The association between ethnic background, socio-economic deprivation and COVID-19 testing, status and health outcomes: a multi-state cohort analysis

---

Tomás Urdiales Bartolomé

28th June 2021

Universitat Politècnica de Catalunya

Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona

Bachelor's Degree in Engineering Physics, Final Thesis

The association between ethnic background, socio-economic deprivation and COVID-19 testing, status and health outcomes: a multi-state cohort analysis

Tomás Urdiales Bartolomé

Co-director Dr. Clara Prats
Computational Biology and Complex Systems
Universitat Politècnica de Catalunya

Co-director Dr. Daniel Prieto-Alhambra
Centre for Statistics in Medicine, NDORMS
University of Oxford

Additional supervisors Albert Prats-Uribe (University of Oxford)
Martí Català (Universitat Politècnica de Catalunya)

28th June 2021

Abstract

Nearly a year and a half have passed since severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) started the global coronavirus pandemic of 2020/2021. Consequently, it is now an optimal time to analyse the disease's health outcomes and their association with socio-economic and biological factors. Here, we use data collected through the UK Biobank program to curate a comprehensive database of 406,408 randomly sampled English individuals that includes information on COVID-19 health outcomes such as testing, infection, hospitalisation and mortality; as well as key population characteristics. A multi-state model has been developed to track the disease's progression and Cox Proportional Hazards methodology has been applied to obtain confounder adjusted hazard ratios. Testing prevalence has been studied in parallel using Poisson regression count analysis.

In the current state of affairs, certain risk factors have already been well established as adverse predictors of coronavirus-related health outcomes. However, its association with ethnicity and socio-economic status has not been fully understood yet. Previous research conducted early on in the pandemic has provided evidence that ethnic minorities and socio-economically deprived groups have been disproportionately affected by the pandemic. Nevertheless, empirical studies on the topic remain limited, hence the need for further research.

Our study results reveal severely increased risk of infection for Asian, Black, Mixed and *'other'* ethnicity individuals, as well as a much higher test count on average. In terms of hospitalisation, increased risk is found across all ethnic groups, and Blacks are the worst-off overall with up to twice the hazard for Whites, followed by Asians. Mortality results indicate increased risk for Asians, most notably, and Blacks.
Regression coefficients for socio-economic strata display a steadily proportional relationship between economic deprivation and hazard. Increased risk is found across all health outcomes, with same-day hospitalisation appearing up to three times as likely for the most deprived section of the population when compared with the least. Furthermore, there is a severely increased risk of infection for the poorer quintiles, as well as many more tests taken on average. Mortality also increases with socio-economic deprivation, and the risk is estimated to be up to 10% higher for the bottom quintile. These findings demonstrate the continued need to protect those at high risk of poor outcomes due to coronavirus disease.

Acknowledgement

This thesis is the result of five months of research carried out as part of a remote collaboration between researchers at the Polytechnic University of Catalonia and the University of Oxford. It is the final work of my Engineering Physics bachelor's degree at UPC.

I would like to sincerely thank Dr. Clara Prats, Dr. Daniel Prieto-Alhambra, Albert Prats-Uribe and Martí Català for their continued guidance and support all throughout the development of this thesis. Although the circumstances have not permitted face-to-face interaction, they have made it an amazing academic and personal experience nonetheless. This project would not have occurred had they not trusted in me, despite my complete lack of experience in the field, and been there with me all along the way, and I am very grateful for that.

Contents

1 Introduction

1.1 Background & rationale

1.1.1 The COVID-19 pandemic

Nearly a year and a half have passed since severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) started the global coronavirus pandemic of 2020/2021 [1]. It was first identified in Wuhan in December of 2019 and by the 23rd January 2020, the first cases within the United Kingdom were already confirmed [2]. At the time of writing of this thesis (28th June 2021), there have been more than 180 million reported COVID-19 cases worldwide, and more than 3.91 million confirmed deaths attributed to the disease [@3], making it one of the deadliest pandemics in history. In the United Kingdom alone, about 4,717,811 people have tested positive, and 128,089 have died [@3].
The pandemic has also caused considerable social and economic disruption. Furthermore, it has highlighted conflicts of racial and geographic discrimination and health equity [4], which partly motivate this project's objectives.

While countries and their medical systems now race to implement population-wide vaccination programs, data on COVID-19's impact keep being collected and put to use for scientific analysis. As such, it is now an optimal time to analyse the disease's health outcomes and their association with socio-economic and biological factors. A better understanding of this relationship is essential for effective health service planning and, possibly, for future risk prevention efforts.

1.1.2 Ethnic inequality in the context of the pandemic

In the current state of affairs, certain risk factors have already been well established as predictors of adverse coronavirus-related health outcomes. Some of these are age, morbid obesity and male sex [5]. However, its association with ethnicity and socio-economic deprivation has not been fully understood yet. There is evidence that ethnic minorities have been disproportionately affected in past pandemics [6].

Research has now shown that the same phenomenon appears to be occurring for the COVID-19 pandemic [4][7][8]. Nevertheless, empirical studies on the topic remain limited, hence the need for further research. Part of the reason behind it is that data recollection on ethnic background and socio-economic status is sensitive and sparse, and many programs do not collect it when carrying out COVID-19 cohort research. This issue was faced early on in the project when in the preliminary results phase it was difficult to obtain said information to guide the study. Fortunately, the UK Biobank program put special effort into collecting population characteristics at the time of registration, and ethnicity and socio-economic deprivation status are well documented for most of its participants.

The treatment of ethnicity in this study is completely conditioned by UK Biobank's standard of data recollection, as explained in detail in the *Variables* section within Methods 2.5. The Journal of the American Medical Association's publication on *The Reporting of Race and Ethnicity in Medical and Science Journals* [9] informed the standards of writing when discussing topics of ethnicity throughout the thesis.

1.1.3 State of the art in UK Biobank research

UK Biobank is a large, non-commercial, long-term biobank project carried out in the United Kingdom since 2006 [@10]. It was created with the goal of investigating the role of genetic predisposition and environmental exposure towards the development of disease [11]. It stores medical data on about half a million people, aged between 40 and 69 at the time of registration, and blood, urine and saliva samples as well as information on their lifestyle linked to health outcomes. All volunteers remain in follow-up for at least 30 years.

So far, some studies have conducted research into ethnic and socio-economic differences in COVID-19 health outcomes using UK Biobank's data. Researchers at the Social & Public Health Sciences Unit of the University of Glasgow university found an increased risk of infection for south Asian and Black individuals [12]. Their study was conducted early on in the pandemic, which warrants further research into the topic given the larger amount of data available today.

Previous work from the Centre for Statistics in Medicine of the University of Oxford also obtained results for infection risk along the same lines [13]. They found higher relative risk rates for Black, Chinese, Asian and '*other*' ethnicities when compared with White participants.

The particular objectives and statistical methodology employed in this project have not been found in other UK Biobank-based research into ethnic and socio-economic differences in COVID-19 health outcomes.

1.1.4 The collider bias problem

Observational studies that attempt to identify risk factors for COVID-19 infection and disease outcomes may be based on non-representative samples that induce collider bias [14].

In statistics, a collider is defined as a variable that is causally influenced by two other variables. In this scenario, both the risk factor and the outcome affect the probability of being sampled. As a result, the association between these two variables may be distorted, a phenomenon that is usually referred to as collider bias.

COVID-19 studies tend to collect their data from patient hospitalisations or people who test for active infection. The resulting sample population may not be representative of the reality of the situation, since the probability of being sampled depends greatly on a patient's COVID-19 risk profile and health outcomes. In other words, observational data could be biased because only individuals who already suffer from coronavirus disease or are at high risk of it are observed. This implies that sampling could be non-random. When analysing UK Biobank data, evidence was found that participants tested for COVID-19 were highly selected for a range of traits (genetic, demographic, behavioural, etc) [14]. Throughout the pandemic this issue has become prevalent, particularly at the early stages when there were very few tests and they were mostly reserved for ill individuals [15]. Some programs have put special effort into carrying out truly random sampling to study the prevalence of infection among the general population, like Imperial College's REACT experiments (Real-time Assessment of Community Transmission Findings) [16].

Analysis over a wide cohort of patients like UK Biobank's, and not just those who test positive for COVID-19 or are hospitalised, circumvents the majority of issues that collider bias may induce otherwise. Herein lies part of the strength behind this study's design.

1.2 Terminology

There is a variety of terms frequently used throughout this document that are specific to the fields of epidemiology and, more specifically, to cohort studies of the kind carried out in this thesis:

Stratification/stratum. A stratum refers to a subset of the population that is being sampled. The process of stratification may be done on a geographical basis or by reference to certain population characteristics. Thus, when referring to ethnicity or socio-economic deprivation strata, one is breaking up the cohort into different categories according to these variables, for the sake of analysis.

Confounder. In statistics, a confounding variable is one that influences both the dependent and independent variable [17]. In other words, confounders are all the factors influencing both exposure and outcome, causing a spurious association between the two. In the context of this study an example would be how socio-economic deprivation confounds the relationship between ethnic background (exposure) and COVID-19 hospitalisation (outcome). In this scenario, there may be substantial differences in the socio-economic deprivation distribution depending on ethnicity, whilst socio-economic status simultaneously affects health outcomes. Confounders ought to be accounted for in statistical analysis, hence the phrase *adjusting for confounders* or *minimising confounders*.

Follow-up is a key concept in cohort studies survival analysis that refers to the time a given patient is monitored (observation window). When a study commences, all individuals enter follow-up and remain in that status until they are censored, or

Censoring refers to a missing data problem or analytical decision, in which the time to event is not observed. In survival analysis, it could be due to termination of study before all participants have undergone the events of interest, or because a subject leaves the study prior to the event occurring.
Left-censoring refers to all events which are not tracked because they occur prior to the beginning of follow-up, the **left-censoring date**. Similarly, the **right-censoring date** marks the date at which follow-up ends and events are no longer registered.

n refers to group or population size. It is commonly used in the context of large or small *n* to discuss whether there are sufficient individuals within a given strata for statistical analysis to be viable.

Furthermore, some abbreviations are often used to avoid redundancy:

UKB is used to shorten UK Biobank, and to refer to data that comes from any of their associated sources.

HES stands for Hospital Episode Statistics, a program of the United Kingdom's National Health Service that stores data on hospital admissions.

COVID-19 is used to refer to SARS-CoV-2 or to the coronavirus disease it causes, always in the context of health outcomes (i.e., COVID-19 hospitalisation).

1.3 Introduction to multi-state modelling

In standard survival data studies, the time until the occurrence of a certain event of interest is measured. Competing risk models [18] are used when there exist multiple possible events, as in the generalised situation of Figure 1.1. This framework understands that progression towards any given event is dependent on how every other possible event's risk "competes" with that of interest. Thus, caution is needed when estimating the probability of any given event occurring in the presence of other possible competing risks.

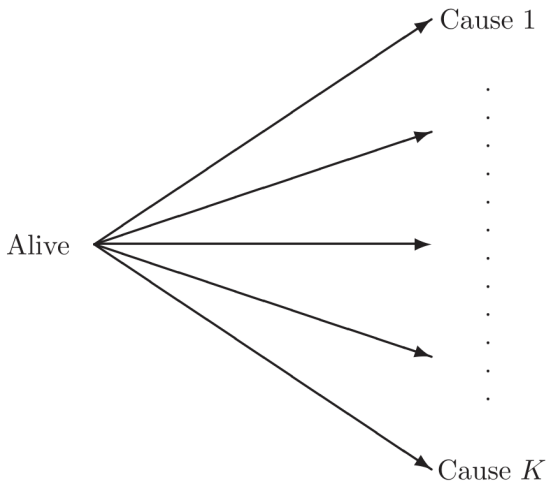


Fig. 1.1: General structure of a competing risks model. A patient in follow-up is simultaneously susceptible to a number of risk events.

To put it in perspective, an example of such a scenario could be epidemiological research on the time from HIV infection to the development of acquired immunodeficiency syndrome (AIDS) or the time to leukaemia relapse after a bone marrow transplant operation. If these were the events of interest, an event that may prevent

their occurrence being observed would be the death of a patient before diagnosis. Therefore, death is treated as one of the competing risks.

There are usually also intermediate events that may significantly affect the risk of the event of interest occurring. One would usually be interested in what happens after one such non-fatal event because it may provide detailed information on the disease or recovery process. In the case of HIV infection, the event of an individual developing AIDS significantly conditions prognosis, and therefore ought to be treated as an intermediate event with an associated competing risk (Fig. 1.2).

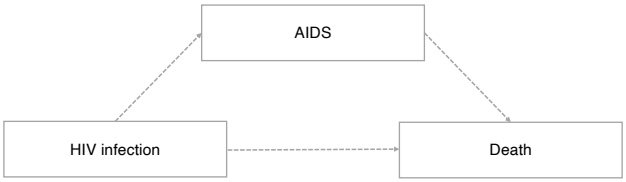


Fig. 1.2: Example of a simple illness-death model for HIV/AIDS.

