer and common autoimmune conditions); and emerging risk factors for severe outcomes among COVID-19 cases (such as raised blood pressure).

Age, sex, body mass index (BMI; kg/m^2), and smoking status were considered as potential risk factors. Where categorised, age groups were: 18-<40, 40-<50, 50-<60, 60-<70, 70-<80, 80+ years. BMI was ascertained from weight measurements within the last 10 years, restricted to those taken when the patient was over 16 years old. Obesity was grouped using categories derived from the World Health Organisation classification of BMI: no evidence of obesity <30 kg/m^2; obese I 30-34.9; obese II 35-39.9; obese III 40+. Smoking status was grouped into current, former and never smokers.

The following comorbidities were also considered potential risk factors: asthma, other chronic respiratory disease, chronic heart disease, diabetes mellitus, chronic liver disease, chronic neurological diseases, common autoimmune diseases (Rheumatoid Arthritis (RA), Systemic Lupus Erythematosus (SLE) or psoriasis), solid organ transplant, asplenia, other immunosuppressive conditions, cancer, evidence of reduced kidney function, and raised blood pressure or a diagnosis of hypertension.

Disease groupings followed national guidance on risk of influenza infection,[34] therefore "chronic respiratory disease (other than asthma)" included COPD, fibrosing lung disease, bronchiectasis or cystic fibrosis; chronic heart disease included chronic heart failure, ischaemic heart disease, and severe valve or congenital heart disease likely to require lifelong follow-up. Chronic neurological conditions were separated into diseases with a likely cardiovascular aetiology (stroke, TIA, dementia) and conditions in which respiratory function may be compromised such as motor neurone disease, myasthenia gravis, multiple sclerosis, Parkinson's disease, cerebral palsy, quadriplegia or hemiplegia, and progressive cerebellar disease. Asplenia included splenectomy or a spleen dysfunction, including sickle cell disease. Other immunosuppressive conditions included HIV or a condition inducing permanent immunodeficiency ever diagnosed, or aplastic anaemia or temporary immunodeficiency recorded within the last year. Haematological malignancies were considered separately from other cancers to reflect the immunosuppression associated with haematological malignancies and their treatment. Kidney function was ascertained from the most recent serum creatinine measurement, where available, converted into estimated glomerular filtration rate (eGFR) using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation,[35] with reduced kidney function grouped into defined as eGFR 30-<60 or <30 mL/min/1.73m^2. History of kidney dialysis or end-stage renal failure was separately explored in a secondary analysis. Raised blood pressure (BP) was defined as either a prior coded diagnosis of hypertension or the most recent recording indicating systolic BP ≥140 mmHg or diastolic BP ≥90 mmHg.

ACCELERATED ARTICLE PREVIEW

Asthma was grouped by use of oral corticosteroids as an indication of severity. Diabetes was grouped according to the most recent Hba1c measurement within the last 15 months (Hba1c <58 mmols/mol, ≥58 mmols/mol, no recent measure available). Cancer was grouped by time since the first diagnosis (within the last year, 2-<5 years, ≥5 years).

Other covariates considered as potential upstream risk factors were deprivation and ethnicity. Deprivation was measured by the Index of Multiple Deprivation (IMD, in quintiles, with higher values indicating greater deprivation), derived from the patient's postcode at lower super output area level for a high degree of precision. Ethnicity was grouped into White, Black, South Asian, Mixed, or Other. In sensitivity analyses, a more detailed grouping of ethnicity was explored. The Sustainability and Transformation Partnership (STP, an NHS administrative region) of the patient's general practice was included as an additional adjustment for geographical variation in infection rates across the country.

Information on all covariates were obtained from primary care records by searching TPP SystmOne records for specific coded data. TPP SystmOne allows users to work with the SNOMED-CT clinical terminology, using a GP subset of SNOMED-CT codes. This subset maps on to the native Read version 3 (CTV3) clinical coding system that SystmOne is built on. Medicines are entered or prescribed in a format compliant with the NHS Dictionary of Medicines and Devices (dm+d),[36] a local UK extension library of SNOMED. Code lists for particular underlying conditions and medicines were compiled from a variety of sources. These include BNF codes from OpenPrescribing.net, published codelists for asthma,[37–39] immunosuppression,[40–42] psoriasis,[43] SLE,[44] RA[45,46] and cancer,[47,48] and Read Code 2 lists designed specifically to describe groups at increased risk of influenza infection.[18] Read Code 2 lists were added to with SNOMED codes and cross-checked against NHS QOF registers, then translated into CTV3 with manual curation. Decisions on every code list were documented and final lists reviewed by at least two authors. Detailed information on compilation and sources for every individual codelist is available at https://codelists.opensafely.org/ and the lists are available for inspection and re-use by the broader research community.[49]

Statistical Analysis

Patient numbers are depicted in a flowchart. The Kaplan-Meier failure function was estimated by age group and sex. For each potential risk factor, a Cox proportional hazards model was fitted, with days in study as the timescale, stratified by geographic area (STP), and adjusted for sex and age modelled using restricted cubic splines. Violations of the proportional hazards assumption were explored by testing for a zero slope in the scaled Schoenfeld residuals. All potential risk factors, including age (again modelled as a spline), sex, BMI, smoking, index of multiple deprivation quintile, and comorbidities listed above were then included in a single multivariable Cox proportional hazards model, stratified by STP. Hazard ratios from the age/sex adjusted and fully adjusted models are reported with 95% confidence intervals. Models were also refitted with age group fitted as a categorical variable in order to obtain hazard ratios by age group.

