## Minimising bias in ethnicity data v0.1

## Version history

|  |  |  |
| --- | --- | --- |
| v0.1 | 17/01/2022 | First protocol draft including the two objectives of the project CCU0037 |
| v0.2 | 15/02/2022 | Draft splitting and creation of the individual protocols: CCU037\_001 and CCU037\_002.  From now on, following versions in this history will belong to CCU037\_002. |
|  |  |  |

## Lay summary

Inequality in health has been highlighted by the COVID-19 pandemic, where people from ethnically diverse backgrounds were disproportionately affected. But we know inequity is not limited to the pandemic as it is a long-standing, multi-faceted issue.

An example is technology for predicting a person’s future health risks. This involves routinely collected health information, which is fed into a computer model which in turn produces a health risk score for a patient, and that is used by doctors to decide patient care. If there is bias in the data or bias in the model, the doctor can potentially make wrong decisions and patients can get the wrong care or no care, meaning some groups of patients might inappropriately be prioritised over other for booster vaccines, hospital beds, or life-saving treatments, which in turn can affect patient and public trust, and cost the NHS.

This proposal aims to improve existing technology for predicting personalised future risk of health conditions, particularly those affecting overlooked groups of patients. We aim to do so by a) improving the way recorded ethnicity is used in research, and b) improving the modelling process to build risk prediction models tailored to ethnicity groups and therefore more reliable in practice.

We propose to develop a calculator to predict cardiovascular disease (CVD) in COVID-19 patients as an example to demonstrate our approach, as here ethnic biases are known to exist. As inequity in data and models affects all disease areas, our approach is sustainable and can be applied to other clinical areas in the NHS. The calculator can be used by public to guide lifestyle choices, and by doctors to provide better care.

The publicly available algorithm for improving information on ethnicity can be used by researchers nationwide doing health research involving ethnicity.

This work will be based on anonymised health information that represents almost everyone currently living in England and Wales. By extending to Northern Ireland and Scotland in future, we hope that this work will help to make health equal and fair for everyone in the UK.

## AUTHORS

Sara Khalid, Marta Pineda Moncusí, Antonella Delmestri…

## TITLE

Improving methods to minimise bias in ethnicity data for more representative and generalizable models, using CVD in COVID-19 as an example.

## SHORT TITLE

Minimising bias in ethnicity data

## BACKGROUND

The importance of ethnicity in understanding and addressing inequalities in healthcare access, patient experience, and patient outcomes is well-recognised [1-5], but it has been highlighted by the COVID-19 pandemic, where people from ethnically diverse backgrounds were disproportionately affected [6].

Much health research and clinical practice, including predicting risk of developing health conditions, relies on data collected by healthcare professionals in primary (e.g., GP practices) and secondary (e.g., hospitals) care settings. Patient self-reported variables (e.g., smoking, ethnicity) are not always collected, as individuals can decline to share them, and healthcare professionals may not ask for or record them. At least one-third of patients are missing ethnicity records [7]. When recorded, ethnicity is often inaccurately coded [7-9]. Literature focuses on a subset of ethnicity Read codes [10, 11], often collapsed into five to nine categories [4, 12, 13]. This oversimplification, differences in census classifications over time [14], conflicts in individuals’ recorded ethnicity, and speculative recording contribute to inaccuracies.

Many prediction models do not consider ethnicity [15] or estimate/impute missing ethnicity using invalid assumptions. Although ethnicity’s missingness can vary by ethnicity, many well-known clinical risk prediction tools [16-19] assume missingness at random. The result is inaccurate, biased ethnicity estimates. Prediction models are generally not trained, tested, and externally validated on data representative of all ethnic groups [15, 16, 18, 19]. They may not capture differences between minority/ethnic groups in the magnitude of predictor-outcome associations or assess predictor-ethnicity interactions. Consequently, these tools can cause over- or under-estimated outcome risks [20] and thus inequality in healthcare provision:   
if there is bias in the data or bias in the model, the doctor can potentially make wrong decisions and patients can get the wrong care or no care, meaning some groups of patients might inappropriately be prioritised over other for booster vaccines, hospital beds, or life-saving treatments, which in turn can affect patients’ lives and public trust, and cost to the NHS.

