# Using national electronic health records for pandemic preparedness: validation of a parsimonious model for predicting excess deaths among those with COVID-19, a data-driven retrospective cohort study.

Short title: Parsimonious data-driven modelling for pandemic preparedness

Mehrdad A Mizani[1,9], Ashkan Dashtban[1], Laura Pasea[1], Alvina G Lai[1], Johan Thygesen[1], Chris Tomlinson[1], Alex Handy[1], Jil B Mamza[2], Tamsin Morris[2], Sara Khalid[3], Francesco Zaccardi[4], Mary Joan Macleod[5], Fatemeh Torabi[6], Dexter Canoy[7], Ashley Akbari[6], Colin Berry[8], Thomas Bolton[9], John Nolan[9], Kamlesh Khunti[4], Spiros Denaxas[1], Harry Hemingway[1], Cathie Sudlow[9], Amitava Banerjee[1], on behalf of the CVD-COVID-UK Consortium.

[1]Institute of Health Informatics, University College London, London, UK.
[2]Medical and Scientific Affairs, BioPharmaceuticals Medical, AstraZeneca, Cambridge, United Kingdom.
[3]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK.
[4]Leicester Diabetes Centre, University of Leicester, Leicester, UK.
[5]School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, UK.
[6]Faculty of Medicine, Health and Life Science, Swansea University
[7]Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK
[8]Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK
[9]BHF Data Science Centre, Health Data Research UK, London, UK.

*Corresponding Author*
Amitava Banerjee, 222 Euston Rd, London NW1 2DA, ami.banerjee@ucl.ac.uk

*Declaration of competing interests*
JBM and TM are employees of AstraZeneca. KK is chair of the ethnicity subgroup of the Independent Scientific Advisory Group for Emergencies (SAGE) and director of the University of Leicester Centre for Black Minority Ethnic Health. KK and AB are trustees of the South Asian Health Foundation (SAHF). CS is Director of the BHF Data Science Centre. All other authors report no competing interests.

*Funding*
The British Heart Foundation Data Science Centre (grant No SP/19/3/34678, awarded to Health Data Research (HDR) UK) funded co-development (with NHS Digital) of the trusted research environment, provision of linked datasets, data access, user software licences, computational usage, and data management and wrangling support, with additional contributions from the HDR UK data and connectivity component of the UK Government Chief Scientific Adviser's National Core Studies programme to coordinate national Covid-19 priority research. Consortium partner organisations funded the time of contributing data analysts, biostatisticians, epidemiologists, and clinicians. AB, MAM, MHD and LP were supported by research funding from AstraZeneca. AB has received funding from the National Institute for Health Research (NIHR), British Medical Association, and UK Research and Innovation. AB, SD and HH are part of the BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No 116074. K.K. is supported by the

National Institute for Health Research (NIHR) Applied Research Collaboration East Midlands (ARC-EM) and NIHR Lifestyle BRC.

*Information Governance and ethics*

Approval for the study in CPRD was granted by the Independent Scientific Advisory Committee (20_074R) of the Medicines and Healthcare products Regulatory Agency in the UK in accordance with the Declaration of Helsinki. The North East-Newcastle and North Tyneside 2 research ethics committee provided ethical approval for the CVD- COVID-UK research programme (REC No 20/NE/0161).

The data used in this study are available in NHS Digital's TRE for England, but as restrictions apply they are not publicly available (https://digital.nhs.uk/coronavirus/coronavirus-data-services-updates/trusted-research-environment-service-for-england). The CVD-COVID-UK/COVID-IMPACT programme led by the BHF Data Science Centre (https://www.hdruk.ac.uk/helping-with-health-data/bhf-data-science-centre/) received approval to access data in NHS Digital's TRE for England from the Independent Group Advising on the Release of Data (IGARD) (https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/independent-group-advising-on-the-release-of-data) via an application made in the Data Access Request Service (DARS) Online system (ref. DARS-NIC-381078-Y9C5K) (https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services). The CVD-COVID-UK/COVID-IMPACT Approvals & Oversight Board (https://www.hdruk.ac.uk/projects/cvd-covid-uk-project/) subsequently granted approval to this project to access the data within NHS Digital's TRE for England. The de-identified data used in this study was made available to accredited researchers only.

The open-source code and utilised phenotype code-lists used this study are available in a repository in the British Hearth Foundation Data Science Centre's GitHub organisation (https://github.com/BHFDSC/CCU003_03).

*Guarantor*

Mehrdad A Mizani, 222 Euston Rd, London NW1 2DA, m.mizani@ucl.ac.uk

*Contributorship*

Research question, approach, and study oversight: AB. Leading data engineering, coding and analysis, and guarantorship: MAM. Data analysis, quality assurance and phenotyping: AD, JT, CT, AH, TB, JN. Study design and review: LP, SD, HH, CS, MJM, DC, CB, KK. Data visualisation: AGL. Coordinating approval for and access to data within NHS Digital's TRE for England for CVD-COVID-UK/COVID-IMPACT: CS. Drafting initial and final versions of manuscript: AB and MAM. Critical review of early and final versions of manuscript: All authors.

*Acknowledgements*
This work is carried out with the support of the BHF Data Science Centre led by HDR UK (BHF Grant no. SP/19/3/34678) and makes use of de-identified data held in NHS Digital's TRE for England, made available via the BHF Data Science Centre's CVD-COVID-UK/COVID-IMPACT consortium. This work uses data provided by patients and collected by the NHS as part of their care and support. We would also like to acknowledge all data providers who make health relevant data available for research.

## Abstract

*Objectives*
We use national, pre- and post-pandemic electronic health records (EHR) to develop and validate a scenario-based model incorporating baseline mortality risk, infection rate (IR) and relative risk (RR) of death for prediction of excess deaths.