In the primary analysis, those with missing BMI were assumed non-obese and those with missing smoking information were assumed to be non-smokers on the assumption that both obesity and smoking would be likely to be recorded if present. A sensitivity analysis was run among those with complete BMI and smoking data only. Ethnicity was omitted from the main multivariable model due to 26% of individuals having missing data; hazard ratios for ethnicity were therefore obtained from a separate model among individuals with complete ethnicity only. Hazard ratios for other risk factors, adjusted for ethnicity, were also obtained from this model and are presented in the sensitivity analyses to allow assessment of whether estimates may have been distorted by ethnicity in the primary model. We conducted an additional sensitivity analysis using a population-calibrated imputation approach to handle missing ethnicity,[50,51] with marginal proportions of each ethnicity group within each of nine broad geographical regions of England (East, East Midlands, London, North East, North West, South East, South West, West Midlands, Yorkshire and The Humber) taken from Annual Population Survey (APS) data (pooled 2014-2016).[52] Five imputed datasets were created with estimated hazard ratios combined using Rubin's rules.

The C-statistic was calculated as a measure of model discrimination. Due to computational time, this was estimated by randomly sampling 5000 patients with and without the outcome and calculating the C-statistic using the random sample, repeating this 10 times and taking the average C-statistic.

All p-values presented are two-sided.

Information governance and ethics

NHS England is the data controller; TPP is the data processor; and the key researchers on OpenSAFELY are acting on behalf of NHS England. This implementation of OpenSAFELY is hosted within the TPP environment which is accredited to the ISO 27001 information security standard and is NHS IG Toolkit compliant;[53,54] patient data has been pseudonymised for analysis and linkage using industry standard cryptographic hashing techniques; all pseudonymised datasets transmitted for linkage onto OpenSAFELY are encrypted; access to the platform is via a virtual private network (VPN) connection, restricted to a small group of researchers, their specific machine and IP address; the researchers hold contracts with NHS England and only access the platform to initiate database queries and statistical models; all database activity is logged; only aggregate statistical outputs leave the platform environment following best practice for anonymisation of results such as statistical disclosure control for low cell counts.[55] The OpenSAFELY research platform adheres to the data protection principles of the UK Data Protection Act 2018 and the EU General Data Protection Regulation (GDPR) 2016. In March 2020, the Secretary of State for Health and Social Care used powers under the UK Health Service (Control of Patient Information) Regulations 2002 (COPI) to require organisations to process confidential patient information for the purposes of protecting public health, providing healthcare services to the public and monitoring and managing the COVID-19 outbreak and incidents of exposure.[56] Taken together, these provide the legal bases to link patient datasets on the OpenSAFELY platform. GP practices, from which the primary care data are obtained, are required to share relevant health information to support the public health response to the pandemic, and have been informed of the OpenSAFELY analytics platform. This study was approved by the Health Research Authority (REC reference 20/LO/0651) and by the LSHTM Ethics Board (reference 21863).

Patient and Public Involvement

Patients were not formally involved in developing this specific study design. We have developed a publicly available website https://opensafely.org/ allowing any patient or member of the public to contact us regarding this study or the broader OpenSAFELY project. This feedback will be used to refine and prioritise our OpenSAFELY activities.

Reporting summary

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

Data availability

All data were linked, stored and analysed securely within the OpenSAFELY platform https://opensafely.org/. All code is shared openly for review and re-use under MIT open license. Detailed pseudonymised patient data is potentially re-identifiable and therefore not shared. We rapidly delivered the OpenSAFELY data analysis platform without prior funding to deliver timely analyses on urgent research questions in the context of the global Covid-19 health emergency: now that the platform is established we are developing a formal process for external

ACCELERATED ARTICLE PREVIEW

users to request access in collaboration with NHS England; details of this process will be published shortly on OpenSAFELY.org.

Code availability

Data management was performed using Python 3.8 and SQL, with analysis carried out using Stata 16.1/ Python. All code for data management and analysis is archived online at https://github.com/opensafely/risk-factors-research. All clinical and medicines codelists are openly available for inspection and reuse at https://codelists.opensafely.org/.

