**Title**

Using national electronic health records for pandemic preparedness: validation of a parsimonious model for predicting excess COVID-19 deaths

**Version history**

|  |  |  |
| --- | --- | --- |
| V0.1 | 15/12/2021 | First version for GitHub |
|  |  |  |
|  |  |  |

**Lay summary**

Predicting the number of COVID-19 deaths has been the focus of many studies since the start of the pandemic. These predictions provide crucial information to guide government policies related to lockdown and social distancing measures. The most common prediction models for infectious diseases are based on parameters not known at the start of a pandemic (for example, the attack rate, and R0) rather than existing historical patient data in Electronic Health Records (EHR). In March 2020, we developed a prediction model for COVID-19 deaths in one year based on chronic conditions of patients1. We used the anonymised EHR of 3.8 million patients provided by Clinical Practice Research Datalink (CPRD)2-5. We made predictions in different scenarios based on the spread of COVID-19 and its severity in terms of increased risk of death.

In this project, we aim to validate our prediction model on approximately 44.9 million patients aged 30 and older in the Trusted Research Environment (TRE) for England. Our approach predicts one year of death in people infected with COVID-19 from March 2020 to March 2021 based on pre-pandemic (from March 2018 to March 2019) health data in people with high-risk chronic diseases. We are specifically interested in chronic diseases listed in the shieling list by NHS6-7.

**Background**

There have been over 5 million deaths due to COVID-19 globally, with more than 144,000 in the UK8. Estimates of mortality have been the most reported and used statistics since the start of the pandemic, influencing policy and planning. Use of EHR informed early identification of risk factors for COVID-19 mortality, leading to social distancing and lockdown in the UK9-10. One of the crucial predictors of mortality in infectious disease is the existence of underlying chronic diseases. However, baseline mortality risk in people with underlying conditions is not used in common models for estimating the mortality of infection diseases. Most of the predictive models for mortality of infectious diseases, including COVID-19, are based on transmissibility, the severity of the infection, case fatality ratio, infection fatality ratio and R-0, many of which are unknown at the start of the pandemic.

In March 2020, we estimated 1-year COVID-19 mortality in England using the EHR of 3.8 million patients provided by Clinical Practice Research Datalink (CPRD). The main parameters in our predictive model were baseline all-cause mortality in people with underlying chronic conditions and combinations of scenario-based values of relative risk (RR) of mortality and infection rate (IR) of COVID-19.

This project aims to validate our previous model in 44.9 million patients aged 30 and older in Trusted Research Environment (TRE) for England. In our development and validation models, we are interested in the direct effects of the pandemic on 1-year all-cause mortality in people with underlying chronic conditions listed at high-risk factors for COVID-19.

**Research aims and question(s)**

1. What is the 1-year all-cause mortality in patients with CVD and high-risk chronic conditions from March 2018 to March 2019?
2. What are the observed RR and IR in individuals with high-risk chronic conditions from March 2020 to March 2021?
3. Based on our model, what is the estimated excess mortality in people with COVID-19 infection?
4. What is the ratio of estimated excess mortality to observed COVID-19 mortality?

**Data sources**

NHS England (TRE)

HES: Hospital Episode Statistics

ONS: Deaths

GDPPR: GPES Data for Pandemic Planning and Research

COVID-19 SGSS: Second Generation Surveillance System (+ve results from pillars 1 and 2)

CHESS: COVID-19 Hospitalisation in England Surveillance System

NHS BSA: Dispensed Medicines

Vaccination: COVID-19 vaccine dataset

**Study design**

The study design is a retrospective cohort. The whole cohort will be randomly divided into two sub-samples. Cases will be selected from the pandemic period (March 2020 to March 2021) in the first (then second) sub-sample. Controls will be selected from the pre-pandemic period (March 2018 to March 2019) in the second (then first) sub-sample. The results will be cross-validated and averaged across two subsamples.

**Study population and cohort specification**

Our study is limited to individuals linked across all sources.

Case individuals: All individuals aged 30 or older at 1st March 2018 infected with COVID-19 from 1st March 2020 to 1st March 2021, with existing values for date of birth, date of death (if dead), date of COVID-19, date of the first diagnosis of selected chronic conditions.

Control individuals: All individuals aged 30 or older at 1st March 2018 with existing values for date of birth, date of death (if dead), date of the first diagnosis of selected chronic conditions.

