## 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. |
| V0.3 | 01/01/2023 | To simplify the study understanding, this protocol is focused on women disparities across covid outcomes while predictive models for both sex are now reported on CCU037\_05 |

## 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.

The main objective of the overall 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.

In order to develop a calculator to predict mortality and cardiovascular disease (CVD) in COVID-19, we first need to observe health disparities across the England population. This subproject aims to observe mortality and CVD disparities among women infected by SARS-CoV-2.

This work will be based on anonymised health information that represents almost everyone currently living in England. 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…

## STUDY TITLE

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

## SHORT STUDY TITLE

Minimising bias in ethnicity data

**CCU037\_02 SUBPROJECT TITLE**

Women’s health and ethnic disparities: a population-wide analysis of 3.6 million digital health records for mortality and cardiovascular risk in women diagnosed with COVID-19

## 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.

Underlying disparities were particularly highlighted by the COVID-19 pandemic where people from ethnically diverse backgrounds were disproportionately affected [15-19]. In addition to that, gender gaps in health are well-known where many women receive poorer care than men. This can often be intertwined with ethnic disparity affecting women of colour. In the US, a cross-sectional survey of 3,200 women show that Black, East or Southeast Asian, and Hispanic women had a larger risk to increase health-related socioeconomic vulnerability (food, housing, utilities and transportation difficulties, and interpersonal violence) early in the COVID-19 pandemic than white women.[20] On the other hand, a study from England show larger disparities in health-related quality of life among women from different ethnic background than men [21].

In order to develop a calculator to predict mortality and cardiovascular disease (CVD) in COVID-19, we first need to observe health disparities across the England population.   
This proposal will observe ethnic disparities in women diagnosed with COVID-19 disease in the England population by exploring the risk of mortality and cardiovascular disease during different phases of the pandemic.

## RESEARCH QUESTION

To estimate age-standardised incidence rates for mortality and CVD across different ethnic groups observable in women diagnosed with COVID-19 disease in the England population

## RESEARCH PLAN

## Objective 2. To estimate age-standardised incidence rates for mortality and CVD in women diagnosed with COVID-19

*Study population*:

All individuals meeting the following inclusion criteria will be recruited:

* aged between ≥30 years and ≤ 100 years
* >1 year of records available
* confirmed diagnosis of COVID-19 (i.e., record of a diagnosis or positive PCR test in SGSS, or hospital admission due to COVID)

Individuals will be excluded when:

* aged <30 years or <100 years
* <1 year of data before index date
* No confirmed diagnosis of COVID-19

*Outcomes*: CVD and death

*Baseline characteristics:* A pre-specified list of baseline characteristics (i.e.., extracted at the time of COVID-19 diagnosis) to characterize the study population across the different ethnic groups. 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.- Demographic data:

1. Age
2. Ethnicity
3. Index of multiple deprivation (IMD)
4. Smoking
5. Ethnicity
6. LSOA

2. – Clinical characteristics (previous clinical history ever unless stated):

1. pregnancy status (at time of COVID-19 diagnosis)
2. atrial fibrillation
3. alcohol problems
4. bipolar disorder
5. cancer
6. chronic kidney disease
7. chronic obstructive pulmonary disease
8. dementia
9. depression
10. diabetes
11. hypertension
12. obesity
13. osteoporosis
14. rheumatoid arthritis
15. schizophrenia
16. history of CVE
17. diagnosis CVE during the previous year

3. – Use of medication in the year prior to COVID-19 diagnosis:

1. antidiabetic
2. anticoagulant
3. antihypertensive
4. antiplatelet
5. antipsychotic medication
6. statins

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

*Methods:*

2.1: Assessing prevalence of ethnic diversity and baseline characteristics.

We will calculate the prevalence of demographic and clinical characteristics of the selected individuals at the time of COVID-19 diagnosis for each of the observed ethnic groups.

The diversity of the selected individuals will be explored by:

1. Calculating the overall prevalence of the different ethnic groups and compare it to all women in GP dataset without any exclusion criteria (except invalid age and sex).
2. Calculate the prevalence of the selected individuals by the period of enrolment to the study cohort (i.e., date of COVID-19 diagnosis). Periods will be defined as every 6 months since the 2020 until end of data availability.

2.2: Estimating Age-standardised incidence rates

Mortality risk will be evaluated at 28 and 90 days.

Cardiovascular risk will be evaluated at 30 days and one year.

To estimate the age-standardised incidence rates (IR) for each of the ethnicities, we will follow the ONS methodology [22]:

1. We will calculate the age-specific IR quintile bands from 30 to 100 years (i.e., 30 to 34, 35 to 39, etc.).
2. The age-specific IR will be combined into the age-standardised IR using the 2013 European Standard Population (2013 ESP) weights from 30 to 90+ age groups.

This process will be repeated stratifying by period of enrolment (i.e., every 6 months since the 2020 until end of data availability).

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

* NHS Digital (55 million records) will be used for estimating the incidence of the study outcomes among women from diverse ethnic background

## REQUESTED DATASETS

### 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

## 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. Baloch, S., Baloch, M. A., Zheng, T. & Pei, X. The Coronavirus Disease 2019 (COVID-19) Pandemic. Tohoku J Exp Med 250, 271-278 (2020). https://doi.org:10.1620/tjem.250.271

16. Fang, X. et al. Epidemiological, comorbidity factors with severity and prognosis of COVID-19: a systematic review and meta-analysis. Aging (Albany NY) 12, 12493-12503 (2020). https://doi.org:10.18632/aging.103579

17. Ramirez-Soto, M. C., Ortega-Caceres, G. & Arroyo-Hernandez, H. Sex differences in COVID-19 fatality rate and risk of death: An analysis in 73 countries, 2020-2021. Infez Med 29, 402-407 (2021). https://doi.org:10.53854/liim-2903-11

18. Doerre, A. & Doblhammer, G. The influence of gender on COVID-19 infections and mortality in Germany: Insights from age- and gender-specific modeling of contact rates, infections, and deaths in the early phase of the pandemic. PLoS One 17, e0268119 (2022). https://doi.org:10.1371/journal.pone.0268119

19. Fabiao, J. et al. Why do men have worse COVID-19-related outcomes? A systematic review and meta-analysis with sex adjusted for age. Braz J Med Biol Res 55, e11711 (2022). <https://doi.org:10.1590/1414-431X2021e11711>

20 Vu, M. et al. Racial and Ethnic Disparities in Health-Related Socioeconomic Risks During the Early COVID-19 Pandemic: A National Survey of U.S. Women. J Womens Health (Larchmt) 30, 1375-1385 (2021). https://doi.org:10.1089/jwh.2021.0230

21 Watkinson, R. E., Sutton, M. & Turner, A. J. Ethnic inequalities in health-related quality of life among older adults in England: secondary analysis of a national cross-sectional survey. Lancet Public Health 6, e145-e154 (2021). <https://doi.org:10.1016/S2468-2667(20)30287-5>

22 Rachel Woods, A. C. *User guide to mortality statistics*, <<https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017#introduction>> (2022).