27. Coronavirus (COVID-19) cases in the UK. https://web.archive.org/web/20200502045059/https://coronavirus.data.gov.uk/ (2020).
28. GP Systems of Choice - NHS Digital. *NHS Digital* https://digital.nhs.uk/services/gp-systems-of-choice.
29. Future GP IT systems and services - NHS Digital. *NHS Digital* https://digital.nhs.uk/services/future-gp-it-systems-and-services.
30. Clegg, A. *et al*. Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age Ageing* **45**, 353–360 (2016).
31. Harcourt, S. *et al*. Estimating primary care attendance rates for fever in infants after meningococcal B vaccination in England using national syndromic surveillance data. *Vaccine* **36**, 565–571 (2018).
32. Lewis, J. D., Bilker, W. B., Weinstein, R. B. & Strom, B. L. The relationship between time since registration and measured incidence rates in the General Practice Research Database. *Pharmacoepidemiol. Drug Saf*. **14**, 443–451 (2005).
33. Public Health England. Guidance on social distancing for everyone in the UK. *GOV.UK* https://web.archive.org/web/20200429043059/https://www.gov.uk/government/publications/covid-19-guidance-on-social-distancing-and-for-vulnerable-people/guidance-on-social-distancing-for-everyone-in-the-uk-and-protecting-older-people-and-vulnerable-adults (2020).
34. Public Health England. UK immunisation schedule: the green book, chapter 11. *GOV.UK* https://www.gov.uk/government/publications/immunisation-schedule-the-green-book-chapter-11 (2013).
35. Levey, A. S. *et al*. A New Equation to Estimate Glomerular Filtration Rate. *Ann. Intern. Med*. **150**, 604 (2009).
36. MacKenna, B. What is the dm+d? The NHS Dictionary of Medicines and Devices. *EBM DataLab* https://web.archive.org/web/20200502143707/https://ebmdatalab.net/what-is-the-dmd-the-nhs-dictionary-of-medicines-and-devices/ (2019).
37. Nissen, F. *et al*. Validation of asthma recording in the Clinical Practice Research Datalink (CPRD). *BMJ Open* **7**, e017474 (2017).
38. Morton, C. & Douglas, I. OpenSAFELY Codelists: Asthma Diagnosis. https://codelists.opensafely.org/codelist/opensafely/asthma-diagnosis/.
39. MacKenna, B. & Douglas, I. OpenSAFELY Codelists: Asthma Oral Prednisolone Medication. *OpenSAFELY Codelists* https://codelists.opensafely.org/codelist/opensafely/asthma-oral-prednisolone-medication/.
40. Grint, D. J. *et al*. Safety of inadvertent administration of live zoster vaccine to immunosuppressed individuals in a UK-based observational cohort analysis. *BMJ Open* **10**, e034886 (2020).
41. McDonald, H. & Smeeth, L. OpenSAFELY Codelists: Permanent Immunosuppresion. *OpenSAFELY Codelists* https://codelists.opensafely.org/codelist/opensafely/permanent-immunosuppresion/.
42. Smeeth, L. & McDonald, H. OpenSAFELY Codelists: Temporary Immunosuppresion. *OpenSAFELY Codelists* https://codelists.opensafely.org/codelist/opensafely/temporary-immunosuppresion/.
43. Wong, A., Schmidt, S. A. J. & Langan, S. Clinical Code List-Psoriasis-Read Codes. (2019).
44. Forbes, H. *et al*. Clinical code list - SLE codes. (2014) https://doi.org/10.17037/DATA.162.
45. Pujades-Rodriguez, M. *et al*. Rheumatoid Arthritis and Incidence of Twelve Initial Presentations of Cardiovascular Disease: A Population Record-Linkage Cohort Study in England. *PLoS One* **11**, e0151245 (2016).
46. OpenSAFELY Codelists: RA / SLE / Psoriasis. https://codelists.opensafely.org/codelist/opensafely/ra-sle-psoriasis/.
47. Strongman, H. *et al*. Medium and long-term risks of specific cardiovascular diseases in survivors of 20 adult cancers: a population-based cohort study using multiple linked UK electronic health records databases. *Lancet* **394**, 1041–1054 (2019).
48. OpenSAFELY Codelists: Cancer excluding lung and haematological. https://codelists.opensafely.org/codelist/opensafely/cancer-excluding-lung-and-haematological/.
49. OpenSAFELY Codelists. https://codelists.opensafely.org/.
50. Carpenter, J. & Kenward, M. *Multiple imputation and its application*. (Chichester: John Wiley & Sons, 2012).
51. Pham, T. M., Carpenter, J. R., Morris, T. P., Wood, A. M. & Petersen, I. Population-calibrated multiple imputation for a binary/categorical covariate in categorical regression models. *Stat. Med*. **38**, 792–808 (2019).
52. ONS. Population characteristics research tables. *ONS* https://web.archive.org/web/20200513113451/https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationcharacteristicsresearchtables (2019).
53. BETA – Data Security Standards - NHS Digital. *NHS Digital* https://digital.nhs.uk/about-nhs-digital/our-work/nhs-digital-data-and-technology-standards/framework/beta–data-security-standards.
54. Data Security and Protection Toolkit - NHS Digital. *NHS Digital* https://digital.nhs.uk/data-and-information/looking-after-information/data-security-and-information-governance/data-security-and-protection-toolkit.
55. ISB1523: Anonymisation Standard for Publishing Health and Social Care Data - NHS Digital. *NHS Digital* https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/isb1523-anonymisation-standard-for-publishing-health-and-social-care-data.
56. Secretary of State for Health and Social Care - UK Government. Coronavirus (COVID-19): notification to organisations to share information. https://web.archive.org/web/20200421171727/https://www.gov.uk/government/publications/coronavirus-covid-19-notification-of-data-controllers-to-share-information (2020).