Note how in this framework such non-fatal events can be understood as transitions from one state to another. As such, there is a time origin from which the patient enters follow-up, in the above examples these would be bone marrow transplantation and HIV infection respectively, and an endpoint, or final absorbing transition, where the patient is censured, i.e., exits the study. Whereas competing risk models usually deal only with several mutually exclusive absorbing states, multi-state models generalize the concept to incorporate and describe possible intermediate events [19][20].

The Markov property. In practice, most multi-state models are assumed to be Markov models, including the one implemented in this study. In Markov systems, past history conditions the occurrence of future events exclusively through the present [21]. In other words, the next state and the time of transition towards it depend only on the present state, and not on any past history of events.

While there are multiple important approaches to these statistical methods (fully parametric models, additive hazards, etc), this project has employed the Cox model framework, a non-parametric hazards multi-state model.

1.4 Objectives

The overarching objective is to study the association between ethnicity and socio-economic deprivation with respect to the risk of SARS-CoV-2 infection, hospitalisation, and mortality in UK Biobank's English cohort. The project seeks to quantify how an individual's socio-economic status and ethnic background affects their probabilities of experiencing these COVID-19 health outcomes.

An indispensable parallel goal of the project is to study COVID-19 testing prevalence separately using different statistical methodology that allows for *'count'* analysis instead of just the probability of being tested as a binary property. It is an essential requirement in order to properly analyse the differences in the risk of infection.

The PICO framework [22] is commonly used in evidence-based Medicine to pose and answer clinical questions. It is an acronym for patient, intervention, comparator and outcome, the four stages with which to establish a research goal. According to this process, the population or patients in the study are all UKB users who reside in England and satisfy certain, straightforward, eligibility criteria. The intervention or investigated conditions are ethnic background and socio-economic deprivation, treated both separately and as mutual confounders. For ethnicity, the comparator is the set of White patients in the cohort; and for socio-economic deprivation, it is the uppermost quintile of the population (i.e., the least deprived). The COVID-19 outcomes of interest in this project are testing, infection, hospital admission and death.

According to the stated framework, the specific objectives of this thesis can be summed up as:

1. To curate an informative, structured and refined dataset from UK Biobank sources that fulfills all the requirements to carry out analysis and modelling.

2. To analyse population characteristics within the cohort and quantify COVID-19 health outcomes.

3. To develop and implement a multi-state model that allows for transition-specific assessment of hazard in the context of COVID-19 health outcomes while minimizing the bias introduced by the relevant confounders.

4. To determine relative hazard ratios for ethnic background and socio-economic deprivation within the cohort, using the set of White patients and the uppermost quintile of the deprivation index as reference. This is to be done for COVID-19 testing, infection, hospitalisation and death.

5. To study COVID-19 testing prevalence and its relationship with ethnicity and socio-economic deprivation using specially adapted statistical methodology.

1.5 Outline

The project's structure largely follows the STROBE guideline for reporting observational cohort studies [23]. It is a widely used schematic in epidemiology for publication in medical journals. The acronym stands for *Strengthening the Reporting of Observational Studies in Epidemiology*. The reason for adopting this particular guideline is partly educational, to gain experience with the scientific standards of publication in this field, but also practical, because once finalised, the work carried out in this project is aimed at publishing an article. There have been some sections added that would normally not be included because this document's goal is to also report how the thesis has developed over time and the work that has gone into it.

The project is structured in four parts. After this introductory section, **Methods**, presents all the relevant information on how the study has been designed and conducted, as well as the reasons behind it. All study variables, the way they are treated and quantified, their data sources and their utility within the model are clearly stated and explained. It provides a summary of the multi-state model and statistical methods that constitute the theoretical backbone behind the methodology applied, and all the virtual resources that have been required for their implementation.

The **Results** chapter presents all tables and figures of the results obtained from the data and the modelling carried out. It provides all the numbers behind the model's final database as well as analysis on population characteristics and COVID-19 health outcomes. This section's function is to present all observations and results obtained and to provide brief summary and interpretation of these.

Discussion & conclusions, summarises the key results with reference to the study's objectives. It provides analysis and interpretation of the most important findings and of how they fit into the larger narrative observed. It also discusses the study's limitations, addresses possible biases in the data or the methodology employed and examines the generalisability (external validity) of the results obtained. Some speculation of the causality behind the phenomena observed is provided.

2 Methods

2.1 Study design

This prospective cohort study was informed by UK Biobank studies carried out in the United Kingdom since 2006 [11], with particular attention naturally given to COVID-19 events that occurred during the global pandemic of 2020/2021 [24]. Participants of the program living in England were followed from the 1st February 2020 to the 20th March 2021. Combined with NHS's Hospital Episode Statistics [@25], a comprehensive database was built that gathered not only information on COVID-19 testing, hospitalisation and mortality, but also key variables such as age, sex, ethnicity, socio-economic status and previous comorbidities for the entire cohort.

A Cox-regression-based multi-state model served as the general framework of analysis, accepting (and later on testing) a proportional hazards assumption and studying each transition between possible COVID-19 events separately, as a Markov system. An exploration of the theoretical backbone behind the statistics employed has been an essential part of the project and is extensively discussed in the *Statistical Methods* section (2.6).

2.2 Data sources

2.2.1 UK Biobank

The primary source of data has been UK Biobank [@10]. It is a large, non-commercial, long-term biobank project carried out in the United Kingdom since 2006 created with the goal of investigating the role of genetic predisposition and environmental exposure towards the development of disease [11]. It stores medical data on about half a million people, aged between 40 and 69 at the time of registration, and blood, urine and saliva samples as well as information on their lifestyle linked to health outcomes. All volunteers remain in follow-up for at least 30 years. It was in 2010 that the recruitment target of 500000 was reached, much earlier than

the beginning of the COVID-19 pandemic. For this particular study, the program provided not only population characteristics (age, sex, ethnicity and socio-economic deprivation), but also data for COVID-19 testing and mortality.

Through UKB's program essentially three different data-sets have been obtained. The first is the baseline file with the population characteristics of all patients. The second corresponds to a new COVID-19 testing database linked to the program from Public Health England [26]. The third is a death registry for their participants that they also update often.

2.2.2 Hospital Episode Statistics

The United Kingdom's National Health Service stores data on all hospital admissions, outpatient appointments, and attendances as part of its Hospital Episodes Statistics program [@25]. All information on COVID-19 hospital admissions and relevant comorbidities was obtained from HES Admitted Patient Cara data (HES APC), through UKB's portal. Registered users of UKB can be traced within HES's records, so the two databases are connectable.

2.2.3 Additional sources

Additional data sources were used at the early stages of the project to contrast early findings within Biobank's data with national statistics. Imperial College's REACT program (Real-time Assessment of Community Transmission Findings) [16] provided great information on community transmission and COVID-19 prevalence in wider society, not just UKB's cohort. It was particularly useful because it also tracked its participants' population characteristics, so it was possible to contrast early findings from our data on ethnic background and COVID-19 prevalence with Imperial's estimates.

A study that included testing data stratified by ethnicity was carried out by the United Kingdom's government from 28th May 2020 to 26th August 2020 [@27], when evidence of an ethnic disparity in the consequences of the pandemic was already beginning to be studied. We were able to compare our early findings on COVID-19 testing differences with those the government measured for the entire population. The same source was again used to obtain macro numbers on the country's ethnic composition, in order to compare the proportions in our cohort with those of England.

2.3 Setting

The main location of interest in this study is England. Despite the fact that UK Biobank has data for members all across the United Kingdom, England has its own particular multiple economic deprivation index [28] different from that of Scotland, Wales, etc. For this reason, we have restricted the area of study to participants exclusively within England.

The relevant date chosen for the beginning of follow-up (and left-censoring in the model) is the 1st February 2020, when it can be considered that inhabitants of the United Kingdom were under considerable exposure to COVID-19. Because Biobank users are tracked since registration, which occurred long before this project's follow-up began, up until the present moment, they were still in follow-up at the time the study was carried out.
The study's end of follow up (or right-censoring) has been chosen to be the 20th March 2021, or earlier in the event of death. The criterion behind this decision has been to settle on the oldest available data entry common to all different sources (testing, hospitalisation episodes, mortality records, etc), so as to have epidemiological consistency when studying the transitions from one COVID-19 state to another. All other posterior data entries have been neglected. In this case, it was the HES database that imposed the right-censoring date. Since each data source is collected and posted independently, the end of follow-up date has been updated numerous times and will continue to be in the coming weeks after the presentation of this thesis, for the sake of future publication. This process has required streamlining of all the programming involved so as to be capable of accepting an input of all the different raw data files and processing everything accordingly from there.

2.4 Participants & study size

Eligibility criteria have been simple and straightforward, based on the available data and the required variables of study. The final cohort size is 406,408, and all the particular numbers at each stage of participant selection are outlined in the Results section 3.1.

The first major "cut-off" in cohort size occurs in the restriction of using only England as the location of interest, starting from the wider United Kingdom. Since the majority of registered users reside there, it is not a large number to neglect in comparison to the total. The second largest reduction comes from censoring all those users who

died before the beginning of follow-up (1st February 2020). Only users alive at this time are eligible for study, naturally.
An additional required step is to censor those users who have requested their data not be used for studies. UK Biobank readily provides a list of said user IDs that are to be excluded. The number is again not significant given the study size (a reduction of 21 participants).
The last exclusion occurs depending on whether data on ethnic background is reported or not. Since this variable is self-reported by each registered user, some prefer not to answer and show up in the data as NA (*Not available*). These participants ought to be excluded due to the fact that our model requires this variable to be known for the entire cohort. Furthermore, in the data exploration phase two other small groups with a different ethnicity categorisation from the rest were found. These were individuals who categorised their ethnic background as: *'Do not know'* or *'Prefer not to say'*. Besides the fact that both groups were too small so as to allow for statistical modelling, they had to be excluded because their information was outside the area of interest of this study.

2.5 Variables

The essential exposure variables in this study are: sex, age, ethnic background and socio-economic deprivation, with the latter two being the main focus points of analysis. Needless to say, most of these are potential confounders when it comes to COVID-19 risk assessment, so all have to be accounted for in the model.

Ethnicity has been categorized into five different groups in accordance with the Government Statistical Service's Ethnicity harmonized standard (GSS) [@29] for England. Their official recommendation is that ethnic group data be gathered by asking an individual what group they belong to from the following:

– Asian / Asian British –

1. Indian
2. Pakistani
3. Bangladeshi
4. Chinese
5. Any other Asian background, please describe

– Black / African / Caribbean / Black British –

1. African
2. Caribbean
3. Any other Black background, please describe

– Mixed / Multiple ethnic groups –

1. White and Black Caribbean
2. White and Black African
3. White and Asian
4. Any other Mixed ethnic background, please describe

– White / White British –

1. English / Welsh / Scottish / Northern Irish / British
2. Irish
3. Gypsy or Irish Traveller
4. Any other White background, please describe

– Other ethnic group –

1. Arab
2. Any other ethnic group, please describe

At the time UK Biobank carried out recruitment, the official standard used for the population census was slightly different: Chinese ethnicity was a category on its own, and not part of the wider Asian ethnicity as it is today. Thus, all the granularity available within UKB's ethnic background data was grouped into these five primary categories: Asian, Black, Chinese, Mixed, White and *'other'*.

As previously mentioned, a small number of registered users had self-reported their ethnicity as "do not know", "prefer not to answer" or showed up as "not available". They have been excluded from the study's cohort since they do not allow for the study of ethnicity as a statistically viable covariate. The majority of participants are of White ethnicity (∼94.1%), and the two following largest groups are Asian (∼2.2%) and Black (∼1.8%). Detailed breakthrough of each group's n at every stage in the model is available in the Results section.

Socio-economic deprivation was studied according to the official measure of relative deprivation in England: the English Indices of Deprivation (IoD2) [28]. The distinction between poverty and deprivation is that people are considered to be living in poverty if they lack the financial resources to meet their needs, but are considered deprived if they lack any kind of resources, including income. Thus, the main variables behind the estimates are income and employment but it also takes into account: education and skills training, health, crime, barriers to housing and services, etc (for a total of 39 indicators that make up the final index). It provides measures at a "Lower Layer Super Output Area" (LSOA) level, small territorial units usually of the size of a neighbourhood or postal code, dividing the whole of England into 32,844 units and ranking them relatively from most to least deprived. Participants are therefore assigned a given deprivation index depending on their designated residence. Even though the index is a continuum of values, the participants in this study have been grouped into quintiles (according to England's standards, not just the study's sample population) [@30].