*Design*
A data-driven retrospective cohort study.

*Setting*
Linked EHR in Clinical Practice Research Datalink (CPRD); and linked EHR and COVID-19 data in England provided in NHS Digital Trusted Research Environment (TRE).

*Participants*
In development (CPRD) and validation (TRE) cohorts, we included 3·8 million and 35·1 million individuals aged ≥30 years, respectively.

*Main outcome measures*
One year all-cause excess deaths related to COVID-19 from March 2020 to March 2021.

*Results*
From 1st March 2020 to 1st March 2021, there were 127,020 observed excess deaths. Observed RR was 4·34 (4·31-4·38, 95% CI) and IR was 6·27% (6·26-6·28, 95%CI). In the validation cohort, predicted one year excess deaths were 100,338 compared with the observed 127,020 deaths with a ratio of predicted to observed excess deaths of 0.79.

*Conclusions*
We show that a simple, parsimonious model incorporating baseline mortality risk, one year infection rate and relative risk of the pandemic can be used for scenario-based prediction of excess deaths in early stages of a pandemic. Our analyses show that EHR could inform pandemic planning and surveillance, despite limited use in emergency preparedness to-date. Although infection dynamics are important in prediction of mortality, future models should take greater account of underlying conditions.

## Introduction

Mortality estimates of COVID-19 have been the most reported and followed statistics at local, regional, national, and international levels since early in the pandemic, influencing policy and health service planning. Electronic health record (EHR) data informed early identification of risk factors for COVID-19 severity and mortality, leading to UK lockdown and shielding policies.[1–3] Moreover, EHR linkage enabled both specialist registry data and pragmatic clinical trials of new treatments at scale.[4,5]

Prediction of all-cause and disease-specific mortality in research and clinical practice has included underlying conditions or "baseline mortality risk", often derived and validated using EHR.[6–8] Underlying non-communicable diseases (NCDs) are important predictors of mortality in infectious diseases[9–11], but baseline mortality risk based on underlying NCDs is largely neglected in pandemic

preparedness, which emphasises transmissibility and severity of infection, using metrics such as case fatality ratio, infection fatality ratio and reproduction number.[12–16] Although the COVID-19 pandemic is increasingly conceptualised as a "syndemic"[17] (with interaction between infectious diseases and NCDs, requiring cross-speciality expertise), efforts to predict excess mortality have focused on dynamic transmission modelling without detailed consideration of baseline risk. Moreover, anonymised, individual-level, population-scale EHR have rarely been used for this purpose[18, 19].

On 22nd March 2020, before the first UK lockdown, we released a preprint (published on 12th May 2020)[1], estimating one year COVID-19 mortality using a model developed in pre-pandemic population-based linked EHR from 3·8 million people in the UK, obtained via the Clinical Practice Research Datalink (CPRD). Our generic model included baseline one year mortality risk for a range of underlying health conditions derived from the EHR. It incorporated scenario-based assumptions regarding relative risk (RR) of mortality during the pandemic compared to baseline, and population infection rate (IR). This approach requires validation in three ways. First, the actual RR and IR need to be established, to update scenario-based assumptions. Second, after incorporating observed IR and RR values, accuracy of model predictions needs to be assessed.

The NHS Digital Trusted Research Environment (TRE) for England, which became available during 2020 offers the opportunity to validate our approach at the whole population level in England, with longitudinal, individual-level data.[20,21] Therefore, using these data, we: (i) ascertained observed IR of COVID-19 and RR of one year COVID-19 mortality; (ii) compared predicted versus observed COVID-19 mortality.

## Methods

*Data sources*

*Abstract model development:* We used a pre-pandemic linked CPRD dataset, including EHR across primary care, hospital data and death registry with follow-up from 1997 to 2017.[1]

*Model validation:* We used the NHS Digital TRE for England which provides secure, remote access to linked, individual-level EHR data[20,21], including primary care, hospital episodes, registered deaths, COVID-19 laboratory tests, dispensed medicines, and COVID-19 vaccinations. We used General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR), Hospital Episode Statistics Admitted Patient Care (HES APC), Second Generation Surveillance System (SGSS), COVID-19 Hospitalisation in England Surveillance System (CHESS), Civil Registry Deaths, NHS Business Services Authority (NHSBSA) dispensed medicines, and COVID-19 vaccine datasets, prior to 15 May 2021.[21]

*Cohort specifications*

Both model development and validation involved population-based, retrospective cohort analyses with a range of high-risk conditions as exposures and one year all-cause mortality as outcome. In the

validation study, a further exposure was SARS-CoV-2 infection. In the development study, eligible individuals were aged ≥30 years, registered with a GP between 1st January 1997 and 1st January 2017, (**Figure S1.A)** with ≥1 year of follow-up.

In the validation study, eligible individuals were aged ≥30 years on 1st March 2018. The high-risk conditions for COVID-19-related outcomes was based on the Public Health England (PHE) guidance[22]. We considered all-cause mortality after COVID-19 as direct pandemic effect. Deaths in those without COVID-19 include baseline mortality and deaths attributable to indirect pandemic effects. To evaluate direct COVID-19 effects on one year all-cause mortality, we specified two time periods (**Figure S1.B** and **S1.C)**. The pre-pandemic period (1st March 2018-1st March 2019) was used for baseline characteristics and outcome (mortality) in the non-exposed (non-COVID-19) group. The pandemic period (1st March 2020-1st March 2021) was used to study COVID-19 cases and deaths in the exposed group (i.e. COVID-19 with or without high-risk conditions). Underlying conditions were assessed on 1st March 2018 in the validation study, minimising effect of age difference between pre-pandemic and pandemic periods (**Figure S2**).