Acknowledgements All authors are from The OpenSAFELY Collaborative. We are very grateful for all the support received from the TPP Technical Operations team throughout this work; for generous assistance from the information governance and database teams at NHS England / NHSX; and for additional discussions on disease characterisation, codelists, and methodology with Henry Drysdale, Brian Nicholson, Nick DeVito, Will Hulme, Ieva Lipska, Jess Morley, Jenni Quint and Tra Pham. No dedicated funding has yet been obtained for this work. TPP provided technical expertise and infrastructure within their data centre *pro bono* in the context of a national emergency. BG's work on better use of data in healthcare more broadly is currently funded in part by: NIHR Oxford Biomedical Research Centre, NIHR Applied Research Collaboration Oxford and Thames Valley, the Mohn-Westlake Foundation, NHS England, and the Health Foundation; all DataLab staff are supported by BG's grants on this work. LS reports grants from Wellcome, MRC, NIHR, UKRI, British Council, GSK, British Heart Foundation, and Diabetes UK outside this work. KB holds a Sir Henry Dale fellowship jointly funded by Wellcome and the Royal Society. HIM is funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Immunisation, a partnership between Public Health England and LSHTM. AYSW holds a fellowship from BHF. RM holds a Sir Henry Wellcome fellowship. EW holds grants from MRC. RG holds grants from NIHR and MRC. ID golds grants from NIHR and GSK. RM holds a Sir Henry Wellcome Fellowship funded by the Wellcome Trust. HF holds a UKRI fellowship. The views expressed are those of the authors and not necessarily those of the NIHR, NHS England, Public Health England or the Department of Health and Social Care. Funders had no role in the study design, collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication. This study was approved by the Health Research Authority (REC reference 20/LO/0651) and by the LSHTM Ethics Board (ref 21863). No further ethical or research governance approval was required by the University of Oxford but copies of the approval documents were reviewed and held on record. Guarantor: BG/LS.

Author contributions BG conceived the platform and the approach; BG and LS led the project overall and are guarantors; SB led the software; EW KB led the statistical analysis; CM AJW led on codelists and implementation; AM led on IG; Contributions are as follows: Data curation CB JP JC SH SB DE PI CM; Analysis EW KB AJW CM; Funding acquisition BG LS; information governance AM BG CB JP; Methodology EW KB AJW BG LS CB JP JC SH SB DE PI CM RP; Disease category conceptualisation and codelists CM AJW PI SB DE CB JC JP SH HD HC KB SB AM BM LT ID HM RM HF JQ; Ethics approval HC EW LS BG; Project administration CM HC CB SB AM LS BG; Resources BG LS FH; Software SB DE PI AJW CM CB FH JC SH; Supervision BG LS SB; Writing (original draft) HC EW KB BM CM AM BG LS; Writing (review & editing) CB CM HC EW KB SB AM BM LT ID HM RM AJW SE. All authors were involved in design and conceptual development and reviewed and approved the final manuscript. Authors EW AW KB SB CB contributed equally.

Competing interests All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare the following: CB JP FH JC SH are employees of TPP. AM was interim Chief Medical Officer NHS Digital April-Sept 2019 (left NHS Digital end Jan 2020) and Digital Clinical Champion NHS England 2014-2015. All other authors have no competing interests.

Additional information

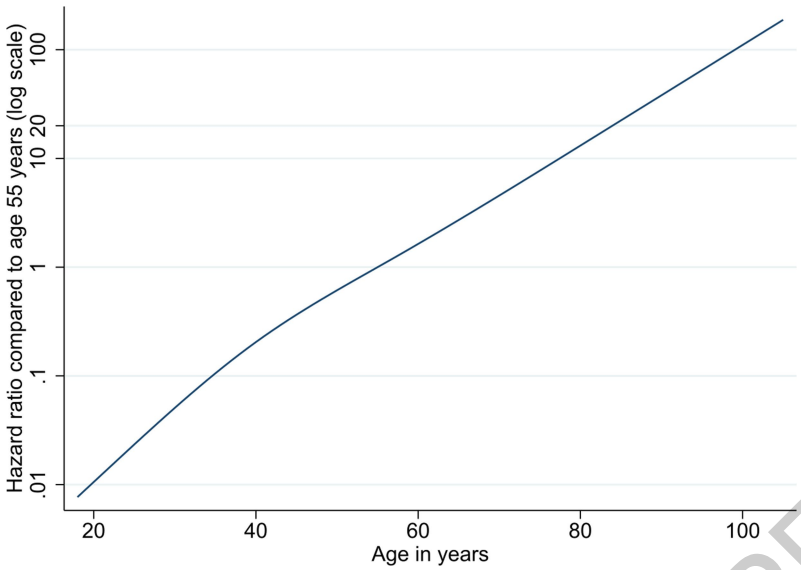
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2521-4.