Age was deduced from the provided dates of birth (only the month and year, for privacy purposes) calculated with respect to the date of beginning of follow-up (1st February 2020). All ages are stored with up to two decimal figures, to have a more continuous spectrum. The median age in our cohort study is 70.25, with an interquartile range of [62.6,75.7]. The youngest registered user is 51 and the oldest 87.

Age has been studied first as a continuous variable with a linear relationship with COVID-19 hazard and later on as polynomial dependency. It is now a well-known fact that COVID-19 risk is not exactly linearly dependent with age, which is the reason behind the exploration of a different polynomial relationship. After careful study and given the advanced age of most of the cohort, it was deemed unnecessary to implement non-linear relationships for age and risk in the model, since most transitions were actually best fitted linearly, or very close to linearly.

Sex was provided with no further information on gender or sexuality available.

2.6 Data curation

Data curation has been an essential part of the project and has required the most working hours by far. Essentially, it has consisted on gathering all the files from the aforementioned data sources, analysing and exploring their content to see what was possible to do with modelling, manipulate and filter them, and make them all

compatible with each other. Furthermore, it was important to streamline all the programming to allow for regular updates to the data. By the end of the project, it was possible to download the updated files from the source and have a series of codes curate the data and carry out all the modelling required with no manual intervention. The goal behind this is to have the ability to easily update results for posterior publication of an article, as more data is gathered and published.

Most data registries from UKB and HES have a multiplicity of entries for an individual user. In the case of testing data, for example, there would be one for every test the patient has taken. This applies for all other databases, including mortality records where there is an entry for each diagnosis at the time of death of a patient. For multi-state modelling it was necessary to gather all data sources and combine them into one large dataframe with one row per patient. As such, each entry is a row containing all COVID-19 events of interest and population characteristics of a given patient in the cohort.

The end-product is a large file containing a 406,408x19 one-row-per-patient dataframe with all the relevant information for multi-state expansion and modelling.

2.7 Statistical methods

2.7.1 Modelling COVID-19

To model the virus' health outcomes, extensive data exploration was required beforehand to know precisely what could and could not be done with the data at hand. One of the first challenges to arise was to determine how to count hospital episodes and deaths based on the data at hand.

The 28-day principle. There are two different methodologies that are commonly used for COVID-19 data recollection and analysis to establish how to count coronavirus related deaths and hospitalisations. The simplest option is to select COVID-19 patients by medical diagnosis: when a doctor registers the hospital episode or death as being caused by the virus. Although, it is straightforward as far as data recollection goes, the problem is that it requires unification between the different hospital and medical organisations in the way they report it. It is well-known that at the beginning of the pandemic no such standard existed, as it was a new virus. In the case of the United Kingdom, it was not until the ICD-10 (International Statistical Classification of Diseases and Related Health problems) [31] registered new codes for coronavirus disease (U071 & U072) that such a standard began to be adopted. Although its use is now widespread, many cases were not registered in the beginning of the pandemic. Furthermore, it is not always trivial to diagnose a hospital episode or death as being caused by coronavirus. We found that in HES's hospitalisation data there were very few episodes registered as being caused by COVID-19 and many who were left as unspecified or unknown and were from around the time when the virus was taking its heaviest toll on the medical system. Furthermore, many patients were found to have tested positive at around the time of hospital admission. All in all, this methodology created too many complications to be practical. If this principle is summarised as *due to* COVID, the second principle would then be *with* COVID. Following this methodology, any hospitalisation or death that occurred within 28 days of an individual having tested positive is counted as *with* COVID. Many countries and institutions have adopted this principle throughout the pandemic as it circumvents many of the problems of using the diagnosis principle. It is important to note that this methodology relies solely on testing records being comprehensive and readily available. Fortunately, after the first months of the pandemic most European countries began to do so as more tests became available for public use, the United Kingdom included. After much deliberation and data exploration, this project settled on using this methodology for all states in the model.

After it became evident which principle to use for defining each state, a number of different multi-state structures were tested. Through weekly meetings I would present the week's work and show preliminary results as well as early model results to discuss and shape the project. Figure 2.1 displays the final transition diagram used for modelling.

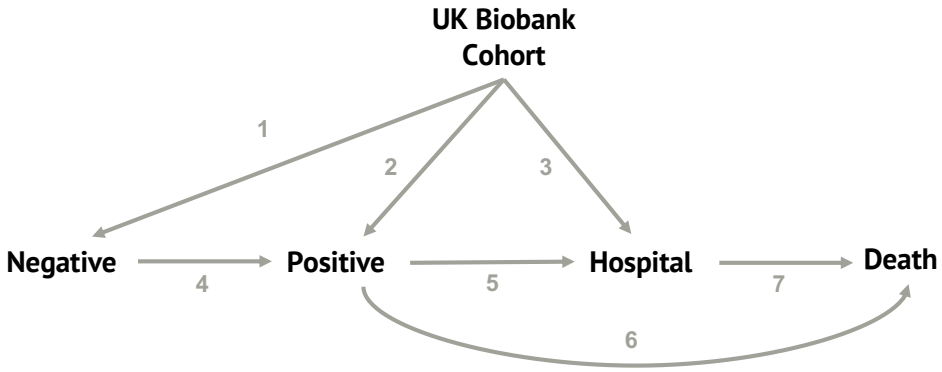


Fig. 2.1: Transition structure of the multi-state model created. All eligible patients start from the initial UKB state and may transition from there if any of the shown COVID-19 risk events occur. The only absorbing state is Death.

All participants begin follow-up as part of the wider UKB cohort and may transition to testing negative, testing positive or direct hospitalisation. One may quickly realise that, by implementing the 28-day principle, the transitions to Hospital/Death would have to start from the *testing states*. After all, it is necessary to have a positive test before the transition to Hospital or Death is possible. The explanation behind the third transition (UKB-Hospital) is that there were a number of *with* COVID-19 hospitalisations were the positive test occurred in the same day as hospital admission. Since this leads to numerical divergence and breaks the assumption that event times are distinct, the third transition was artificially put in place. Those cases were then manually set to bypass the second transition towards the Positive state and go directly from UKB to Hospital.

The same phenomenon was present for immediate transitions from Hospital to Death or Positive to Death, so another artificial transition UKB-Death was also experimented with. Besides having too little n for hazard ratio confidence intervals to actually have a physical interpretation, this transition was deemed unrepresentative of the reality of the situation. What was done instead was to manually add half a day to the date of death for said instances and keep those cases in the original transitions (6 & 7). This way numerical convergence was achieved without distorting the data in a transcendental way. It was agreed upon that this necessary solution

more accurately portrayed those cases than an artificial UKB-Death transition since it is more realistic to expect that those deaths occurred soon after testing positive or being hospitalised, but not immediately.

For patients with multiple COVID-19 tests, only the first negative test and subsequent positive test are counted. If any other tests were taken after having entered the Positive state, they are neglected. Since all transitions are unidirectional by design, an individual who has their first negative test after having tested positive goes straight to Positive from UKB (transition 2) and the negative test is neglected.

Another '*absorbing state*' for recovered patients was discussed and experimented with. In the end, it was deemed unnecessary since it added no real information to the results obtained from all other hazard ratios and it only served to further complicate the modelling and programming.

All in all, the combination the of Cox proportional hazards methodology with this COVID-19 model made it possible to obtain hazard ratios for each transition and covariate of interest.

2.7.2 The Cox proportional hazards model

The Cox proportional hazards model [32] is a regression model widely used in epidemiological and statistical medicine research for analysing the association between a set of patient covariates and their survival times.

Common covariates in medical studies are sex, age, previous comorbidities, drug treatments, operations, etc. What is generally understood by survival times is the time that passes before certain events of interest occur. The hazard or risk of an event occurring, $\lambda(t)$, is probabilistic and varies over time. It can be interpreted as being indirectly proportional to the survival function, $S(t)$, which describes how an individual's chances of not experiencing the event (i.e. surviving it) decrease as time passes. These functions are unique to every patient depending on their characteristics but share a common time-dependent term representative of the proportional hazards nature of the model.

Mathematically, hazard rates (or transition intensities) can be defined as:

$$\lambda_{ij}(t) = \lim_{\Delta t \to 0} \frac{Prob(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2.1}$$

Where T denotes the time of reaching sate j from state i in the $i \to j$ transition. We can therefore easily define the cumulative hazard for the $i \to j$ transition as:

$$\Lambda_{ij}(t) = \int_0^1 \lambda_{ij}(s)\, ds \tag{2.2}$$

And, because hazard completely describes the survival distribution, it can also be derived from the survival function:

$$\lambda_{ij}(t) = \frac{1}{S_{ij}(t)} \lim_{\Delta t \to 0} \frac{S_{ij}(t) - S_{ij}(t + \Delta t)}{\Delta t} = -\frac{d \log S_{ij}(t)}{dt} \tag{2.3}$$

So, finally, combining equations 2.2 and 2.3 the following relation is obtained:

$$S_{ij}(t) = \exp(-\Lambda_{ij}(t)) \tag{2.4}$$

Proportional hazards assumption. In proportional hazards models, like Cox's, the key underlying assumption is that the effect of covariates is only multiplicative with respect to hazard. In other words, all participants share the same **baseline hazard curve**, $\lambda_0(t)$, but it is multiplied by the coefficients associated to each covariate (which are different for every individual). It is more simply explained by its mathematical expression:

$$\lambda_{ij}(t|Z) = \lambda_{ij,0}(t) \exp(\beta_{ij}^T Z) \tag{2.5}$$

Where $\lambda_{ij,0}(t)$ is the aforementioned baseline hazard function for the $i \to j$ transition and β_{ij} the **vector of regression coefficients** that adjust the effect of the **covariate vector** Z for the same transition. Naturally, $\beta_{ij}^T Z$ denotes $\sum_{K=1}^{p} \beta_{ij,k} \times Z_k$ (with p being the total number of covariates implemented).

There are multiple methods for testing the proportional hazards assumption beforehand in order to see whether a proportional hazards model is adequate for the data. This is a necessary procedure in multi-state modelling of the kind performed in this study because it is a crucial assumption in the derivation of its formulas and it underpins the very concept of hazard ratios. One of the most widely methods consists on graphically plotting survival curves for individuals with different characteristics using the Kaplan-Meier estimator [33]. This is explained in more detail in the last section under *Statistical methods* 2.7.4.

The vector of regression coefficients β_{ij} is obtained by maximising the Cox partial likelihood. All event times are assumed to be distinct in order to perform this

calculation. It is easy to check said assumption computationally because if it is not the case one obtains divergence in the results. From now on the ij transition indices will be omitted for ease of notation purposes, with the understanding that what is presented would be performed for each transition separately.

Breslow's estimate of the baseline hazard function. The Cox partial likelihood is obtained using Breslow's estimate of the baseline cumulative hazard:

$$\widehat{\Lambda}_0(t) = \sum_{j:t_j \leq t} \frac{1}{\sum_{l \in R_j} \exp(\widehat{\beta}^T Z_l)} \tag{2.6}$$

Where R_j denotes the set at risk at event time t_j and j is now used as a general index (unrelated to the previous transition numbers).

By plugging 2.6 into the full likelihood of the event being observed occurring for a subject at a given time, the result is the product of two factors. The first factor has no dependence on the regression coefficients and only depends on the censoring pattern in the data. The second factor is the following partial likelihood:

$$L(\beta) = \prod_{j=1}^{N} \frac{\exp(\beta^T Z_j)}{\sum_{l \in R_j} \exp(\widehat{\beta}^T Z_l)} \tag{2.7}$$

Recalling the hazard function's expression (2.5), the last equation is simply the product, over the event times, of quotients that compare the hazard of an individual with the event occurring at t_j with the hazard of all individuals at risk of that event at t_j (including the subject himself).

By maximising the partial likelihood in 2.7, one obtains the vector of regression coefficients β and from there a series of hazard ratio coefficients (essentially e^{β}) that quantify the relative effect of each covariate on hazard. Since they are statistical regression coefficients, they have associated confidence intervals and p-values that depend on the n at each transition and on the nature of the event times in the data.

One of the greatest insights of Cox's model is that the effect of said covariates can be estimated without having to model the baseline hazard function $\lambda_0(t)$. Thanks to the proportional hazards assumption, the common baseline hazard function cancels out in the quotients in equation 2.7.

Confounder adjustment. Another powerful aspect of Cox's model is that it is very simple to see how confounders are adjusted for. By looking at the $\exp(\beta_{ij}^T Z)$ term

in equation 2.5, it is clear that when all confounder variables are included in the regression, risk is assigned to each of them jointly. If only one covariate was used for modelling, the results would be fully confounded hazard ratios where partial likelihood maximisation would have occurred over just one parameter. If, on the other hand, all covariates are inputted simultaneously, the model distributes risk taking into account the separate effect of each of them on the overall hazard function.