All of the >100 published COVID-19 prediction models are at risk of bias and unsuitable for triaging patients in clinical practice [20]. Against the backdrop of health inequalities exposed by the COVID-19 pandemic [21-25], these issues motivate the need to fill gaps in ethnicity data, identify ethnicity-specific predictors, and build accurate clinical risk prediction models that are suitable for minority and ethnic groups.

This proposal will develop an approach for modelling observational data that addresses ethnicity coding and builds representative models suitable for ethnic groups by predicting cardiovascular disease (CVD) in COVID-19 patients as an example to demonstrate our approach, as here ethnic biases are known to exist. As inequity in data and models affects all disease areas, our approach is sustainable and can be applied to other clinical areas in the NHS. The calculator can be used by public to guide lifestyle choices, and by doctors to provide better care.

## RESEARCH QUESTION

To improve the correctness, completeness, and granularity of ethnicity in routinely collected data such that more representative and generalizable models can be developed, using CVD in COVID-19 as an example.

## RESEARCH PLAN

## Objective 2. To build ethnicity-specific CVD-in-COVID-19 risk prediction models using the improved ethnicity data

*Study population*:

All individuals meeting the following inclusion criteria will be recruited:

* aged ≥18 years
* >1 year of records available
* confirmed diagnosis of COVID-19

Individuals will be excluded when:

* aged <18 years
* <1 year of data before index date
* No confirmed diagnosis of COVID-19

*Outcomes*: CVD (venous and arterial thromboembolic events), death, hospital admission

*Candidate risk factors:* A pre-specified list of risk factors from published studies will be included for the analysis of key predictors of cardiovascular disease. The list will be limited to factors that are available in primary and secondary care records, as these will be the basis for future risk assessments based on the proposed prediction tool. This list will include:

1.- Generic/risk factors included in existing cardio-vascular prediction tools and identified in existing literature:

a) Age (at covid diagnosis date)

b) Sex

c) Body mass index

d) Smoking and Alcohol drinking

e) History of cardiovascular disease (ischaemic heart disease, stroke)

f) History of Type 2 diabetes (or use of anti-diabetic therapies)

g) History of hypertension (or use of anti-hypertensive drug/s)

h) Previous thrombo-embolic events (deep vein thrombosis, pulmonary embolism) or use of anti-thrombotic or anti-coagulant treatment/s

i) Chronic kidney disease

j) Estimated renal function

k) Hyperlipidaemia (recorded diagnosis), use of statin/s, or serum lipid levels

l) Charlson co-morbidity index

m) Concomitant medications (number of different ATCs received in the year before inclusion)

n) Socio-economic deprivation (patient-level index of multiple deprivation quintile).

o) Systolic blood pressure (if available)

p) Ethnicity

2. – Other potential risk factors

a) Prior fractures

b) Recorded diagnosis of osteoporosis (marker of low bone mass)

*Follow-up:* Study participants will be followed from index date (date of COVID-19 diagnosis) to the earliest of the following: death, transfer-out, or event of interest.

*Methods:*

2.1: Assessing existing tools on ethnicity-specific cohorts.

We will evaluate the performance of existing models e.g. QRisk3 [18] and CVD-COVID-UK project CCU004 (as discussed with project lead Dr Wood) on ethnicity-specific cohorts, both for the 5 most commonly used ethnicity groups (White, Black, Asian, Mixed, Other), and the more granular ethnicity sub-groups derived from Objective1.

Cardiovascular risk will be evaluated at 90 days, 180 days, one year, and two years according to follow-up availability (COVID –19 is a new disease that has been recorded since early 2020).