*Exposures and outcomes of interest*

Exposures were presence (versus absence) of high-risk conditions for COVID-19[22] including cardiovascular disease (CVD), chronic kidney disease (CKD), diabetes, chronic obstructive pulmonary disease (COPD), body mass index (BMI) over 40kg/m$^2$, chronic liver disease, age>70 years, and history of oral steroid therapy. For all conditions, except steroid therapy, minimum period between earliest diagnosis date and baseline date (1st March 2018) was one year. For steroid therapy, event date was based on first dispensing date between 1st March 2018 and 1st March 2019, since prescription/dispensed medication data were only available from April 2018 onwards. Outcome was one year all-cause mortality.

To define underlying conditions, we used extended CALIBER phenotyping algorithms[23]. Phenotypes with earliest diagnosis dates between 1st March 2017 and 1st March 2018 were excluded, to allow ≥1 year history of conditions prior to cohort entry. The CVD phenotype was a composite, including heart failure, stroke (non-specified, ischaemic, haemorrhagic, transient ischaemic attack, subarachnoid haemorrhagic), arrhythmias, acute myocardial infarction, cardiomyopathy, atrial fibrillation, deep vein thrombosis, isolated calf vein thrombosis, and pulmonary embolism. The dispensed oral corticosteroid phenotype was determined based on the CALIBER phenotype mapped to British National Formulary codes.[24] To define COVID-19 cases, we used positive swab testing results and Public Health England labs and NHS hospitals, community swab testing results, primary care and hospital episode data, vaccination, and death registration.[25]

*Model development and validation*

Our prediction model in the development study was an abstract model based on baseline mortality, RR of death in those exposed to COVID-19 vs those not exposed to COVID-19 (pre-pandemic) and IR of COVID-19:

$$\frac{COVID\text{-}19\ related\ all\text{-}cause\ excess\ death\ count}{Baseline\ death\ count} = IR(RR-1)$$

In the development study, we calculated scenario-based COVID-19 excess deaths using baseline mortality by high-risk underlying conditions and plausible RR/IR (0·001%, 1%, 10% and 80% for total, partial, moderate, and no suppression)[2]. For each IR scenario, we applied RRs (1·2, 1·5, and 3), and scaled up to mid-2018 population of England aged ≥30 using estimates of the Office for National Statistics[26].

Validation of our approach involved use of observed IR and RR values (TRE for England; **Figure S1.B)** in the abstract model to predict COVID-19 deaths in development and validation cohorts. To capture direct COVID-19 mortality effects, we selected the unexposed and exposed groups in the pre-pandemic and pandemic periods respectively. We estimated baseline one year mortality in the pre-pandemic period (**Figure S1.B**) by Kaplan-Meier survival analysis. To calculate RR, we used and cross-validated pre-pandemic and pandemic periods to calculate baseline and COVID-19 mortality risk, respectively, by high-risk conditions. To calculate IR in each sub-sample, we divided the COVID-19 population by those at-risk at the start of the period. The final IR was the average of IRs of two sub-samples (refer to **Supplementary materials)**.

## Results

In validation cohort, we included 35,098,810 individuals (as shown in the CONSORT-based diagram in **Figure S2**) aged ≥30 years at baseline respectively. Of all individuals aged ≥30 years on 1st March 2018, 18,361,665 (52·3%) were female; mean age was 55·0 [SD 16·2] in both sexes; 28,049,984 (79·9%) were aged ≤70 years (mean 48·7 [SD 11·6] years in females and mean age 49·1 [SD 11·5] years in males) and 7,048,826 (20·1%) were >70 (mean 79·7 [SD 6·8] years in females and mean 78·5 [SD 6·1] years in males). Prevalence for CVD, diabetes, CKD, COPD, BMI>40, chronic liver disease and steroid therapy was 5.56% and 2.76%, 4.59% and 3.75%, 2.03% and 2.84%, 1.83% and 1.81%, 1.41% and 2.07%, 0.15% and 0.10%, and 3.52% and 5.07% in males and females, respectively. Prevalence of 0, 1, 2 and ≥3 underlying conditions was 35.57% and 39.95%, 8.15% and 8.48%, 8.82% and 2.79%, and 1.13% and 1.09% in males and females, respectively**.** Prevalence of all underlying conditions was higher in individuals >70 years and in males (**Figure 1** and **Table 1**).

*One year mortality*

Among individuals with at least one high risk condition, estimated pre-pandemic one year mortality risk was observed to be 3·55% (3·54-3·57). One year mortality risk in individuals >70 years was 9·24% (9·17-9·31), 3·37% (3·34-3·40), 8·36% (8·32-8·40) and 6·38% (6·34-6·42) for COPD, CKD, CVD

and diabetes, respectively. In individuals >70 years, one year mortality risks in men were 9·45% (9·35-9·55), 3·91% (3·85-3·96), 7·92% (7·98-9·20), 6·48% (6·42-6·54) for COPD, CKD, CVD, diabetes, respectively; and in women, 9·02 % (8·92-9·11), 3·00% (2·96-3·04), 8·84% (8·78-9·11), and 6·27% (6·21-6·33), respectively.

*Validation and replication of the abstract model*

In March 2020, we predicted 73,498 one year COVID-19 related deaths for the population of England, by scaling from the development cohort (3,862,012 aged ≥30) to the mid-2018 population of England and assuming a scenario of IR=10% and RR=3.[2] In the validation study, from March 2020 until March 2021, we ascertained 127,020 COVID-19 related all-cause deaths. We estimated pre-pandemic one year mortality risk by age group, sex, and number of high-risk conditions in the absence of COVID-19.