Time scales. There are two frequently used approaches for scaling time to events in multi-state models: *'clock forward'* and *'clock reset'* [19]. In the latter, the time argument t in $\lambda_{ij}(t)$ refers to the time since entry to state i, so the clock is reset whenever the patient enters a new state. This method is commonly referred to as backward recurrence time. This project adopted the former approach. In *'clock forward'* systems, times are always valued with respect to the origin: when a patient enters the initial state and begins follow-up. So the clock only moves forward, even when transitions occur.

2.7.3 Poisson regression

A different statistical methodology was adopted for analysing testing data as stratified by socio-economic deprivation and ethnic background. The idea arose from the fact that UKB's data holds records for all COVID-19 tests patients in the cohort have taken and that it was possible to count them and incorporate it into the larger database created for the study. With multi-state modelling it is only possible to calculate hazard ratios for one test per person, which is why only the first negative and positive tests were taken.

Assuming a Poisson distribution, the Poisson regression generalised linear model [34] was used to fit the data. The Poisson distribution is as follows:

$$Pr(Y = y) = \frac{(\lambda V)^y}{y!} e^{-\lambda V} \tag{2.8}$$

Where V may be understood as the typical time parameter despite actually representing patients-time.

One essential characteristic of Poisson distribution and regression is the **equidispersion** property: the expectation value and variance of the random variable Y are equal.

$$E(Y) = Var(Y) = \lambda V \tag{2.9}$$

The comparison between different strata of the cohort is derived as follows:

$$ln[E(Y)] = ln(\lambda V) = ln(\lambda) + ln(V) \tag{2.10}$$

And assuming:

$$ln(\lambda) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{2.11}$$

So that, for analysing sex differences, for example:

$$ln(\lambda) = \beta_0 + \beta_1 Male + \beta_2 Age + ... \tag{2.12}$$

And:

$$\lambda = e^{\beta_0 + \beta_1 Male + \beta_2 Age + ...} \tag{2.13}$$

Finally, it is possible to define the ***risk ratio*** as:

$$RR(male\ Vs\ female) = \frac{e^{\beta_0 + \beta_1 + \beta_2 Age + ...}}{e^{\beta_0 + \beta_2 Age + ...}} = e^{\beta_1} \tag{2.14}$$

By regression with respect to the data it is possible to obtain such relative β coefficients for different groups, with different levels of confounding (note how the example only shows Age but could also include the other covariates). In practice, this methodology allowed for proper analysis of testing differences in the UKB cohort depending on a patient's characteristics.

It is important to note that, due to over-dispersion in the distribution (i.e., variance greater than the mean), the ***Quasi-Poisson*** model [35] was shown side by side with the regular model for more accurate estimates of the standard error. Essentially, it is a generalisation of the Poisson regression which assumes that the variance is a linear function of the mean.

2.7.4 Kaplan-Meier estimator

As previously mentioned, the Kaplan-Meier estimator [36] may be used to graphically check the proportional hazards assumption. It is based on a simple concept: so long as the independent censoring assumption holds, $\lambda(t_j)$ can be estimated simply by

the at risk sample proportion that may fail (have the event of interest occur) at t_j. Mathematically:

$$\widehat{\lambda}(t_j) = \frac{d_j}{n_j} \tag{2.15}$$

Where d_j represents the number of observed events at t_j and n_j the size of the set of patients at risk of the event at t_j.

Consequently, the probability of survival up to time t_j would then be the product of the probability of survival up to t_{j-1} (given by the survival function) and the conditional probability of survival at t_j:

$$\widehat{S}(t_j) = \widehat{S}(t_{j-1})(1 - \widehat{\lambda}(t_j)) = \widehat{S}(t_{j-1})\left(1 - \frac{d_j}{n_j}\right) \tag{2.16}$$

By applying the previous concept repeatedly, one obtains the Kaplan-Meier estimator:

$$\widehat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.17}$$

The larger the sample size, the more it approaches a continuous distribution.

This method is straightforward, it can even be done by hand, and provides useful estimation of the survival function. In this study in particular, it makes it possible to plot survival curves stratified by the covariates of interest. The proportional hazards assumption establishes that these curves should be proportional and therefore never cross. If the graphs do not exhibit this behaviour, the Cox proportional hazards model may not be right for the data. Some of the estimated survival plots for this study are presented in the Results section 3.6.

2.8 R resources used

Although it is widely used in medical data science and statistics, the **R programming language** [@37] is not taught as part of the Engineering Physics Bachelor's degree of which this thesis is the final work. At the beginning of the project, other, more familiar, programming languages were used (MATLAB and Python) for early analysis. After the first weeks, all work began to be carried out with R. A learning process was required to learn to use the language proficiently and to incorporate all the external software packages needed (listed below).

Most data work was carried out through the *tidyverse* collection of R packages [@38]. It is a library of data science tools with shared grammar and data structures. Among the most used in this project are: *dplyr*, *tidyr* and *readr*. These all provide dataframe manipulation tools. The package *ggplot2* was also used for graphical representation, though often extended with other survival analysis graphical packages. The combination of *tidyverse* with the visualisation tools of the *RStudio* IDE [@39] provided strong data science capabilities without which this project would not have been possible.

For multi-state modelling, the *mstate* package [@40] allowed for building the necessary data structures for regression. Taking the one-row-per-patient dataframe, it expands it into 'long format' and includes the transition-specific covariates [41]. Then, the Cox proportional hazards coefficients are obtained through regression with the *survival* package [@42]. Kaplan-Meier estimates of the survival function were obtained with *survivalAnalysis* [@43]. Furthermore, the *mfp* package [@44] (Multivariable Fractional Polynomials) was used for experimentation with non-linear treatment of age with respect to hazard.

Additional resources

Stock date handling capabilities in the R language are limited, so *lubridate* [@45] was used throughout all data curation. It served to make all date formats from the different sources compatible with each other and create time-dependent filters for censoring, among other things.

The *reader* package [@46] was used to simplify reading data from files. For specially large data frames that could not be be loaded onto the RAM directly, the *data.table* package [@47] made it possible to read only the sections of interest from the data.

A different set of resources were used for graphical results and tabulation, often in conjunction with R Markdown software (*knitr*). All tables were produced with *kable* and *kableExtra* [@48]. For Kaplan-Meier survival plots, the *survminer* package [@49] was used. Radar charts were created using *viridis*, *patchwork*, *hrbrthemes*, *fmsb* and *colormap*.

The *pacman* package manager [@50] was necessary given the number of external resources used.

Results

3

3.1 Participants & population flow diagram

From a starting cohort size of just over half a million participants, a sizeable reduction is made in order to arrive at a final, usable, cohort of 406,408. The principles behind these exclusions, as shown in Figure 3.1, are:

- Requests that personal data not be used for studies after registering with UK Biobank. The program readily provides a list of these mandatory exclusions.
- Participants who died before the beginning of follow-up: 1st February 2020.
- Participants who live outside England are not ranked within the English Indices of Deprivation (IoD). Since socio-economic deprivation is one of the key variables of analysis, the location of interest of this study has been forcibly reduced to England and all other participants have been excluded.
- Self-reported ethnicity as: '*Do not know*' or '*Prefer not to answer*'.
- Ethnicity shows NA (*not available*). Data was not recorded or is missing.

No other eligibility criteria have been applied besides these necessary exclusions. All patients who are available for study and have the necessary variables recorded have been included in the cohort.

The possibility of non-participation throughout follow-up is not really an option, besides the event of death or the request to be excluded from any studies whatsoever. Otherwise, all tests and hospital records are recorded by either HES or UK Biobank from the beginning of follow-up to the end (20th March 2021) for all users.

The final cohort size of 406,408 allows for different levels of stratification depending on group identity. When it comes to ethnic background, however, some groups within the study population do not have enough participants suffering from an event for precise regression estimates. This is particularly problematic for the latter transitions, like the Positive-Death transition, where there are too few individuals for calculation and sometimes none at all (in which case model results diverge). The exact numbers are presented in the following sections.

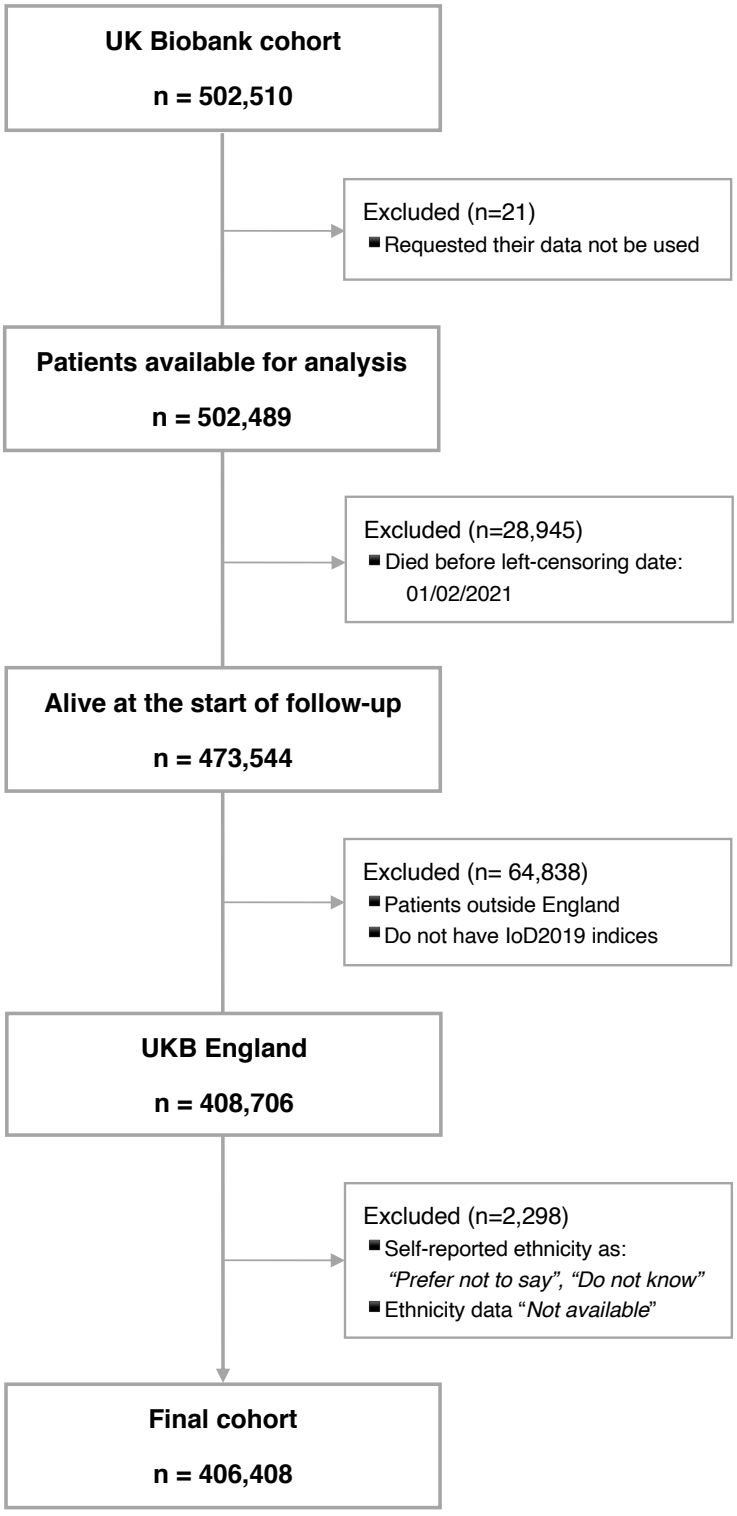


Fig. 3.1: Population flow chart outlining step exclusions in cohort size.

3.2 Descriptive data

All participants in the study share some demographic characteristics: they are all English, of advanced age, and, chronologically, had the same exposure to COVID-19. They differ considerably in other traits like socio-economic status, although there is higher prevalence among the upper quintiles of the socio-economic deprivation index (least deprived): 29.7% in Q1 and 23.9% in Q2. Ethnically, the sample is varied but shows a large White majority (94.1%).

By design, no participants in the final cohort have missing data for the variables of interest and the covariates, since it is required for multi-state modelling. Therefore, the exclusions shown in the population flow diagram (Figure 3.1) also reflect all the necessary reductions in cohort size to fulfill this condition.

The maximum follow-up time for any participant is 413 days, from left-censor to right-censor dates. The only event to definitively end follow-up prematurely is death.

3.2.1 Population characteristics

Table 3.1 summarises the cohort's population characteristics.

Tab. 3.1: Cohort's population characteristics. Stratified by ethnicity and displaying sex, age and socio-economic deprivation characteristics. Age is presented by group median age alongside interquartile range. All percentages (in brackets) are calculated with respect to group size *n*.