Models will be evaluated in terms of calibration and discrimination by calculating the area under the curve (AUC) and by plotting the observed versus predicted probabilities, respectively.

2.2: Develop and externally validate ethnicity-specific models whilst minimising algorithmic bias.

Cardiovascular risk will be evaluated at 90 days, 180 days, one year, and two years according to follow-up availability.

All the available candidate risk factors will be fitted into a lasso regression to select the key predictors. Then, these key predictors will be combined into a final logistic regression equation to build a prediction model.

A prediction algorithm will be fitted to each ethnic group identified in Objective1, and, but not limited to, by the following groups\*:

- un/fully vaccinated

- sex

- age groups (decyls)

- prior/non-prior cardiovascular events (CVE)

- comorbidities

- covid variant/strain using calendar time as proxy 🡪 temporal confounding as limitation?

\* Subject to sample size

Using recent sample-size guidance [29, 30], sample sizes were determined to minimise overfitting. We will also develop a global model trained on the full dataset and recalibrated for the smallest ethnic sub-groups.

Measures for reducing algorithmic bias will include evaluating known and hidden interactions using both statistical testing and machine learning. Multiple feature selection methods will be used complementarily [31]. To minimise overfitting bias, we will test several approaches for splitting data into training and testing sets including random split, split by GP practice, split by time period, x:y split ratios (x:50-90, y:50-10), repeated cross-validation, and bootstrapping.

All prediction modelling will be based on best-practice guidelines [20, 32] and the UK AI code [33].

Models will be evaluated in terms of calibration and discrimination by calculating the area under the curve (AUC) and by plotting the observed versus predicted probabilities, respectively.

2.3: PPI case studies.

### PPI members will be invited to volunteer for case studies in two half-day meetings of 6-8 participants each.

## DATA SOURCES

For both Objectives 1 and 2, the same approach will be used:

* NHS Digital (55 million records) will be used for model development (and internal validation)
* SAIL (3.2 million) will be used for external validation.

## REQUESTED DATASETS

Note 1: we are aware that not all of the datasets in black font are available currently  
Note 2: please comment on whether any of the datasets in grey font might be relevant to this study

### NHS Digital TRE for England

* Primary care data
  + GPES Data for Pandemic Planning and Research (GDPPR)
* Secondary care data:
  + Hospital episode statistics Admitted Patient Care (HES APC)
  + Adult Critical Care
  + Outpatients
  + Accident & Emergency
  + SUS: Secondary Uses Service
  + SUS/Uncurated Low Latency Hospital Data (Admitted Patient Care, Outpatients, Critical Care)
  + Emergency Care Data Set (ECDS)
* COVID testing:
  + COVID-19 SGSS: Second Generation Surveillance System
  + Pillar 2 Antigen
  + Pillar 3 Antibody
  + Variant strain data (COG-UK)
  + Vaccination Status
  + Vaccination Adverse Reactions
* Death registers:
  + Office for National Statistics (ONS) death registration records
  + NHSD mortality data review
* ITU
  + ICNARC: Intensive Care National Audit and Research Centre
* ITU/HDU admissions
  + COVID-19 SARI-Watch (formerly CHESS: COVID-19 Hospitalisation in England Surveillance System)
* Community dispensing data
  + Medicines Dispensed in Primary Care (NHS BSA)
  + Secondary Care Prescribed Medicines (EPMA)
* NICOR CVD audits
  + NICOR – MINAP: Myocardial Ischaemia National Audit Project
  + NICOR – PCI: Percutaneous Coronary Interventions
  + NICOR – NHFA: National Heart Failure Audit
  + NICOR – NACSA: National Adult Cardiac Surgery Audit
  + NICOR – NACRM: National Audit of Cardiac Rhythm Management
  + NICOR – NCHDA: National Congenital Heart Disease Audit
  + NICOR – TAVI: Transcatheter Aortic Valve Implantation
* Stroke audit
  + SSNAP: Sentinel Stroke National Audit Programme
* National Vascular Registry
  + National Vascular Registry Audit
* Others:
  + Diagnostic Imaging Dataset
  + Improving Access to Psychological Therapies (IAPT)
  + Maternity Services Data Set
  + Mental Health Services Data Set
  + Mental Health of Children and Young People
  + Patient Reported Outcome Measures (PROMs)