We calculated cross-validated one year (March 2020-2021) RR and IR of COVID-19 as 4·34 (4·31-4·38, 95% CI) and 6·27% (6·26-6·28, 95%CI), respectively. **Table S1** and **S2** show cross-validated IR and RR, respectively, across two random subsamples of the cohort as shown in **Figure S1**. **Table S3** shows sensitivity analysis for underfitting and further cross-validation. We found that the effect of vaccination on the overall RR or IR between December 2020 and March 2021 was negligible compared to the effects of under-reported COVID-19 cases from the pre-vaccination period (**Table S4)**. We applied our prediction model using observed values for RR (4.34) and IR (6.27) and baseline mortality risk data in validation cohort (**Table S5 and S6**).

**Figures 3** and **S4** show predicted one year COVID-19 related all-cause deaths, based on baseline mortality risk (March 2018-2019 for validation cohort), RR=4·34, and IR=6·27% compared to observed excess deaths (March 2020-2021). Observed COVID-19 deaths numbered 127020. Model-predicted COVID-19 deaths were 100338 , 79.0% of the observed value (**Table 2, Figure 3**).

## Discussion

In anonymised, individual-level, population-scale, national EHR data between March 2020 and March 2021, we conducted the first study to predict and validate one year mortality among those with COVID-19 using baseline (pre-pandemic) mortality risk. We provide the first detailed, scenario-based mortality risk assessment before and during the pandemic, based on absolute risk estimates in national population data. We show that a simple, parsimonious model incorporating baseline risk of mortality, infection rate and relative risk of the pandemic can be used to predict one year COVID-19 mortality.

*Strengths and weaknesses*

Our analysis uses anonymised, national, individual-level EHR data with unprecedented scale and whole population inclusivity and validated EHR phenotypes. It highlights the importance of EHR data, baseline mortality, and scenario-based assumptions in risk assessment at early stages of a pandemic where dynamics of the new infectious disease are not yet known.

Our analysis used only the most frequent high-risk conditions. Our simple model made assumptions regarding static RR and IR over the course of the pandemic and did not incorporate infectivity or population dynamics of the original or later strains of SARS-CoV2, the impact of COVID-19-related policies or vaccination rates. Generalisability of our findings to other countries and contexts requires further validation. Our study only investigated COVID-19 and applicability to other infectious diseases or pandemics is unknown. There are differences between development and validation cohorts in terms of data coding systems (e.g. lack of standardised one-to-one mapping between coding terminologies), , and limited availability of fields in CPRD (e.g. ethnicity) and in the TRE for England (e.g. medication use before 2018 and multiple index of deprivation), which restricted analyses. Overall, national mortality estimates in people with COVID-19 were similar in development and validation cohorts, with differences in mortality risk at baseline in stratified analyses. For example, mortality risk was similar for younger people in both cohorts, but mortality risk was relatively higher in the development cohort for individuals >70 years due to the earlier cohort entry date in the CPRD study population.[1] Also, number of estimated deaths was lower in the development cohort in all age categories, perhaps because one year mortality in CPRD data was calculated after study entry date, when these individuals were younger (mean age 43·5 [SD 11·7] years), compared to the validation cohort in March 2018-2019 (mean age 55·0 [SD 16·2] years). Another explanation is that actual IR over one year is higher than our observed rate (and probably greater than the 10% we used in prediction), due to incomplete availability of testing, especially during early months of the pandemic.

*Strengths and weaknesses in relation to other studies*

We searched for systematic reviews published after March 2020 in PubMed for combinations of equivalent Mesh terms of “COVID”, “prediction”, “mortality, “model”, “underlying condition”, “relative risk”, and “infection rate”. A systematic review of 107 multi-variate prediction models for COVID-19 mortality showed that variables were selected from signs, symptoms, and risk factors from COVID-19 patients during the pandemic[27]. All models had unclear or high risk of bias, including non-representative data sources, unreliable COVID-19 case definition, excluding patients who had not experienced the outcome of interest, and model overfitting. We found no studies of excess mortality prediction based on pre-pandemic mortality in people with high-risk underlying conditions and RR and IR associated with COVID-19. In our study design, all patients, regardless of outcome of interest, were included in analyses. Furthermore, we conducted model cross-validation to minimise overfitting (**Table S3**).

We used EHR data of the whole population in England to validate our model for predicting one year excess mortality in people exposed to COVID-19. The data used in our study is derived from anonymised, individual-level, and linked EHR of the whole population in England, making our model highly representative. We have used validated phenotype definitions for high-risk underlying conditions and COVID-19 cases. Our study highlights the significance of pre-pandemic longitudinal EHR data to predict direct effects of the pandemic for preparedness and early response.

Our model is an abstract simple model for formulating worst-case to best-case scenarios at the start of the pandemic. We developed the model in CPRD data with assumed parameters and replicated the model in NHS Digital TRE using observed RR and IR values. Hence, our model is more suitable for risk assessment for pandemic preparedness and early response rather than high-precision estimation of the mortality.

*Meaning of the study: possible mechanisms and implications*

Pre-pandemic mortality risks: Baseline mortality risk can be used to predict COVID-19 related mortality over one year at national level, and underlying conditions and age are major determining factors of the risk. We show that national data EHR, such as the NHS Digital TRE, and sampled less complete data, such as CPRD, can be used to estimate and monitor baseline risk at scale. Such data are available across diseases, risk factors and countries via the Global Burden of Disease Study and other efforts and have already been used to project high-risk populations for COVID-19[28]. There is public demand for such information and it can be provided in an interpretable, usable format employing open phenotypes, coding and standards[20-23, 29].