	All	Asian	Black	Chinese	Mixed	Other	White
n	406408	8810	7299	1327	2541	3965	382466
Sex=Male (%)	181894 (44.8)	4654 (52.8)	3026 (41.5)	469 (35.3)	926 (36.4)	1686 (42.5)	171133 (44.7)
Age (median [IQR])	70.2 [62.6,75.7]	65 [58.4,72.2]	62.8 [57.7,69.8]	64.8 [58.7,70.8]	62.5 [57.3,70.2]	64.3 [58.1,71.1]	70.7 [63,75.8]
Socio-economic deprivation							
Q1 (%)	120663 (29.7)	1347 (15.3)	390 (5.3)	370 (27.9)	492 (19.4)	567 (14.3)	117497 (30.7)
Q2 (%)	97174 (23.9)	1454 (16.5)	586 (8)	253 (19.1)	502 (19.8)	559 (14.1)	93820 (24.5)
Q3 (%)	72874 (17.9)	1831 (20.8)	1111 (15.2)	283 (21.3)	437 (17.2)	703 (17.7)	68509 (17.9)
Q4 (%)	63557 (15.6)	2275 (25.8)	2075 (28.4)	229 (17.3)	527 (20.7)	963 (24.3)	57488 (15)
Q5 (%)	52140 (12.8)	1903 (21.6)	3137 (43)	192 (14.5)	583 (22.9)	1173 (29.6)	45152 (11.8)

The age variable is represented by each group's median age and its interquartile range, to show what the bulk of the population's age is. The advanced age of most the cohort (by UKB's design) is evident: 70.25, IQR=[62.6,75.7]. It is also notable

that Whites (70.7 [63,75.8]) are about 5 years older than the other ethnic groups, which gives further indication of the importance of treating age as a significant confounder.

The relationship between socio-economic deprivation and ethnic background in the sample can be shown using radar charts, as in Figure 3.2, by quantifying the percentage of individuals that belong to a given socio-economic deprivation quintile with respect to the group's size.

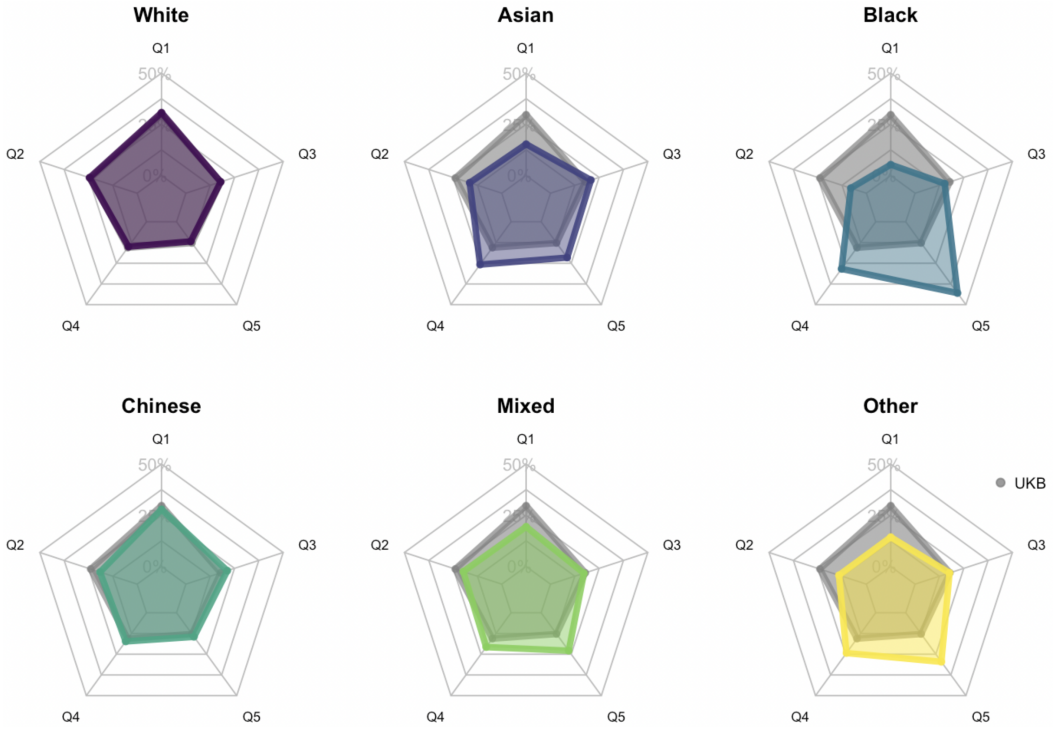


Fig. 3.2: Radar charts for socio-economic deprivation quintile population distribution, stratified by ethnicity. Percentages are calculated as the fraction of group members within a given a quintile. Shown in grey is the distribution of the entire cohort, without stratifying.

As mentioned in the Methods section, socio-economic deprivation quintiles are set according to the national ranking, not the study's cohort. Thus, if UK Biobank's English participants had been selected in a completely unbiased manner the overall distribution would much more closely resemble a perfect pentagon. Instead, what is observed is a slightly higher concentration in the upper quintiles, displaying the socio-economic bias in the sample population.

Despite the sample's overall tendency for the less deprived quintiles, inequalities between the ethnic groups are still stridently evident. All groups except Chinese participants are generally more deprived than Whites, with the Black and *'other'*

groups having the biggest gap. These differences in socio-economic status point to the importance of studying both socio-economic deprivation and ethnic background as variables of interest and mutual confounders separately.

3.3 Outcome data

Health outcomes in the study's model are measured directly by the population size at each transition, depending on the stratification of interest. All of the following percentages shown in Table 3.2 are relative to the population size that was susceptible to undergo each particular transition. In other words, it is the fraction of those who transitioned with respect to those who could have. The two age classifications are in fact chosen according to the median age of the cohort: 70.25.

Tab. 3.2: Outcome event numbers. Stratified by ethnicity, socio-economic deprivation quintile, sex and age. All percentages (in brackets) are calculated as the fraction of individuals who undergo a particular transition from the total who were in the states preceding it (susceptible to transitioning). Age is divided into those younger and older than the median, 70.25.

	Total	UKB-Negative	Negative-Positive	UKB-Positive	Positive-Hospital	Positive-Death	UKB-Hospital	Hospital-Death
All (%)	406408	56884 (14)	1316 (2.3)	11628 (2.9)	1199 (9.3)	117 (0.9)	2257 (0.6)	728 (21.1)
Ethnicity								
Asian (%)	8810	1140 (12.9)	55 (4.8)	554 (6.3)	63 (10.3)	3 (0.5)	80 (0.9)	29 (20.3)
Black (%)	7299	982 (13.5)	49 (5)	347 (4.8)	52 (13.1)	0 (0)	93 (1.3)	28 (19.3)
Chinese (%)	1327	111 (8.4)	1 (0.9)	25 (1.9)	3 (11.5)	0 (0)	9 (0.7)	2 (16.7)
Mixed (%)	2541	344 (13.5)	15 (4.4)	99 (3.9)	9 (7.9)	0 (0)	13 (0.5)	3 (13.6)
Other (%)	3965	522 (13.2)	34 (6.5)	168 (4.2)	17 (8.4)	1 (0.5)	29 (0.7)	5 (10.9)
White (%)	382466	53785 (14.1)	1162 (2.2)	10435 (2.7)	1055 (9.1)	113 (1)	2033 (0.5)	661 (21.4)
Socio-conomic deprivation								
Q1 (%)	120663	17038 (14.1)	273 (1.6)	2492 (2.1)	232 (8.4)	28 (1)	455 (0.4)	147 (21.4)
Q2 (%)	97174	13493 (13.9)	263 (1.9)	2489 (2.6)	208 (7.6)	20 (0.7)	403 (0.4)	110 (18)
Q3 (%)	72874	9930 (13.6)	249 (2.5)	2180 (3)	224 (9.2)	17 (0.7)	406 (0.6)	140 (22.2)
Q4 (%)	63557	8951 (14.1)	252 (2.8)	2230 (3.5)	246 (9.9)	29 (1.2)	445 (0.7)	142 (20.5)
Q5 (%)	52140	7472 (14.3)	279 (3.7)	2237 (4.3)	289 (11.5)	23 (0.9)	548 (1.1)	189 (22.6)
Sex								
Male (%)	181894	26246 (14.4)	568 (2.2)	5305 (2.9)	683 (11.6)	69 (1.2)	1289 (0.7)	466 (23.6)
Female (%)	224514	30638 (13.6)	748 (2.4)	6323 (2.8)	516 (7.3)	48 (0.7)	968 (0.4)	262 (17.7)
Age								
Younger than 70 (%)	202637	23949 (5.9)	791 (1.4)	8315 (2)	592 (4.6)	12 (0.1)	680 (0.2)	146 (4.2)
Older than 70 (%)	203771	32935 (8.1)	525 (0.9)	3313 (0.8)	607 (4.7)	105 (0.8)	1577 (0.4)	582 (16.8)

In the following sections, hazard ratios for the Positive-Death transition are omitted. It is clear from Table 3.2 why that is: when stratifying by ethnic background most groups have no individuals undergoing that transition. Consequently, regression calculations diverge or result in confidence intervals that are too wide to provide any real information about hazard ratios.

3.4 Main results

3.4.1 Ethnicity: hazard ratio coefficients

The results from the Cox proportional hazards model are relative hazard ratio coefficients that display the effect of belonging to a given group on the baseline hazard.

The numbers presented below are the exponential of the components of the vector of regression coefficients (e^{β_i}) that adjust the effect of the covariate vector Z, as explained in the *Statistical methods* 2.7.2 section. Thus, they can be interpreted as the direct multiplicative effect on hazard that belonging to a given ethnicity is associated to (with respect to the reference majority group: Whites). Table 3.3 displays these transition-specific coefficients for different levels of confounder adjustment.

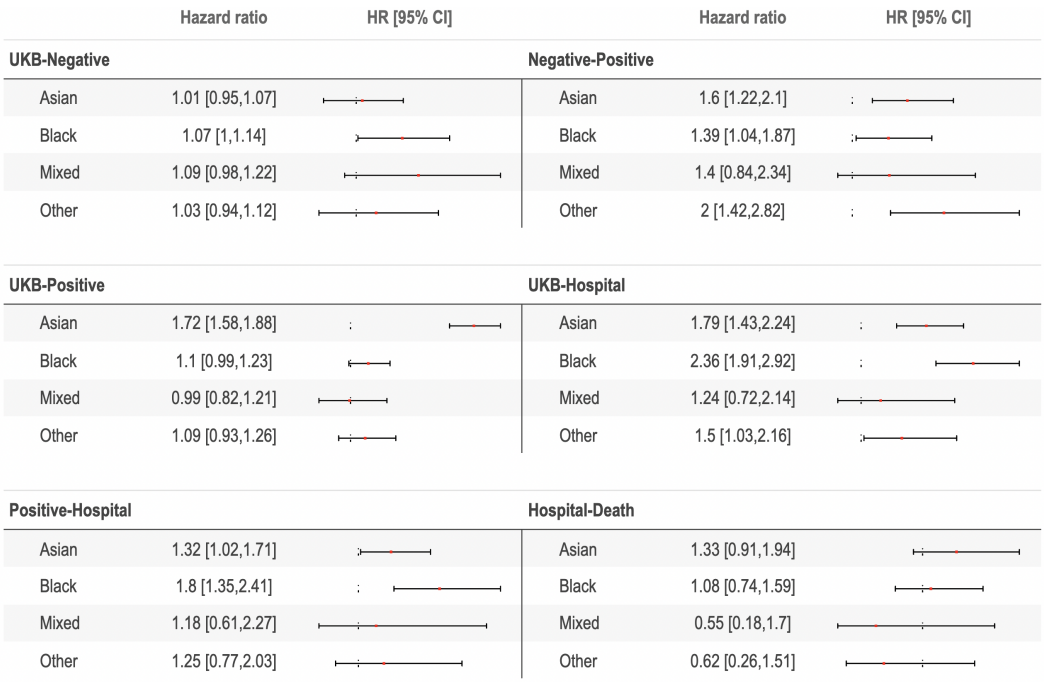
For no confounder adjustment at all, *'unadjusted'* is written. In these cases regression is calculated with hazard depending exclusively on ethnicity. Then, regression is calculated taking into account age and sex as relevant covariates. Finally, *'fully adjusted'* stands for regression coefficients that are computed taking into account ethnicity, age, sex and socio-economic deprivation.

Tab. 3.3: Hazard ratio coefficients stratified by ethnicity, shown at different levels of confounding[a,b]. All coefficients are relative to the reference group, Whites.