### Secure Anonymised Information Linkage Databank (SAIL) for Wales

* Primary care data
  + Welsh Longitudinal General Practice (GPCD) - daily COVID codes only.
  + Welsh Longitudinal General Practice (WLGP)
* Secondary care data:
  + Critical Care Dataset (CCDS)
  + Emergency Department Dataset Daily (EDDD)
  + Emergency Department Dataset (EDDS)
  + Outpatient Dataset for Wales (OPDW)
  + Outpatient Referral Dataset (OPRD)
  + Patient Episode Dataset for Wales (PEDW)
* COVID testing:
  + PATD: COVID-19 Test Results (Laboratory Information Management System  
     [Pillar 1&2 NHS/Lighthouse Labs Results & Pillar 3 Antibody Results])
  + CTTP: COVID-19 Test, Trace and Protect
  + CVSP: COVID-19 Shielded People List
  + CVSD: COVID-19 Sequence Data
  + CVVD: Covid Vaccination Dataset
* Death registers:
  + ADDD: Annual District Death Daily (ONS Deaths)
  + ADDE: Annual District Death Extract (ONS Deaths)
  + CDDS: COVID-19 Consolidated Deaths
* ITU:
  + ICCD: ICNARC – Intensive Care National Audit & Research Centre   
     (COVID-19 only admissions)
  + ICNC: ICNARC – Intensive Care National Audit & Research Centre   
     (All admissions)
* Prescribing/ Dispensing
  + Wales Dispensing Dataset (WDDS)
* NICOR CVD audits
  + NICOR Audits and Registers (NICO)
* Stroke audit
  + HQIP Stroke Audit (HQIP)
* National Vascular Registry
  + National Vascular Registry (NVR)
* Others:
  + ADBE: Annual District Birth Extract
  + MIDS: Maternity Indicators Dataset
  + CARE: Care Homes Index
  + CENW: Office of National Statistics Census (2011)
  + RTTD: Referral to Treatment Times
  + SDEC: SAIL Dementia e-Cohort
  + WASD: Welsh Ambulance Services NHS Trust
  + WDSD: Welsh Demographic Service Dataset
  + WRRS: Welsh Results Reporting Service

## REFERENCES

1. Evans, N., et al., *Appraisal of literature reviews on end-of-life care for minority ethnic groups in the UK and a critical comparison with policy recommendations from the UK end-of-life care strategy.* BMC Health Serv Res, 2011. **11**: p. 141.

2. Wilkinson, E. and G. Randhawa, *An examination of concordance and cultural competency in the diabetes care pathway: South Asians living in the United Kingdom.* Indian J Nephrol, 2012. **22**(6): p. 424-30.

3. Bhattacharyya, S. and S.M. Benbow, *Mental health services for black and minority ethnic elders in the United Kingdom: a systematic review of innovative practice with service provision and policy implications.* Int Psychogeriatr, 2013. **25**(3): p. 359-73.

4. Mathur, R., et al., *Completeness and usability of ethnicity data in UK-based primary care and hospital databases.* J Public Health (Oxf), 2014. **36**(4): p. 684-92.

5. Pan, D., et al., *The impact of ethnicity on clinical outcomes in COVID-19: A systematic review.* EClinicalMedicine, 2020. **23**: p. 100404.

6. Mathur, R., et al., *Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: an observational cohort study using the OpenSAFELY platform.* Lancet, 2021. **397**(10286): p. 1711-1724.

7. Wolf, A., et al., *Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum.* Int J Epidemiol, 2019. **48**(6): p. 1740-1740g.

8. Saunders, C.L., et al., *Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: evidence from the English Cancer Patient Experience survey.* BMJ Open, 2013. **3**(6).