Infection rate over one year: Surveillance of SARS-CoV-2 infection rates has been crucial across countries throughout the pandemic by different methods, including incident or prevalent cases, over weeks or months, by antigen or antibody tests, or by static or dynamic rates. Our model used population IR over one year, which we estimated using comprehensive testing, primary care, hospital data and death data in the NHS Digital TRE in a mostly pre-vaccination era. Our estimates of IR represent nearly the whole English population, consistent with pre-vaccination antibody rates in the UK[30] and a recent study using the same data[25]. However, underestimation is still possible and, moreover, likely, due to initially limited testing capacity and asymptomatic infection. Future research and models should incorporate higher vaccination rates, novel variants, potential impact of reinfection, and dynamic infection rates over time.

Relative risk associated with the pandemic: Excess mortality associated with COVID-19 has been a focus in health policy since the early stages of the pandemic. Comparisons with flu persist until now, including "winter excess deaths" which have been estimated as 20% higher than baseline mortality rate[1]. In our model, we used RR estimates of 1.5, 2 and 3, and in national data, we observed 4.34 in the overall population. Assuming under-estimation of IR, we may have over-estimated RR, but our estimates are in line with a recent time-series analysis of excess mortality in the first pandemic wave in the UK. That study showed that certain underlying conditions were associated with higher RR of excess pandemic mortality, compared with pre-pandemic period[31].

***Implications for public health and policy makers***

There are three public health and policy implications. First, EHR were designed and used for reimbursement, clinical care and quality improvement, with limited use in emergency preparedness. Our analyses show that EHR could and should be part of pandemic planning and surveillance. Second, pre-pandemic mortality risk can be estimated at individual, subgroup, and national levels, and is important in the prediction of mortality during pandemics as well as preparedness including shielding and vaccination prioritisation. Third, our data support the syndemic lens which views COVID-19 not just as an infectious disease, but one with social, environmental, and non-communicable disease

determinants and effects, signalling need for multidisciplinary approaches to public health and policy in this and future pandemics.

Research implications and future steps: There are four unanswered aspects in our research. First, moderate and high-risk conditions for COVID-19, outlined by the UK government[22], number more than 80 diseases, risk factors and underlying conditions. We will be validating estimates of COVID-19 mortality for the comprehensive list of conditions and providing condition-specific IR and RR estimates stratified by conditions, ethnicity, deprivation, and vaccination, which will be useful in refining future models for the analyses of COVID-19 and other pandemics. Second, the need for region- and country-specific data for health policy is well-recognised, and our analyses regarding IR, RR, and validation of our model need to be conducted in other countries and datasets to investigate generalisability of our findings. Third, we have only considered direct impact of the pandemic on mortality. The major indirect and long-term (Long COVID) impact of the COVID-19 public health emergency on health systems need to be studied and incorporated into future models of potential effects of pandemics. Fourth, there is great scope for combining baseline mortality risk estimation (using models such as ours) with existing methods of dynamic transmission modelling to predict and reduce the suffering caused by future pandemics.

## Conclusions

The major impact of the COVID-19 pandemic on excess mortality can be predicted using national electronic health records and is related to baseline mortality risk, population infection rates and pandemic-associated relative risk. There are significant implications for public health, policy and research in terms of expertise, data and resources for future pandemic preparedness.

## References


1. Banerjee A, Pasea L, Harris S, et al. Estimating excess 1-year mortality associated with the COVID-19 pandemic according to underlying conditions and age: a population-based cohort study. *Lancet* 2020; **395(10238):**1715–1725.
2. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020; **584(7821):** 430–436.
3. Clift AK, Coupland CAC, Keogh RH, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ* 2020; **371:** m3731.
4. Docherty AB, Harrison EM, Green CA, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* 2020; **369:** m1985.
5. RECOVERY Collaborative Group. Dexamethasone in Hospitalized Patients with Covid-19. *N Engl J Med* 2021; **384(8):** 693–704.
6. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008; **336(7659):** 1475–82.
7. Vogelsang RP, Bojesen RD, Hoelmich ER, et al. Prediction of 90-day mortality after surgery for colorectal cancer using standardized nationwide quality-assurance data. *BJS Open* 2021; **5(3):** zrab023.