	UKB-Negative	Negative-Positive	UKB-Positive	Positive-Hospital	UKB-Hospital	Hospital-Death
Asian vs White						
Unadjusted	0.94 [0.88,0.99]	2.04 [1.56,2.68]	2.36 [2.17,2.57]	1.19 [0.92,1.53]	1.74 [1.39,2.17]	0.94 [0.65,1.36]
Age-sex adjusted	1.02 [0.96,1.08]	1.71 [1.3,2.24]	1.89 [1.73,2.06]	1.45 [1.12,1.87]	2.11 [1.69,2.64]	1.37 [0.94,1.99]
Fully adjusted	1.01 [0.95,1.07]	1.6 [1.22,2.1]	1.72 [1.58,1.88]	1.32 [1.02,1.71]	1.79 [1.43,2.24]	1.33 [0.91,1.94]
Black vs White						
Unadjusted	0.97 [0.91,1.04]	2.15 [1.62,2.86]	1.77 [1.59,1.97]	1.46 [1.1,1.93]	2.43 [1.97,2.99]	0.79 [0.54,1.16]
Age-sex adjusted	1.1 [1.04,1.18]	1.66 [1.24,2.21]	1.33 [1.2,1.48]	2.06 [1.55,2.73]	3.45 [2.8,4.25]	1.13 [0.77,1.65]
Fully adjusted	1.07 [1,1.14]	1.39 [1.04,1.87]	1.1 [0.99,1.23]	1.8 [1.35,2.41]	2.36 [1.91,2.92]	1.08 [0.74,1.59]
Chinese vs White						
Unadjusted	0.57 [0.48,0.69]	0.45 [0.06,3.2]	0.66 [0.45,0.98]	1.73 [0.56,5.38]	1.23 [0.64,2.37]	0.49 [0.12,1.95]
Age-sex adjusted	0.64 [0.53,0.77]	0.37 [0.05,2.63]	0.53 [0.36,0.79]	2.6 [0.84,8.09]	1.72 [0.9,3.32]	0.58 [0.15,2.34]
Fully adjusted	0.64 [0.53,0.77]	0.36 [0.05,2.57]	0.52 [0.35,0.77]	2.65 [0.85,8.24]	1.66 [0.86,3.19]	0.58 [0.15,2.34]
Mixed vs White						
Unadjusted	0.97 [0.87,1.08]	2.05 [1.23,3.42]	1.43 [1.18,1.75]	0.89 [0.46,1.71]	0.96 [0.56,1.66]	0.48 [0.15,1.49]
Age-sex adjusted	1.1 [0.99,1.23]	1.53 [0.92,2.55]	1.07 [0.88,1.31]	1.23 [0.64,2.37]	1.42 [0.82,2.45]	0.56 [0.18,1.74]
Fully adjusted	1.09 [0.98,1.22]	1.4 [0.84,2.34]	0.99 [0.82,1.21]	1.18 [0.61,2.27]	1.24 [0.72,2.14]	0.55 [0.18,1.7]
Other vs White						
Unadjusted	0.94 [0.86,1.03]	2.9 [2.06,4.08]	1.56 [1.34,1.82]	1.02 [0.63,1.64]	1.38 [0.96,1.99]	0.45 [0.19,1.09]
Age-sex adjusted	1.05 [0.96,1.14]	2.24 [1.59,3.16]	1.23 [1.06,1.43]	1.32 [0.82,2.14]	1.87 [1.3,2.7]	0.62 [0.26,1.49]
Fully adjusted	1.03 [0.94,1.12]	2 [1.42,2.82]	1.09 [0.93,1.26]	1.25 [0.77,2.03]	1.5 [1.03,2.16]	0.62 [0.26,1.51]

[a] Fully adjusted for age, sex and socio-economic deprivation.
[b] [] <- 95% Confidence intervals

For graphical representation of these hazard ratios and their confidence intervals, one may use forest plots. Figure 3.3 only shows the fully adjusted coefficients since they are the main point of interest.



[a] Fully adjusted for age, sex and socio-economic deprivation.
[b] [] <- 95% Confidence intervals

Fig. 3.3: Forest plots for ethnicity-stratified hazard ratios, fully adjusted [a,b]. The scale used for visualisation is linear but differs from transition to transition. The two dots reflect the $x = 1$ axis.

It is clear that for the latter transitions in the model the confidence intervals widen considerably. It is the effect of the lower n at these stages, particularly when stratifying the cohort.
Chinese participants have been excluded from Fig 3.3 because they are one of the groups with the lowest n and their hazard ratio confidence intervals widen too greatly for the other results to be displayed properly. The numbers are displayed in Table 3.3 however.

Higher risk for ethnic minorities is found across the board, with the exception of positive and negative testing for Chinese patients. Poisson regression results, shown in the next section, back up this result. The Hospital-Death transition results are inconclusive for Chinese, Mixed and *'other'* ethnic groups, given the confidence intervals obtained.
Intensive analysis of the results obtained is reserved for the *Discussion* section.

3.4.2 Ethnicity: Poisson regression on testing

To analyse group differences in COVID-19 testing, Poisson regression provides relative estimates that can also be stratified by ethnic background.

Table 3.4 shows the estimated Poisson coefficients, their standard error and the associated risk ratios. The latter is simply derived from the coefficients by taking the exponential, as shown in the equations presented in the *Methods* section 2.14. The results ought to be interpreted much in the same way as with the previous hazard ratios, but in this case understanding Poisson risk ratios as the difference in the average ***tests per person*** of each ethnic group with respect to Whites. As previously mentioned, the Quasi-Poisson model is shown side by side for more accurate estimates of the standard error, since there is overdispersion in the sample.

Tab. 3.4: Poisson regression coefficients on number of tests taken, stratified by ethnicity and shown at different levels of confounding[a]. Risk ratios are the exponential of the regression estimates, with the corresponding confidence intervals[b]. Quasi-Poisson model results are shown due to overdispersion in the sample.

	Poisson			Quasi-Poisson		
	Estimate	S.E.	Risk ratio	Estimate	S.E.	Risk ratio
Asian vs White						
Unadjusted	0.229	0.016	1.26 [1.22,1.3]	0.229	0.037	1.26 [1.17,1.35]
Age-sex adjusted	0.288	0.016	1.33 [1.29,1.38]	0.288	0.037	1.33 [1.24,1.43]
Fully adjusted	0.234	0.017	1.26 [1.22,1.31]	0.234	0.037	1.26 [1.17,1.36]
Black vs White						
Unadjusted	0.312	0.017	1.37 [1.32,1.41]	0.312	0.039	1.37 [1.27,1.47]
Age-sex adjusted	0.407	0.017	1.5 [1.45,1.55]	0.407	0.039	1.5 [1.39,1.62]
Fully adjusted	0.281	0.018	1.32 [1.28,1.37]	0.281	0.040	1.32 [1.23,1.43]
Chinese vs White						
Unadjusted	-0.489	0.060	0.61 [0.55,0.69]	-0.489	0.135	0.61 [0.47,0.8]
Age-sex adjusted	-0.405	0.060	0.67 [0.59,0.75]	-0.405	0.136	0.67 [0.51,0.87]
Fully adjusted	-0.417	0.060	0.66 [0.59,0.74]	-0.417	0.135	0.66 [0.51,0.86]
Mixed vs White						
Unadjusted	0.016	0.034	1.02 [0.95,1.09]	0.016	0.076	1.02 [0.88,1.18]
Age-sex adjusted	0.119	0.034	1.13 [1.05,1.2]	0.119	0.077	1.13 [0.97,1.31]
Fully adjusted	0.074	0.034	1.08 [1.01,1.15]	0.074	0.076	1.08 [0.93,1.25]
Other vs White						
Unadjusted	0.145	0.025	1.16 [1.1,1.21]	0.145	0.057	1.16 [1.03,1.29]
Age-sex adjusted	0.226	0.025	1.25 [1.19,1.32]	0.226	0.058	1.25 [1.12,1.4]
Fully adjusted	0.152	0.025	1.16 [1.11,1.22]	0.152	0.058	1.16 [1.04,1.3]

[a] Fully adjusted for age, sex and socio-economic deprivation.
[b] [] <- 95% Confidence intervals

All ethnic groups except Chinese are found to have a high risk ratio of testing for COVID-19 in comparison to Whites. In the case of Asian and Black patients, the difference is as large as 25%. Chinese patients appear to be take many less tests on average than Whites.

3.4.3 Socio-economic deprivation: hazard ratio coefficients

In the same manner, it is possible to stratify by socio-economic deprivation quintile instead and analyse the effect of belonging to a given group on COVID-19 hazard, as Table 3.5 illustrates through the obtained hazard ratio coefficients.

Tab. 3.5: Hazard ratio coefficients stratified by soico-economic deprivation quintile, at different levels of confounding[a,b]. All coefficients are relative to the reference group, the least deprived quintile Q1.

	UKB-Negative	Negative-Positive	UKB-Positive	Positive-Hospital	UKB-Hospital	Hospital-Death
Q2 vs Q1						
Unadjusted	0.99 [0.96,1.01]	1.2 [1.02,1.43]	1.24 [1.18,1.31]	0.86 [0.72,1.04]	1.1 [0.96,1.26]	0.87 [0.68,1.11]
Age-sex adjusted	0.99 [0.97,1.01]	1.19 [1.01,1.41]	1.23 [1.16,1.3]	0.88 [0.73,1.06]	1.11 [0.97,1.27]	0.9 [0.7,1.15]
Fully adjusted	0.99 [0.97,1.01]	1.19 [1,1.41]	1.22 [1.16,1.29]	0.88 [0.73,1.06]	1.11 [0.97,1.27]	0.89 [0.7,1.14]
Q3 vs Q1						
Unadjusted	0.97 [0.95,0.99]	1.53 [1.28,1.81]	1.46 [1.38,1.54]	1.12 [0.93,1.34]	1.48 [1.3,1.7]	0.95 [0.75,1.2]
Age-sex adjusted	0.98 [0.96,1.01]	1.47 [1.24,1.74]	1.4 [1.32,1.48]	1.13 [0.94,1.36]	1.54 [1.35,1.76]	0.98 [0.78,1.24]
Fully adjusted	0.98 [0.96,1.01]	1.44 [1.21,1.71]	1.38 [1.31,1.47]	1.12 [0.93,1.34]	1.51 [1.32,1.72]	0.97 [0.77,1.23]
Q4 vs Q1						
Unadjusted	1.01 [0.99,1.04]	1.66 [1.4,1.97]	1.73 [1.63,1.83]	1.16 [0.97,1.39]	1.88 [1.65,2.14]	0.94 [0.74,1.18]
Age-sex adjusted	1.04 [1.02,1.07]	1.53 [1.29,1.81]	1.58 [1.49,1.68]	1.22 [1.02,1.46]	2.04 [1.79,2.32]	1.08 [0.85,1.36]
Fully adjusted	1.04 [1.02,1.07]	1.47 [1.24,1.75]	1.55 [1.46,1.64]	1.18 [0.99,1.41]	1.94 [1.71,2.22]	1.06 [0.84,1.34]
Q5 vs Q1						
Unadjusted	1.04 [1.01,1.07]	2.13 [1.81,2.52]	2.14 [2.02,2.26]	1.37 [1.15,1.63]	2.85 [2.51,3.22]	1.04 [0.84,1.29]
Age-sex adjusted	1.09 [1.06,1.12]	1.92 [1.62,2.27]	1.89 [1.79,2]	1.43 [1.21,1.71]	3.17 [2.8,3.6]	1.14 [0.92,1.42]
Fully adjusted	1.08 [1.05,1.11]	1.82 [1.53,2.15]	1.84 [1.74,1.95]	1.35 [1.13,1.61]	2.95 [2.6,3.35]	1.12 [0.9,1.4]

[a] Fully adjusted for age, sex and ethnic background.
[b] [] <- 95% Confidence intervals

The results in Table 3.5 ought to be interpreted in the same manner as the ethnicity-stratified hazard ratios, except that now they are all relative to the least deprived quintile.

There is a clear general trend of augmenting risk for more socio-economically deprived patients. It is most apparent in the risk of COVID-19 hospitalisation, where people who belong to the most deprived quintile are subject to almost three times the risk as those in Q1. Deprivation is also monotonically associated with testing (both negative and positive) and conversion (negative-positive). No association is observed between socio-economic status and mortality following hospital admission, although results do point towards increased risk for the most deprived quintile.

Figure 3.4 graphically displays the fully adjusted hazard ratios in the form of forest plots.