9. Neckles, L. and NHS-England, *Ethnicity coding in health records.* 2020.

10. *Final recommended questions for the 2011 Census in England and Wales*. 2009; Available from: ttps://[www.ethnicity-factsfigures.service.gov.uk/style-guide/ethnic-groups](http://www.ethnicity-factsfigures.service.gov.uk/style-guide/ethnic-groups).

11. Benson, T., *The history of the Read Codes: the inaugural James Read Memorial Lecture 2011.* Inform Prim Care, 2011. **19**(3): p. 173-82.

12. Kumarapeli, P., et al., *Ethnicity recording in general practice computer systems.* J Public Health (Oxf), 2006. **28**(3): p. 283-7.

13. Tippu, Z., et al., *Ethnicity Recording in Primary Care Computerised Medical Record Systems: An Ontological Approach.* J Innov Health Inform, 2017. **23**(4): p. 920.

14. GOV.UK. *List of ethnic groups:* [*www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups*](http://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups). Available from: <https://www.ethnicityfacts-figures.service.gov.uk/style-guide/ethnic-groups>.

15. Brindle, P.M., et al., *The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study.* Br J Gen Pract, 2005. **55**(520): p. 838-45.

16. Collins, G.S., S. Mallett, and D.G. Altman, *Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective independent and external validation of QFractureScores.* BMJ, 2011. **342**: p. d3651.

17. Nashef, S.A., et al., *EuroSCORE II.* Eur J Cardiothorac Surg, 2012. **41**(4): p. 734-44; discussion 744-5.

18. Hippisley-Cox, J., C. Coupland, and P. Brindle, *Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study.* BMJ, 2017. **357**: p. j2099.

19. Clift, A.K., 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**: p. m3731.

20. Collins, G.S., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement.* Ann Intern Med, 2015. **162**(1): p. 55-63.

21. Khunti, K., et al., *Is ethnicity linked to incidence or outcomes of covid-19?* BMJ, 2020. **369**: p. m1548.

22. Sze, S., et al., *Ethnicity and clinical outcomes in COVID-19: A systematic review and meta-analysis.* EClinicalMedicine, 2020. **29**: p. 100630.

23. Apea, V.J., et al., *Ethnicity and outcomes in patients hospitalised with COVID-19 infection in East London: an observational cohort study.* BMJ Open, 2021. **11**(1): p. e042140.

24. Boseley, S., *QCovid: how improved algorithm can identify more higher-risk adults*, in *The Guardian News*. 16 February 2021.

25. Haynes, L., *QCovid tool to be reviewed by NHS Digital after healthy women wrongly advised to shield*, in *GPonline*. 22 February 2021.

26. Delmestri, A. and D. Prieto-Alhambra, *CPRD GOLD and linked ONS mortality records: Reconciling guidelines.* Int J Med Inform, 2020. **136**: p. 104038.

27. Pham, T.M., et al., *Ethnic Differences in the Prevalence of Type 2 Diabetes Diagnoses in the UK: Cross-Sectional Analysis of the Health Improvement Network Primary Care Database.* Clin Epidemiol, 2019. **11**: p. 1081-1088.

28. Petersen, I., et al., *Health indicator recording in UK primary care electronic health records: key implications for handling missing data.* Clin Epidemiol, 2019. **11**: p. 157-167.

29. Riley, R.D., et al., *Calculating the sample size required for developing a clinical prediction model.* BMJ, 2020. **368**: p. m441.

30. Riley, R.D., et al., *Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes.* Stat Med, 2019. **38**(7): p. 1276-1296.

31. Tarassenko, L., *Guide to Neural Computing Applications*. 1998: Elsevier.

32. Sounderajah, V., et al., *Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group.* Nat Med, 2020. **26**(6): p. 807-808.

33. DHSC. *A guide to good practice for digital and data-driven health technologies*. Available from: <https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology>.