Correspondence and requests for materials should be addressed to B.G.

Peer review information *Nature* thanks David Christiani, Jeffrey Morris and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

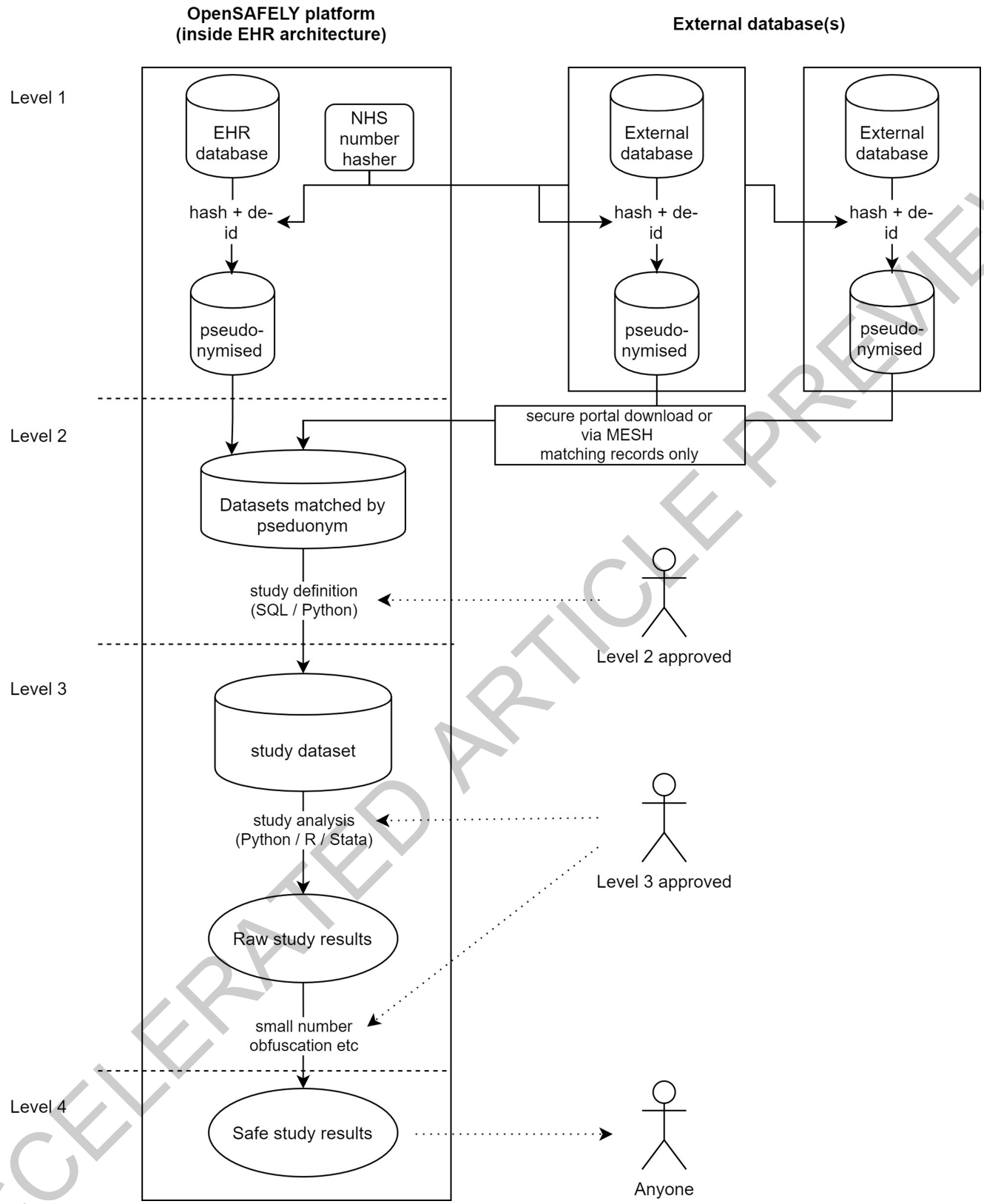
Reprints and permissions information is available at http://www.nature.com/reprints.

ACCELERATED ARTICLE PREVIEW



Extended Data Fig. 1 | Estimated log hazard ratio by age in years. From the primary fully adjusted model containing a 4-knot cubic spline for age, and adjusted for all covariates listed in Table 2 except for ethnicity.