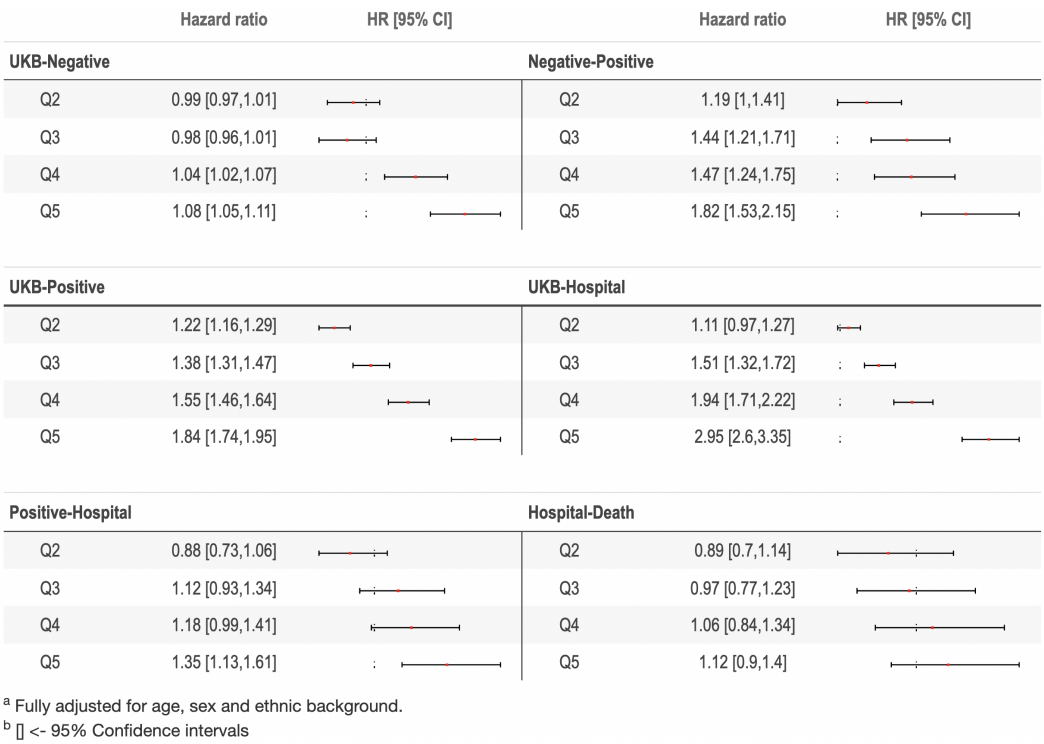


Fig. 3.4: Forest plots for socio-economic deprivation-stratified hazard ratios, shown fully adjusted[a,b]. The scale used for visualisation is linear but differs from transition to transition. The two dots reflect the $x = 1$ axis.

The same effect on confidence intervals for latter transitions appears as for the ethnicity stratification. In this case it is slightly less pronounced due to the larger group sizes when stratifying by socio-economic deprivation quintile.

3.4.4 Socio-economic deprivation: Poisson regression on testing

The same procedure follows for testing differences based on socio-economic deprivation strata, as Table 3.6 shows.

Testing appears to grow steadily more prevalent as socio-economic deprivation increases, with those in the most deprived quintile estimated to take more than 40% as many tests on average as those in Q1.

Tab. 3.6: Poisson regression coefficients on number of tests taken, stratified by socio-economic deprivation quintile and shown at different levels of confounding[a,b].

	Poisson			Quasi-Poisson		
	Estimate	S.E.	Risk ratio	Estimate	S.E.	Risk ratio
Q2 vs Q1						
Unadjusted	0.053	0.008	1.05 [1.04,1.07]	0.053	0.017	1.05 [1.02,1.09]
Age-sex adjusted	0.056	0.008	1.06 [1.04,1.07]	0.056	0.017	1.06 [1.02,1.09]
Fully adjusted	0.054	0.008	1.06 [1.04,1.07]	0.054	0.017	1.06 [1.02,1.09]
Q3 vs Q1						
Unadjusted	0.110	0.008	1.12 [1.1,1.13]	0.110	0.018	1.12 [1.08,1.16]
Age-sex adjusted	0.121	0.008	1.13 [1.11,1.15]	0.121	0.018	1.13 [1.09,1.17]
Fully adjusted	0.114	0.008	1.12 [1.1,1.14]	0.114	0.018	1.12 [1.08,1.16]
Q4 vs Q1						
Unadjusted	0.205	0.008	1.23 [1.21,1.25]	0.205	0.019	1.23 [1.18,1.27]
Age-sex adjusted	0.229	0.008	1.26 [1.24,1.28]	0.229	0.019	1.26 [1.21,1.3]
Fully adjusted	0.214	0.008	1.24 [1.22,1.26]	0.214	0.019	1.24 [1.19,1.28]
Q5 vs Q1						
Unadjusted	0.355	0.008	1.43 [1.4,1.45]	0.355	0.019	1.43 [1.37,1.48]
Age-sex adjusted	0.388	0.008	1.47 [1.45,1.5]	0.388	0.019	1.47 [1.42,1.53]
Fully adjusted	0.363	0.009	1.44 [1.41,1.46]	0.363	0.019	1.44 [1.38,1.49]

[a] Fully adjusted for age, sex and ethnic background.
[b] [] <- 95% Confidence intervals

3.5 Secondary results: sex & age

With the same methodology used to obtain the previous results, it is also possible to set sex or age as the variable of interest and the other variables as covariates to see their effect on COVID-19 hazard. Regression calculations are now remodelled to adjust partially only by age in the case of sex hazard ratios, and vice versa. Table 3.7 shows the Cox hazard ratio coefficients and Table 3.8 the Poisson risk ratios for testing differences.

We find a higher risk of infection for females, both following a negative test (conversion from negative to positive) and directly in their first test. Conversely, men display a higher probability of testing negative, and of severe forms of disease, with a higher risk of hospitalisation and death with COVID-19. Whilst older age is associated with negative testing and with severe disease (hospitalisation and mortality), younger

Tab. 3.7: Hazard ratio coefficients stratified by sex and age, at different levels of confounding[a,b]. Coefficients for sex are for males relative to females. The age coefficients reflects the change in hazard for every year older that a patient is.

	UKB-Negative	Negative-Positive	UKB-Positive	Positive-Hospital	UKB-Hospital	Hospital-Death
Male vs Female						
Unadjusted	1.02 [1.02,1.03]	0.95 [0.94,0.95]	0.94 [0.94,0.94]	1.06 [1.05,1.07]	1.07 [1.06,1.07]	1.08 [1.06,1.09]
Age adjusted	1.02 [1.02,1.03]	0.95 [0.94,0.95]	0.94 [0.94,0.94]	1.06 [1.05,1.07]	1.07 [1.06,1.07]	1.08 [1.06,1.09]
Fully adjusted	1.03 [1.02,1.03]	0.95 [0.95,0.96]	0.94 [0.94,0.95]	1.06 [1.05,1.07]	1.07 [1.07,1.08]	1.08 [1.06,1.09]
Age (+1 year)						
Unadjusted	1.02 [1.02,1.03]	0.95 [0.94,0.95]	0.94 [0.94,0.94]	1.06 [1.05,1.07]	1.07 [1.06,1.07]	1.08 [1.06,1.09]
Sex adjusted	1.02 [1.02,1.03]	0.95 [0.94,0.95]	0.94 [0.94,0.94]	1.06 [1.05,1.07]	1.07 [1.06,1.07]	1.08 [1.06,1.09]
Fully adjusted	1.03 [1.02,1.03]	0.95 [0.95,0.96]	0.94 [0.94,0.95]	1.06 [1.05,1.07]	1.07 [1.07,1.08]	1.08 [1.06,1.09]

[a] Fully adjusted for age, sex, ethnic background and socio-economic deprivation.
[b] [] <- 95% Confidence intervals

age is associated with a higher probability of testing positive. The effect of age was monotonic and quasi-linear in our models.

Tab. 3.8: Poisson regression coefficients on number of tests taken, stratified by sex and age, at different levels of confounding[a,b]. Sex coefficients are for males relative to females. The age coefficients reflects the change in hazard for every year older that a patient is.

	Poisson			Quasi-Poisson		
	Estimate	S.E.	Risk ratio	Estimate	S.E.	Risk ratio
Male vs Female						
Unadjusted	0.100	0.005	1.11 [1.09,1.12]	0.100	0.012	1.11 [1.08,1.13]
Age adjusted	0.096	0.005	1.1 [1.09,1.11]	0.096	0.012	1.1 [1.08,1.13]
Fully adjusted	0.091	0.005	1.1 [1.08,1.11]	0.091	0.012	1.1 [1.07,1.12]
Age (+1 year)						
Unadjusted	0.017	0.000	1.02 [1.02,1.02]	0.017	0.001	1.02 [1.02,1.02]
Sex adjusted	0.017	0.000	1.02 [1.02,1.02]	0.017	0.001	1.02 [1.02,1.02]
Fully adjusted	0.019	0.000	1.02 [1.02,1.02]	0.019	0.001	1.02 [1.02,1.02]

[a] Fully adjusted for age, sex, ethnic background socio-economic deprivation.
[b] [] <- 95% Confidence intervals

Poisson regression results confirm the finding that male sex, and older age are associated with a higher number of tests, in line with the results produced by the multi-state model.

Interestingly, the relationships observed remained after adjusting for confounders and did not vary in any significant way. This suggests that differences may be largely biological.

3.6 Other analyses: Kaplan-Meier plots

As discussed under the Methods section 2.7.4, Kaplan-Meier plots serve to graphically check the proportional hazards assumption, an essential requisite for the Cox proportional hazards model. If the assumption holds then all curves should be proportional to each other and not differ in form or intersect one another.

Stratifying by ethnic background, survival curves and their respective confidence intervals are shown in Figure 3.5 for the events of testing positive and hospitalisation *with* COVID-19. Below the two upper curves are the associated *log-log* plots of the survival function. All other risk events in the model have been studied but only these two are shown in this document to illustrate the usefulness of this methodology. Chinese ethnicity could not be displayed because the wide confidence intervals cluttered the figures and prevented the other curves from being displayed clearly.

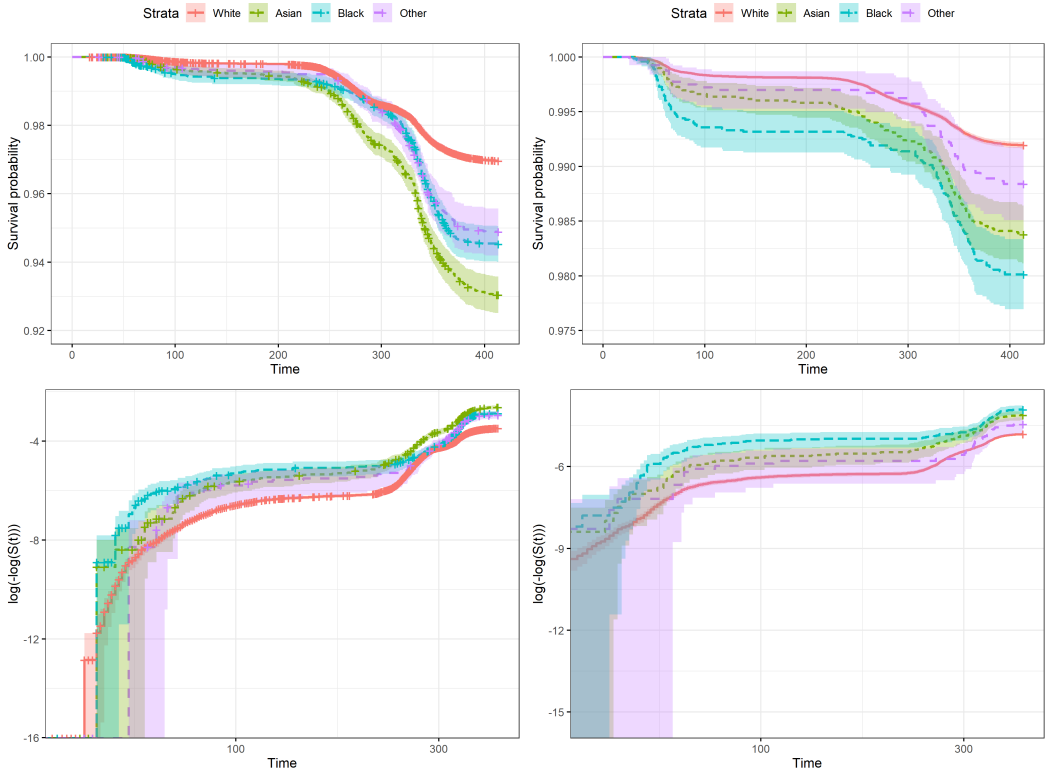


Fig. 3.5: Ethnicity-stratified Kaplan-Meier survival curves for the events of testing positive (left) and hospitalisation (right). The lower plots display the log-log curves of the survival functions above, with logarithmic axes. The x-axis displays time in days since the beginning of follow-up.

It is clear that there is no significant violation of the proportional hazards assumption for the three largest ethnic groups. Though the curves are not perfectly proportional, they all display similar behaviour and do not intersect each other. The curve for participants of *'other'* ethnicity does seem to cross that of Asians for positive testing and slightly touches that of Whites for hospitalisation. Confidence intervals complicate the evaluation of whether or not this is actually problematic for Cox regression. Furthermore, in the positive testing survival plot, the curve behaves correctly with respect to its reference curve, Whites. All in all, we consider it to be a very minor issue that does not warrant the need for a different parametric model.