**Exposures and outcomes of interest**

Exposures of interest are the prevalence of high-risk conditions for COVID-19 complications35 including cardiovascular disease (CVD), chronic kidney disease (CKD), diabetes, chronic obstructive pulmonary disease (COPD), body mass index (BMI) over 40kg/m2, being older than 70, chronic liver disease, and history of oral steroid therapy.

The outcome of interest is 1-year all-cause mortality.

**Covariates**

The main covariates are age, sex, and the number of chronic conditions.

**Model development and statistical analysis**

The model is based on RR and IR, where the ratio of COVID-19 excess death to baseline death is IR time RR-1. We will use Kaplan-Meier (KM) survival analysis to estimate baseline mortality over 1-year (1st March 2018 to 1st March 2019). To calculate the RR, we will use the Chi-Square test using the mortality of COVID-19 in cases in the first sub-sample (refer to study design) versus mortality of pre-pandemic controls in the second sub-sample and vice versa. To calculate IR, we will use simple counts of COVID-19 infections from 1st March 2020 to 1st March 2021.

**Sensitivity analysis and model validation**

1. Vaccination: COVID-19 vaccination programmes started in mid-December 2020. Some cases in the last four months of the study will be vaccinated, which might be a confounding factor in assessing mortality risk. To control for this confounding factor, we will calculate quarterly RR and IR to evaluate the effects of vaccination on quarterly and overall results.
2. We will evaluate the KM survival analysis model by adjusting for age, sex, the number of underlying conditions, and combinations.
3. For internal validation of the estimation of RR against over-fitting and under-fitting, we will train and test the model of various ratios of training to validation sets.

**References**

1. Banerjee A, Pasea L, Harris S, Gonzalez-Izquierdo A, Torralbo A, Shallcross L, Noursadeghi M, Pillay D, Sebire N, Holmes C, Pagel C, Wong WK, Langenberg C, Williams B, Denaxas S, Hemingway H. Estimating excess 1-year mortality associated with the COVID-19 pandemic according to underlying conditions and age: a population-based cohort study. Lancet. 2020 May 30;395(10238):1715-1725.
2. Herrett, E., Gallagher, A.M., Bhaskaran, K., Forbes, H., Mathur, R., van Staa, T., Smeeth, L., 2015. Data Resource Profile: Clinical Practice Research Datalink (CPRD). International Journal of Epidemiology 44, 827–836. <https://doi.org/10.1093/ije/dyv098>
3. Mathur, R., Bhaskaran, K., Chaturvedi, N., Leon, D.A., vanStaa, T., Grundy, E., Smeeth, L., 2014. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. Journal of Public Health 36, 684–692. <https://doi.org/10.1093/pubmed/fdt116>
4. Gallagher, A.M., Dedman, D., Padmanabhan, S., Leufkens, H.G.M., de Vries, F., 2019. The accuracy of date of death recording in the Clinical Practice Research Datalink GOLD database in England compared with the Office for National Statistics death registrations. Pharmacoepidemiol Drug Saf 28, 563–569. <https://doi.org/10.1002/pds.4747>
5. L. McDonald, A. Schultze, R. Carroll, and S. V. Ramagopalan, ‘Performing studies using the UK Clinical Practice Research Datalink: to link or not to link?’, Eur J Epidemiol, vol. 33, no. 6, pp. 601–605, Jun. 2018, doi: [10.1007/s10654-018-0389-5](https://doi.org/10.1007/s10654-018-0389-5).
6. 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/>
7. Guidance for people previously considered clinically extremely vulnerable from COVID-19<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>
8. Worldometer. [https://www.worldggggometers.info/coronavirus/](https://www.worldometers.info/coronavirus/)
9. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, Curtis HJ, Mehrkar A, Evans D, Inglesby P, Cockburn J, McDonald HI, MacKenna B, Tomlinson L, Douglas IJ, Rentsch CT, Mathur R, Wong AYS, Grieve R, Harrison D, Forbes H, Schultze A, Croker R, Parry J, Hester F, Harper S, Perera R, Evans SJW, Smeeth L, Goldacre B. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020 Aug;584(7821):430-436.
10. Clift AK, Coupland CAC, Keogh RH, Diaz-Ordaz K, Williamson E, Harrison EM, Hayward A, Hemingway H, Horby P, Mehta N, Benger J, Khunti K, Spiegelhalter D, Sheikh A, Valabhji J, Lyons RA, Robson J, Semple MG, Kee F, Johnson P, Jebb S, Williams T, Hippisley-Cox J. 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 Oct 20;371:m3731.