ACCELERATED ARTICLE PREVIEW



Extended Data Fig. 2 | Illustration of data flows in the OpenSAFELY platform.

ACCELERATED ARTICLE PREVIEW

Extended Data Table 1 | Adjusted hazard ratios for detailed ethnicity categories

Ethnicity	Fully adjusted hazard ratio*	95% CI
British or mixed British	1.00	(ref)
Irish	1.16	(0.96-1.41)
Other White	0.87	(0.79-0.97)
Mixed ethnicity	1.42	(1.11-1.83)
Indian or British Indian	1.40	(1.23-1.59)
Pakistani or British Pakistani	1.24	(1.05-1.46)
Bangladeshi or British Bangladeshi	1.84	(1.35-2.49)
Other Asian	1.73	(1.44-2.08)
Caribbean	1.28	(1.07-1.53)
African	1.78	(1.42-2.23)
Other Black	1.73	(1.24-2.41)
Chinese	1.22	(0.81-1.84)
Other	1.35	(1.09-1.67)

Estimated from a model restricted to those with recorded ethnicity, adjusted for age using a 4-knot cubic spline age spline, sex, BMI, smoking, IMD quintile, hypertension/high blood pressure, asthma, chronic heart disease, diabetes, non-haematological cancer, haematological malignancy, reduced kidney function, liver disease, stroke/dementia, other neurological disease, organ transplant, asplenia, rheumatoid/lupus/psoriasis, other immunosuppressive condition; all categorisations are as in the primary analysis.

ACCELERATED ARTICLE PREVIEW

Extended Data Table 2 | Hazard Ratios (HRs) and 95% confidence intervals (CI) in sensitivity analyses