8. Ajnakina O, Agbedjro D, McCammon R, et al. Development and validation of prediction model to estimate 10-year risk of all-cause mortality using modern statistical learning methods: a large population-based cohort study and external validation. *BMC Med Res Methodol* 2021; **21(1):** 1–11.
9. Bolge SC, Kariburyo F, Yuce H, Fleischhackl R. Predictors and Outcomes of Hospitalization for Influenza: Real-World Evidence from the United States Medicare Population. *Infect Dis Ther* 2021; **10(1):** 213–28.
10. Ma HM, Tang WH, Woo J. Predictors of in-hospital mortality of older patients admitted for community-acquired pneumonia. *Age Ageing* 2011; **40(6):** 736–41.
11. Bastos HN, Osório NS, Castro AG, et al. A Prediction Rule to Stratify Mortality Risk of Patients with Pulmonary Tuberculosis. *PLoS One* 2016; **11(9):** e0162797.
12. Horton R. Offline: COVID-19 is not a pandemic. *Lancet* 2020; **396(10255):** 874.
13. Portugal L. Mortality and Excess Mortality: Improving FluMOMO. *J Environ Public Health* 2021; **2021:** 5582589.
14. Huppert A, Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. *Clin Microbiol Infect* 2013; **19(11):** 999–1005.
15. Biggerstaff M, Cowling BJ, Cucunubá ZM, et al. WHO COVID-19 Modelling Parameters Group. Early Insights from Statistical and Mathematical Modeling of Key Epidemiologic Parameters of COVID-19. *Emerg Infect Dis* 2020; **26(11):** e1–e14.
16. Laydon DJ, Mishra S, Hinsley WR, et al. Modelling the impact of the tier system on SARS-CoV-2 transmission in the UK between the first and second national lockdowns. *BMJ Open* 2021; **11(4):** e050346.
17. Verity R, Okell LC, Dorigatti I, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis* 2020; **20(6):** 669–77.
18. Banerjee A, Chen S, Pasea L, et al. Excess deaths in people with cardiovascular diseases during the COVID-19 pandemic. *Eur J Prev Cardiol* 2021; **28(14):** 1599–609.
19. Lai AG, Pasea L, Banerjee A, et al. Estimating excess mortality in people with cancer and multimorbidity in the COVID-19 emergency. *BMJ Open* 2020; **10(11):** e043828.
20. Wood, A., Denholm, R., Hollings, S., et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021; **373:** n826.
21. CVD-COVID-UK/COVID-IMPACT TRE Dataset Provisioning Dashboard, British Heart Foundation Data Science Centre, Health Data Research UK, https://www.hdruk.ac.uk/wp-content/uploads/2022/02/220210-CVD-COVID-UK-COVID-IMPACT-TRE-Dataset-Provisioning-Dashboard.pdf (accessed 11 Feb 2022)
22. Who is at high risk from coronavirus (COVID-19), https://www.nhs.uk/conditions/coronavirus-covid-19/people-at-higher-risk/who-is-at-high-risk-from-coronavirus/ (accessed 1 Feb 2022)
23. Denaxas S, Gonzalez-Izquierdo A, Direk K et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER.J Am Med Inform Assoc. 2019 Dec 1;26(12):1545-1559.
24. OpenPrescribing, https://openprescribing.net/bnf/0603/ (accessed 1 Feb 2022)
25. Thygesen JH, Tomlinson C, Hollings S et al. COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records. Lancet Digit Health. 2022 Jun 8:S2589-7500(22)00091-7. doi: 10.1016/S2589-7500(22)00091-7.
26. Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland, Office for National Statistics, https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland (accessed 1 Feb 2022)

27. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020; **369**: m1328.
28. Clark A, Jit M, Warren-Gash C et al. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. Lancet Glob Health. 2020 Aug;8(8):e1003-e1017.
29. Banerjee A, Pasea L, Manohar S et al. 'What is the risk to me from COVID-19?': Public involvement in providing mortality risk information for people with 'high-risk' conditions for COVID-19 (OurRisk.CoV). Clin Med (Lond). 2021 Nov;21(6):e620-e628.
30. Coronavirus (COVID-19) Infection Survey: characteristics of people testing positive for COVID-19 in England and antibody data for the UK: December 2020. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsinthecommunityinengland/december2020 (accessed 02/02/2022)
31. Strongman H, Carreira H, De Stavola BL, Bhaskaran K, Leon DA. Factors associated with excess all-cause mortality in the first wave of the COVID-19 pandemic in the UK: A time series analysis using the Clinical Practice Research Datalink. PLoS Med. 2022 Jan 6;19(1):e1003870.

## Figures and Tables

Figure 1. Prevalence of high-risk conditions for COVID-19 mortality in validation cohort (n=35,098,810) cohort aged ≥ 30 years.

Figure 2. Baseline one year mortality in England (age ≥ 30) according to underlying conditions in validation cohort (n=35,098,810)

Figure 3. Baseline deaths, model-predicted COVID-19 related all-cause deaths, and observed deaths among those with COVID-19 in England (age ≥ 30) over one year, stratified by age and sex in validation cohort (n=35,098,810)

Table 1. Underlying conditions in the validation cohort (NHS Digital TRE, n= 35,098,810, aged 30 years or older)

Table 2. Observed COVID-19 one year mortality in England (NHS Digital TRE; n = 35,098,810 aged ≥30 years; 1st March 2020 to 1st March 2021)

**Figure 1 Prevalence of high-risk conditions for COVID-19 mortality in validation cohort (n=35,098,810) aged ≥ 30 years.**

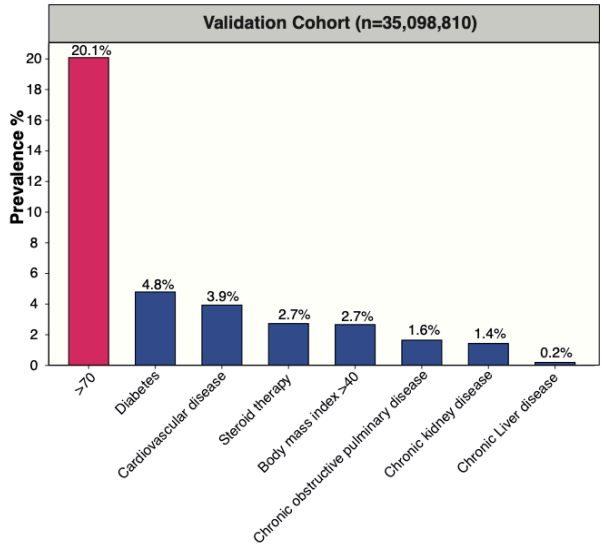
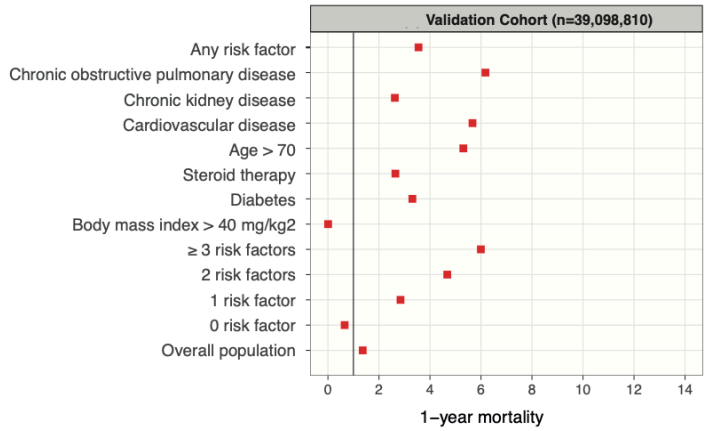


