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 Mizani1,9, Ashkan Dashtban1, Laura Pasea1, Alvina G Lai1, Johan Thygesen1, Chris Tomlinson1, Alex Handy1, Jil B Mamza2, Tamsin Morris2, Sara Khalid3, Francesco Zaccardi4, Mary Joan Macleod5, Fatemeh Torabi6, Dexter Canoy7, Ashley Akbari6, Colin Berry8, Thomas Bolton9, John Nolan9, Kamlesh Khunti4, Spiros Denaxas1, Harry Hemingway1, Cathie Sudlow9, Amitava Banerjee1, on behalf of the CVD-COVID-UK Consortium.

1Institute of Health Informatics, University College London, London, UK.

2Medical and Scientific Affairs, BioPharmaceuticals Medical, AstraZeneca, Cambridge, United Kingdom.

3Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK.

4Leicester Diabetes Centre, University of Leicester, Leicester, UK.

5School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, UK.

6Faculty of Medicine, Health and Life Science, Swansea University

7Nuffield Department of Women’s and Reproductive Health, University of Oxford, Oxford, UK

8Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK

9BHF Data Science Centre, Health Data Research UK, London, UK.

## Corresponding Author

Amitava Banerjee,  222 Euston Rd, London NW1 2DA, [ami.banerjee@ucl.ac.uk](mailto: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

Amitava Banerjee,  222 Euston Rd, London NW1 2DA, [ami.banerjee@ucl.ac.uk](mailto:ami.banerjee@ucl.ac.uk)

## Contributorship

Research question, approach, and study oversight: AB. Leading data engineering, coding and analysis: 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.

# 