		Fully adjusted HR and 95% CI				
Characteristic	**Category**	**Primary analysis**	**Early censoring at 6/4/2020**	**Restricted to those with complete BMI /smoking**	**Adjusted for ethnicity in those where recorded**	**Adjusted for ethnicity using multiple imputation**
N outcome events in analysis		10926	2816	9880	8149	
Age	18-<40	0.06 (0.04-0.08)	0.07 (0.04-0.12)	0.07 (0.05-0.10)	0.06 (0.05-0.09)	0.06 (0.04-0.07)
	40-<50	0.30 (0.25-0.36)	0.32 (0.23-0.45)	0.30 (0.24-0.37)	0.29 (0.24-0.36)	0.29 (0.24-0.35)
	50-<60	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)
	60-<70	2.40 (2.16-2.66)	2.55 (2.11-3.08)	2.38 (2.13-2.67)	2.37 (2.11-2.67)	2.43 (2.19-2.70)
	70-<80	6.08 (5.52-6.69)	5.84 (4.89-6.99)	5.96 (5.38-6.61)	6.06 (5.43-6.76)	6.24 (5.66-6.87)
	80+	20.61 (18.72-22.70)	14.68 (12.24-17.59)	19.97 (18.01-22.15)	20.20 (18.09-22.55)	21.19 (19.23-23.34)
Sex	Female	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)
	Male	1.59 (1.53-1.65)	1.90 (1.75-2.05)	1.65 (1.58-1.72)	1.54 (1.47-1.61)	1.57 (1.52-1.64)
BMI	Not obese	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)	
	30-34.9kg/m2 (Obese class I)	1.05 (1.00-1.11)	1.30 (1.18-1.43)	1.07 (1.02-1.13)	1.05 (0.99-1.11)	1.06 (1.00-1.11)
	35-39.9kg/m2 (Obese class II)	1.40 (1.30-1.52)	1.57 (1.36-1.81)	1.45 (1.34-1.57)	1.41 (1.30-1.54)	1.42 (1.32-1.54)
	≥40 kg/m2 (Obese class III)	1.92 (1.72-2.13)	2.70 (2.26-3.21)	1.99 (1.79-2.21)	1.92 (1.70-2.17)	1.96 (1.76-2.18)
Smoking	Never	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)
	Former	1.19 (1.14-1.24)	1.27 (1.17-1.39)	1.18 (1.13-1.24)	1.22 (1.16-1.29)	1.23 (1.18-1.29)
	Current	0.89 (0.82-0.97)	0.93 (0.79-1.09)	0.91 (0.83-0.99)	0.93 (0.84-1.02)	0.93 (0.85-1.01)
Ethnicity*	White	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)
	Mixed	1.43 (1.11-1.85)	1.01 (0.60-1.72)	1.38 (1.05-1.80)	1.43 (1.11-1.85)	1.44 (1.06-1.95)
	South Asian	1.44 (1.32-1.58)	1.62 (1.38-1.91)	1.51 (1.38-1.66)	1.44 (1.32-1.58)	1.48 (1.33-1.65)
	Black	1.48 (1.30-1.69)	1.76 (1.41-2.20)	1.48 (1.28-1.70)	1.48 (1.30-1.69)	1.53 (1.32-1.77)
	Other	1.33 (1.10-1.61)	1.84 (1.37-2.47)	1.40 (1.15-1.70)	1.33 (1.10-1.61)	1.34 (1.12 1.61)
IMD quintile	1 (least deprived)	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)
	2	1.12 (1.05-1.19)	0.96 (0.85-1.08)	1.12 (1.05-1.19)	1.16 (1.08-1.25)	1.12 (1.05-1.19)
	3	1.23 (1.15-1.30)	1.00 (0.88-1.13)	1.23 (1.15-1.31)	1.26 (1.17-1.36)	1.21 (1.14-1.29)
	4	1.51 (1.42-1.61)	1.26 (1.11-1.41)	1.51 (1.42-1.61)	1.54 (1.43-1.66)	1.48 (1.39-1.57)
	5 (most deprived)	1.80 (1.69-1.91)	1.41 (1.25-1.60)	1.80 (1.69-1.93)	1.77 (1.64-1.91)	1.72 (1.61-1.84)
Blood pressure	Normal	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)
	High bp or diagnosed hyper-tension	0.89 (0.85-0.93)	0.95 (0.87-1.04)	0.88 (0.84-0.92)	0.91 (0.86-0.96)	0.89 (0.85-0.93)
Respiratory disease ex asthma		1.63 (1.55-1.71)	1.86 (1.69-2.04)	1.59 (1.51-1.67)	1.65 (1.56-1.75)	1.64 (1.56-1.72)
Asthma (vs none)**	With no recent OCS use	0.99 (0.93-1.05)	1.08 (0.96-1.21)	0.97 (0.91-1.04)	0.94 (0.87-1.00)	0.98 (0.93-1.05)
	With recent OCS use	1.13 (1.01-1.26)	1.38 (1.13-1.67)	1.09 (0.97-1.22)	1.08 (0.95-1.22)	1.11 (0.99-1.24)
Chronic heart disease		1.17 (1.12-1.22)	1.37 (1.26-1.48)	1.16 (1.11-1.22)	1.16 (1.11-1.22)	1.17 (1.12-1.22)
Diabetes (vs none)***	With HbA1c<58 mmol/mol	1.31 (1.24-1.37)	1.38 (1.26-1.52)	1.29 (1.23-1.36)	1.28 (1.21-1.35)	1.27 (1.21-1.33)
	WithHbA1c>=58 mmol/mol	1.95 (1.83-2.07)	2.33 (2.08-2.61)	1.90 (1.78-2.02)	1.85 (1.72-1.99)	1.87 (1.76-1.99)
	With no recent HbA1c measure	1.90 (1.72-2.09)	1.71 (1.40-2.08)	1.92 (1.74-2.12)	1.86 (1.67-2.08)	1.84 (1.67-2.02)
Cancer (non-haematological, vs none)	Diagnosed < 1 year ago	1.72 (1.50-1.97)	1.66 (1.27-2.16)	1.68 (1.46-1.94)	1.67 (1.43-1.96)	1.74 (1.52-1.99)
	Diagnosed 1-4.9 years ago	1.15 (1.05-1.27)	1.34 (1.13-1.60)	1.16 (1.05-1.28)	1.21 (1.09-1.35)	1.17 (1.06-1.28)
	Diagnosed ≥5 years ago	0.96 (0.91-1.03)	0.92 (0.81-1.04)	0.97 (0.91-1.03)	0.99 (0.92-1.06)	0.97 (0.92-1.04)
Haematological malignancy (vs none)	Diagnosed < 1 year ago	2.82 (2.09-3.81)	2.22 (1.15-4.27)	2.87 (2.11-3.91)	2.35 (1.61-3.43)	2.81 (2.08-3.79)
	Diagnosed 1-4.9 years ago	2.47 (2.06-2.96)	3.50 (2.61-4.69)	2.40 (1.99-2.91)	2.53 (2.06-3.11)	2.48 (2.07-2.97)
	Diagnosed ≥5 years ago	1.62 (1.39-1.88)	1.45 (1.07-1.98)	1.62 (1.38-1.89)	1.56 (1.31-1.86)	1.63 (1.40-1.89)
Reduced kidney function****	Estimated GFR 30-60	1.33 (1.28-1.40)	1.49 (1.36-1.63)	1.33 (1.27-1.39)	1.37 (1.30-1.44)	1.33 (1.27-1.39)
	Estimated GFR <30	2.52 (2.33-2.72)	2.98 (2.57-3.46)	2.47 (2.28-2.68)	2.50 (2.29-2.74)	2.50 (2.31-2.70)
Liver disease		1.75 (1.51-2.03)	1.92 (1.48-2.49)	1.69 (1.44-1.97)	1.75 (1.48-2.07)	1.75 (1.51-2.03)
Stroke/dementia		2.16 (2.06-2.27)	1.74 (1.58-1.93)	2.12 (2.01-2.22)	2.16 (2.05-2.28)	2.16 (2.06-2.27)
Other neurological disease		2.58 (2.38-2.79)	2.26 (1.91-2.68)	2.50 (2.30-2.73)	2.53 (2.31-2.77)	2.58 (2.38-2.80)
Organ transplant		3.55 (2.79-4.52)	2.57 (1.60-4.13)	3.72 (2.91-4.75)	3.48 (2.64-4.58)	3.48 (2.74-4.44)
Asplenia		1.34 (0.98-1.83)	1.87 (1.13-3.11)	1.29 (0.93-1.80)	1.35 (0.95-1.92)	1.33 (0.98-1.82)
Rheumatoid/Lupus/ Psoriasis		1.19 (1.11-1.27)	1.29 (1.14-1.46)	1.17 (1.09-1.26)	1.15 (1.07-1.24)	1.20 (1.12-1.28)
Other immunosuppressive condition		1.70 (1.34-2.16)	1.98 (1.32-2.96)	1.62 (1.26-2.09)	1.66 (1.27-2.16)	1.67 (1.31-2.11)