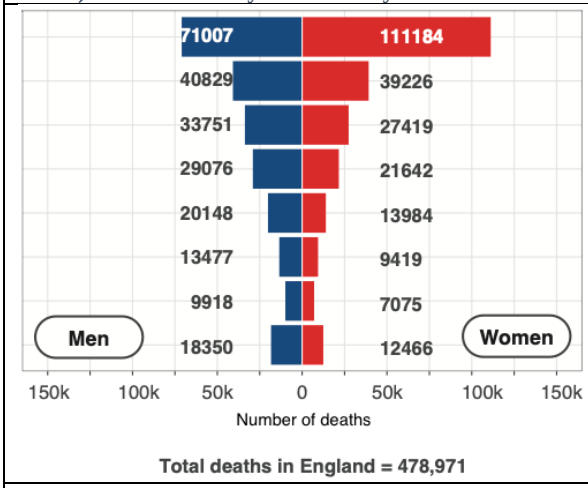

**Figure 2 Baseline one year mortality in England (age ≥ 30) according to underlying conditions in validation cohort (n=35,098,810)**




| n (%) | Deaths | 1-year mortality % (95% CI) |
|---|---|---|
| 8590341 (24.47) | 305121 | 3.55 (3.54-3.57) |
| 1279728 (6.65) | 79005 | 6.17 (6.13-6.22) |
| 1707330 (4.86) | 44790 | 2.62 (2.60-2.65) |
| 3431847 (9.78) | 194500 | 5.67 (5.64-5.69) |
| 7048826 (20.08) | 374134 | 5.31 (5.29-5.32) |
| 3014906 (8.59) | 79760 | 2.65 (2.63-2.66) |
| 2929508 (8.35) | 96993 | 3.31 (3.29-3.33) |
| 1220688 (3.48) | 86 | 0.007 (0.006-0.009) |
| 780273 (2.22) | 46792 | 6.00 (5.94-6.05) |
| 1972993 (5.62) | 92313 | 4.68 (4.65-4.71) |
| 5837075 (16.63) | 173759 | 2.85 (2.83-2.86) |
| 26508469 (75.52) | 173759 | 0.655 (0.652-0.659) |
| 35098810 (100) | 478971 | 1.36 (1.36-1.37) |

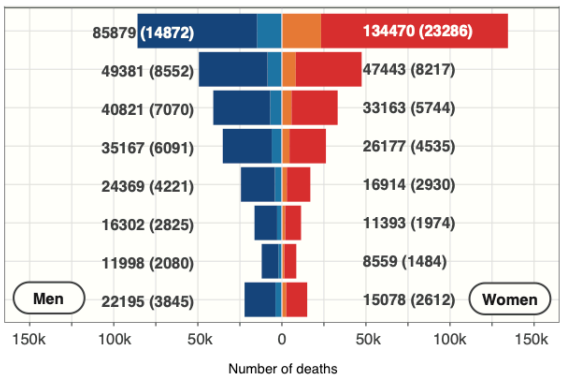
**Figure 3 Baseline deaths, model-predicted COVID-19 related all-cause deaths, and observed deaths among those with COVID-19 in England (age ≥ 30) over one year, stratified by age and sex validation cohort (n=35,098,810)**

a) Baseline one year mortality



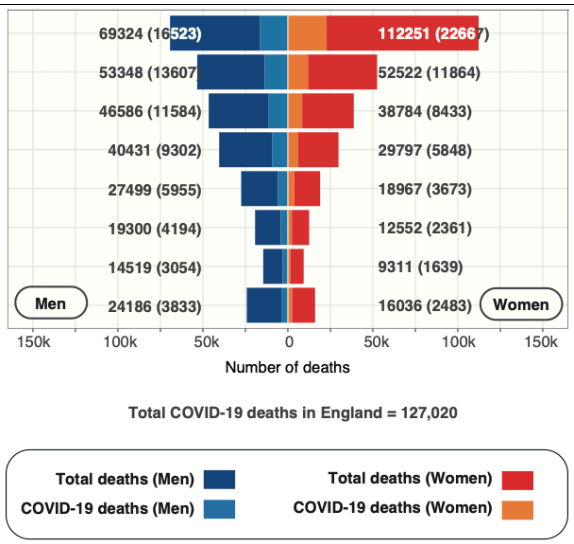

Total deaths in England = 478,971

b) Total (and model-predicted COVID-19) one year mortality based on relative risk=4.34, infection rate=6.27%




Total COVID-19 deaths in England = 100,338

c) Total (and observed COVID-19) one year mortality




Total COVID-19 deaths in England = 127,020

**Table 1 Underlying conditions in the validation cohort (NHS Digital TRE, n= 35,098,810, aged 30 years or older)**

| | Count (% of total population) | | | | | |
|---|---|---|---|---|---|---|
| **Underlying condition** | **Male Age ≤ 70 years N = 13587089** | **Male Age > 70 years N = 3150056** | **Male All ages N= 16737145** | **Female Age ≤ 70 years N =14462895** | **Female Age > 70 years N = 3898770** | **Female All ages N=18361665** |
| **CVD** | 873001 (2·49) | 1080487 (3·08) | 1953488 (5·56) | 509450 (1·45) | 968909 (2·76) | 1478359 (4·21) |
| **Diabetes** | 965436 (2·75) | 647269 (1·84) | 1612705 (4·59) | 716309 (2·04) | 600494 (1·71) | 1316803 (3·75) |
| **CKD** | 227924 (0·65) | 483972 (1·38) | 711896 (2·03) | 274852 (0·78) | 720582 (2·05) | 995434 (2·84) |
| **COPD** | 291294 (0·83) | 351684 (1·00) | 642978 (1·83) | 287287 (0·82) | 349463 (0·99) | 636750 (1·81) |
| **BMI>40** | 373213 (1·06) | 120512 (0·34) | 493725 (1·41) | 561351 (1·60) | 165612 (0·47) | 726963 (2·07) |
| **Chronic liver disease** | 42789 (0·12) | 10966 (0·03) | 53755 (0·15) | 25807 (0·07) | 9875 (0·03) | 35682 (0·10) |
| **Steroid therapy** | 762449 (2·17) | 472571 (1·35) | 1235020 (3·52) | 1183308 (3·37) | 596578 (1·70) | 1779886 (5·07) |
| | | | | | | |
| **0** | 11167965 (31·82) | 1317372 (3·75) | 12485337 (35·57) | 12137332 (35·58) | 1885800 (5·37) | 14023132 (39·95) |
| **1** | 1835747 (5·23) | 1025674 (2·92) | 2861421 (8·15) | 1800831 (5·13) | 1174823 (3·35) | 2975654 (8·48) |
| **2** | 451492 (1·29) | 541211 (1·54) | 992703 (8·82) | 406504 (1·16) | 573786 (1·63) | 980290 (2·79) |
| **≥3** | 131885 (0·38) | 265799 (0·76) | 397684 (1·13) | 118228 (0·34) | 264361 (0·75) | 382589 (1·09) |

**Table 2 Observed COVID-19 one year mortality in England (NHS Digital TRE; n = 35,098,810 aged ≥30 years; 1st March 2020 to 1st March 2021)**

| | Age ≤ 70 years | | | | Age > 70 years | | | | All ages | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **N total (%)** | **Total deaths** | **N COVID (%)** | **COVID deaths** | **N total (%)** | **Total deaths** | **N COVID (%)** | **COVID deaths** | **N total (%)** | **Total deaths** | **N COVID (%)** | **COVID deaths** |
| ≥1 underlying condition excluding age > 70 years | 4634608 (13·58) | 70202 | 314587 (0·92) | 16203 | 3340209 (9·79) | 317114 | 209190 (0·61) | 74276 | 7974817 (23·36) | 70479 | 523777 (1·53) | 90479 |
| Age > 70 years | - | - | - | - | 6299844 (18·46) | 443043 | 317798 (0·93) | 99828 | - | - | - | - |
| Diabetes | 1645037 (4·82) | 29688 | 123984 (0·36) | 8338 | 1087148 (3·18) | 106124 | 75870 (0·22) | 27474 | 2732158 (8·00) | 135902 | 199854 (0·58) | 35812 |
| CVD | 1335614 (3·91) | 30301 | 80174 (0·23) | 6966 | 1710348 (5·01) | 200644 | 121772 (0·36) | 46744 | 3045962 (8·92) | 230945 | 201946 (0·59) | 53710 |
| BMI > 40 | 932120 (2·73) | 8454 | 73399 (0·21) | 2333 | 280331 (0·82) | 19410 | 15690 (0·04) | 4911 | 1212451 (3·55) | 27864 | 89089 (0·26) | 7244 |
| Steroid therapy | 1889695 (5·54) | 44671 | 149685 (0·44) | 7655 | 923584 (2·70) | 111144 | 67321 (0·20) | 24354 | 2813279 (8·24) | 155815 | 217006 (0·64) | 32009 |
| COPD | 549304 (1·61) | 18905 | 29797 (0·09) | 3733 | 574369 (1·68) | 70701 | 43183 (0·13) | 16872 | 1123673 (3·29) | 89606 | 72980 (0·21) | 20605 |
| CKD | 492763 (1·44) | 11102 | 33377 (0·10) | 3255 | 1100918 (3·25) | 121830 | 75622 (0·22) | 29332 | 1593680 (4·67) | 132932 | 108999 (0·32) | 32587 |
| Chronic liver disease | 60270 (0·18) | 3584 | 3769 (0·18) | 556 | 15556 (0·04) | 2291 | 1213 (0·003) | 483 | 75826 (0·22) | 5875 | 4982 (0·01) | 1039 |
| 3+ underlying conditions | 233799 (0·68) | 12645 | 18267 (0·05) | 1470 | 442569 (1·30) | 67507 | 40625 (0·12) | 17304 | 676368 (1·98) | 80152 | 58892 (0·17) | 20774 |
| 2 underlying conditions | 827803 (2·472) | 20516 | 55977 (0·16) | 4885 | 956907 (2·80) | 104452 | 66645 (0·19) | 24693 | 1784710 (5·23) | 124968 | 122622 (0·36) | 29578 |
| 1 underlying condition | 3573006 (10·47) | 37041 | 240343 (0·70) | 7848 | 1940733 (5·68) | 145155 | 101920 (0·30) | 32279 | 5513739 (16·15) | 182196 | 342263 (1·00) | 40127 |
| No underlying condition | 23197624 (47·96) | 72168 | 1615026 (4·73) | 10989 | 2959635 (8·67) | 125929 | 108608 (0·32) | 23869 | 26157259 (76·63) | 198097 | 1723634 (5·05) | 36541 |
| Overall population | 27832232 (81·54) | 142370 | 1929613 (5·65) | 27192 | 6299844 (18·46) | 443043 | 317798 (0·93) | 99828 | 34132076 100) | 585413 | 2247411 (6·58) | 127020 |