Models adjusted for age using a 4-knot cubic spline age spline, except for estimation of age group hazard ratios. *Ethnicity hazard ratios in primary analysis estimated from a model restricted to those with recorded ethnicity. **OCS = oral corticosteroids. Recent OCS use defined as in the year before baseline. ***HbA1c classification based on latest. ****GFR = glomerular filtration rate in ml/min/1.73m2, based on most recent serum creatinine measure.

ACCELERATED ARTICLE PREVIEW

nature research

Corresponding author(s): Ben Godlacre

Last updated by author(s): Jun 16, 2020

Reporting Summary

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

Statistics

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

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data were collected using TPP SystmOne software (14th May maintenance release), for the purpose of direct clinical care. Data management was performed using Python 3.8 and SQL. All code for data management and analysis is archived online at https://github.com/opensafely/risk-factors-research.
Data analysis	Analysis was carried out using Stata 16.1 / Python 3.8. All code for data management and analysis is archived at https://github.com/opensafely/risk-factors-research.

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

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

All data were linked, stored and analysed securely within the OpenSAFELY platform https://opensafely.org/. All code is shared openly for review and re-use under MIT open license. Detailed pseudonymised patient data is potentially re-identifiable and therefore not shared. All clinical and medicines codelists are openly available for inspection and reuse at https://codelists.opensafely.org/.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

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

Study description	We conducted a quantitative cohort study using national primary care electronic health record data linked to COVID-19 death data.
Research sample	We used patient data from general practice (GP) records managed by the GP software provider The Phoenix Partnership (TPP), linked to Office for National Statistics (ONS) death data. The sample of patients represents approximately 40% of the population of England, spread geographically across the whole country.
Sampling strategy	Our study population consisted of all adults (males and females 18 years and above) currently registered as active patients in a TPP general practice in England on 1st February 2020. To be included in the study, participants were required to have at least 1 year of prior follow-up in the GP practice to ensure that baseline patient characteristics could be adequately captured, and to have recorded sex, age, and deprivation (see covariates, below).
Data collection	Data were collected by clinicians (e.g. doctors, nurses) and administrative staff, for the purpose of direct clinical care. This was carried out on computers using TPP SystmOne software. The researchers were not present for data collection into the TPP database. Data were then queried from the TPP database by the researchers, to create the study dataset. This was carried out using Python 3.8 and SQL software (available here https://github.com/opensafely/risk-factors-research). This study did not have an experimental condition or hypothesis.
Timing	Patients were observed from the 1st of February 2020 and were followed until the first of either their death date (whether COVID-19 related or due to other causes) or the study end date, 6th May 2020.
Data exclusions	To be included in the study, participants were required to have at least 1 year of prior follow-up in the GP practice to ensure that baseline patient characteristics could be adequately captured, and to have recorded sex, age, and deprivation. The total number of excluded patients was 6,322,225.
Non-participation	No participants dropped out.
Randomization	Participants were not allocated into experimental groups.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	This study uses data gathered during routine medical practice. We selected all patients except those <18 years old, anyone without a recorded sex, age, or deprivation score, and anyone without a year of prior follow-up (to ensure that baseline patient characteristics could be adequately captured). These inclusive criteria mean that bias is minimised.

Ethics oversight	This study was approved by the Health Research Authority (REC reference 20/LO/0651) and by the LSHTM Ethics Board (reference 21863). In March 2020, the Secretary of State for Health and Social Care used powers under the UK Health Service (Control of Patient Information) Regulations 2002 (COPI) to require organisations to process confidential patient information for the purposes of protecting public health, providing healthcare services to the public and monitoring and managing the COVID-19 outbreak and incidents of exposure. Taken together, these provide the legal bases to link patient datasets on the OpenSAFELY platform and set aside the requirement for patient consent for COVID-19 related public health research. GP practices, from which the primary care data is obtained, are required to share relevant health information to support the public health response to the pandemic, and have been informed of the OpenSAFELY analytics platform.

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