Figure 3.6 shows the survival plots for socio-economic deprivation strata. In this case it is evident that the proportional hazards assumption holds very well. Narrower confidence intervals are obtained due to the larger n of these strata.

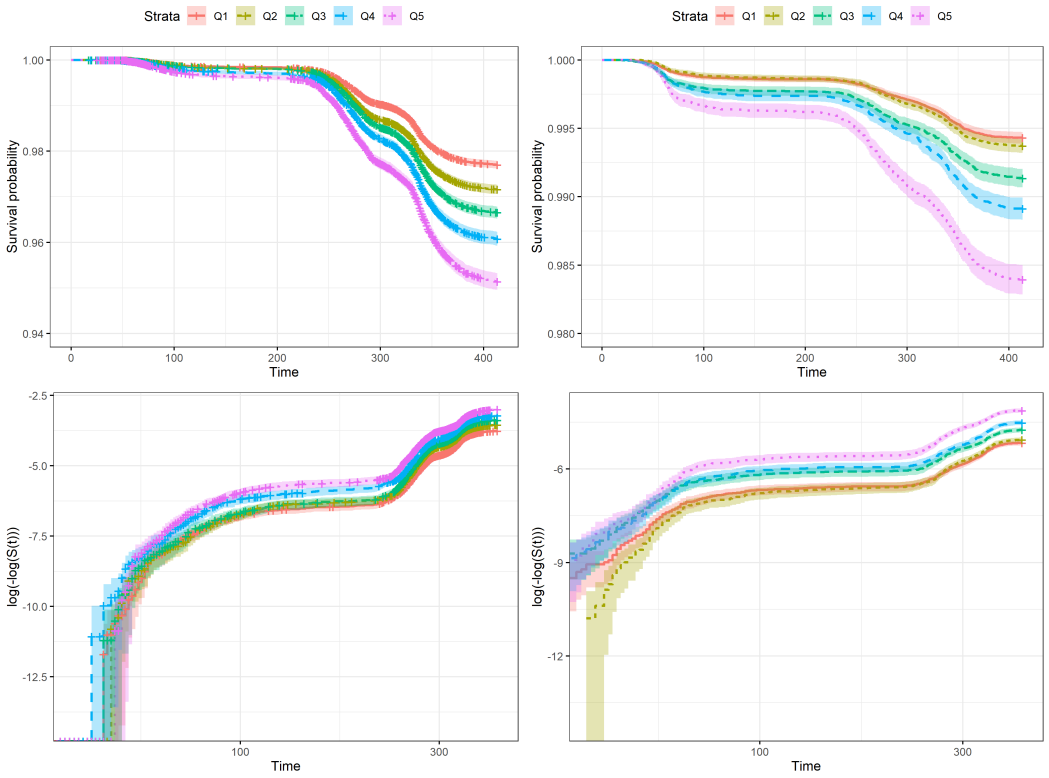


Fig. 3.6: Socio-economic deprivation-stratified Kaplan-Meier survival curves for the events of testing positive (left) and hospitalisation (right). The lower plots display the log-log curves of the survival functions above, with logarithmic axes. The x-axis displays the time in days since the beginning of follow-up.

4 Discussion & conclusions

4.1 Key results & conclusions

Referring back to the study's objectives outlined in Introduction 1.4, all have essentially been achieved.

In regards to the first objective, a database has been generated that contains the relevant information for eligible English patients of UK Biobank's cohort for a final population of 406,408. All patients have been monitored from the 1st February 2020 to the 20th March 2021, to include the better part of the observational window for COVID-19 health outcomes available so far. Combining a series of different sources, a cohesive single file has been created with the dataframe necessary for all statistical analysis. Furthermore, the complete set of codes has been streamlined to allow for automatic updates, as UKB and HES incorporate new data into their system.
The discussion that follows covers the work carried out for the other objectives.

4.1.1 Population characteristics

Population characteristics have been extensively examined and quantified. We find a large White majority, about ∼94.1% of the cohort. The next two largest groups are Asian (∼2.2%) and Black (∼1.8%), followed by *'other'* (∼1.0%), Mixed (∼0.6%) and Chinese (∼0.3%). In terms of socio-economic status, there is an overall tendency for the less deprived quintiles, with about 29.7% of the cohort belonging to the least deprived quintile and only 12.8% in the most. It is important to note, however, that there is a stark difference between the entire cohort's socio-economic distribution and that of ethnic minorities, as this phenomenon seems to only affect White and Chinese individuals. The largest inequality observed in the cohort is that of Blacks when compared with Whites. About 43.0% of Black individuals reside in areas within the most deprived quintile, while only 5.3% are in the top quintile. Such drastic differences may condition analysis, even when confounders are minimised, which will be discussed in the following sections.

In terms of age, we find that the cohort is generally of advanced age. This occurs due to UK Biobank's design, as this was their target population at the time of recruitment. The median age is 70.25, with the interquartile range extending from 63 to 76 years of age. The youngest registered user is 51 and the oldest 87. Again, the overall population characteristics are heavily conditioned by those of Whites, since they make up the large majority of the cohort. This is particularly apparent when measuring age too. All other ethnic groups have a median age at least 5 years younger than Whites, with the youngest overall being Mixed (62.5 median, IQR=[57.3,70.2]) and Black individuals (62.8 median, IQR=[57.7,69.8]).

The cohort is generally majority female (∼55.2%), where only Asian male individuals outnumber their female counterparts (47.2%). The largest difference appears for the Chinese ethnic group, with about 64.7% of all patients being female.

4.1.2 Health outcomes

We find the most transitions at the beginning stages of the multi-state COVID-19 model developed, in the Negative and Positive states. Prior to modelling, there is already evidence that ethnic minorities are more likely than Whites to be infected at some point. The most notable difference is observed in Asians having a positivity rate of 6.3% compared to the cohort-wide 2.9%. Chinese patients are the only group with a lower rate (1.9%). The same phenomenon is clearly observed for the more deprived quintiles of the population, where there is a steady rise in the percentage of infection as socio-economic deprivation increases. Overall, 2.9% of the cohort has tested positive for COVID-19 during the time of follow-up.
The percentage for testing negative is 14% cohort-wide, and there is no clear pattern of differences between ethnic or socio-economic strata, though White individuals have a slightly higher rate.

In terms of hospital admissions, the majority of episodes occur with a positive test prior to the date of admission. Here we find that about 9.3% of patients with a registered positive test end up hospitalised. A further 0.9% of those who enter the hospital on the same day they are tested, and Black individuals appear to suffer the greatest amount of these extreme scenarios (1.3%), as evidenced by the UKB-Hospital transition. Asian, Black and Chinese patients have higher percentages of hospitalisation than Whites, with the largest difference being 4 percentage points higher for Blacks. Those of Mixed and *'other'* ethnicity have lower hospitalisation rates.
The adverse effect of lower socio-economic status is again immediately apparent.

The most deprived quintile of the population has a rate of hospitalisation of 11.5%, and the top quintile 8.4%. The same occurs for same-day positive tests and hospital admission. Interestingly, hospitalisation is least frequent among those in the second least deprived quintile.

Of those patients who do suffer COVID-19 hospital episodes, 21.1% die from it. There does not seem to be any trend of ethnic or socio-economic differences in COVID-19 mortality, except that Whites have the highest at 24.1%. The second least deprived quintile of the population appears again to be the least affected, with 18% mortality. Viewing these results, it is also important to consider the aforementioned differences in the probability of being hospitalised in the first place.

There are no noticeable differences in testing depending on sex. Nevertheless, males have a higher hospitalisation rate (11.6% vs 7.3%) and mortality (23.6% vs 17.7). Patients older than the median age (70.25) do not have a higher hospitalisation rate but do display much higher mortality than those younger than the median (16.8% vs 4.2%).

It is important to note that the percentages discussed above correspond to completely unadjusted calculations, as they are merely based on the percentage of individuals who undergo a given transition with respect to the total group who were susceptible to it.

4.1.3 Ethnic differences in relative risk

Cox regression analysis provides hazard ratios for relative risk assessment. This methodology allows for confounder minimisation, which is why they are the central piece of this thesis' results. The third and fourth objectives of the project revolve around creating the necessary structures to obtain these estimates, and they both have been achieved.

When compared with White patients, the fully adjusted coefficients for the UKB-Positive transition display increased risk of infection for Asian 1.72 [1.58,1.88], Black 1.1 [0.99,1.23] and *'other'* 1.09 [0.93,1.26] ethnicity patients (in brackets are the 95% confidence intervals). Chinese individuals have a significantly decreased risk of infection 0.52 [0.35,0.77] and Mixed ethnicity individuals have no substantial difference when compared with Whites.
The Negative-Positive transition, for those who have tested negative before they test positive, displays even greater risk for Blacks (1.39 [1.04,1.87]) and Mixed (1.4 [0.84,2.34]) individuals than the UKB-Positive transition. *'Other'* ethnicity

participants have a coefficient of 2 [1.42,2.82], much higher than for the previous transition. All in all, there is an increased risk of infection for ethnic minorities all across the board, with the exception of Chinese individuals; and Asians seem to suffer the greatest risk.
As for the UKB-Negative transition, there is a slight variation with respect to Whites, most notably Blacks having a coefficient of 1.07 [1,1.14] and Mixed ethnicity patients 1.09 [0.98,1.22]. Chinese individuals are again the only group to display severely reduced risk, (0.64 [0.53,0.77]). These results point to Chinese patients generally taking fewer COVID-19 tests on average.

Hospitalisation coefficients present much more drastic differences. Blacks patients are the worst off overall, with a coefficient of 1.8 [1.35,2.41] for the Positive-Hospital transition and 2.36 [1.91,2.92] for the more extreme UKB-Hospital transition where testing occurs on the same day as hospital admission. There is also increased risk for Asians: 1.32 [1.02,1.71] and 1.79 [1.43,2.24] (for Positive-Hospital and UKB-Hospital respectively). The confidence intervals widen greatly for the other ethnic groups, due to the lower number of individuals undergoing these transitions. Nevertheless, all of them have increased risk: 2.65 [0.85,8.24] and 1.66 [0.86,3.19] for Chinese patients; 1.18 [0.61,2.27] and 1.24 [0.72,2.14] for Mixed; 1.25 [0.77,2.03] and 1.5 [1.03,2.16] for *'other'* ethnicity patients.
Altogether, there is clear evidence of much increased hazard of hospitalisation for ethnic minorities when compared with Whites, in some cases even doubling it. It is important to state once more that all of the results mentioned above correspond to calculations fully adjusted for socio-economic deprivation, age and sex, which only highlights the gravity of the situation.

Results on mortality hazard ratios reveal a different trend. Only Asians and Blacks appear to be under increased risk, with coefficients of 1.33 [0.91,1.94] and 1.08 [0.74,1.59] respectively. All other groups have significantly decreased risk of dying with COVID-19 when compared with Whites. Chinese patients have 0.58 [0.15,2.34], and *'other'* ethnicity patients have 0.62 [0.26,1.51]. Mixed individuals appear to have the lowest hazard of death at 0.55 [0.18,1.7]. Note that confidence intervals are quite wide for this final transition, which makes drawing conclusions from the results difficult.

4.1.4 Socio-economic differences in relative risk

The following is a discussion of the results on hazard ratios for socio-economic strata obtained from calculations fully adjusted for ethnicity, age and sex. We find a directly

proportional relationship between socio-economic deprivation and COVID-19 hazard that is steady for most coefficients, which is why discussion will mostly cover the results obtained for the most deprived quintile of the population.

There is severely increased risk across all transitions for the most deprived quintile with respect to the least. The most drastic results are observed for hospitalisation: the UKB-Hospital transition returns a hazard rate of 2.95 [2.6,3.35], and Positive-Hospital 1.35 [1.13,1.61]. In other words, patients from the most deprived neighbourhoods are up to three times as likely to enter the hospital on the same day they test positive. The risk of infection is also greatly increased, with the UKB-Positive returning a coefficient of 1.84 [1.74,1.95] and the Negative-Positive transition 1.82 [1.53,2.15]. The risk of testing negative is slightly higher too, at 1.08 [1.05,1.11]. This suggests that poorer individuals generally take more tests but also have a greater positivity rate when they do. Finally, the risk of dying with COVID-19 reflects an increased of 1.12 [0.9,1.4].

An additional observation that may be important to point out is that, what was observed early on in the health outcomes phase of the study in regard to the second most privileged quintile, still stands after proper analysis and confounder adjustment. We find that the second quintile has a reduced risk of hospitalisation 0.88 [0.73,1.06] and death 0.89 [0.7,1.14].

Sex & age differences in relative risk. Our results point to increased hazard for males when it comes to hospitalisation 1.06 [1.05,1.07] and death 1.08 [1.06,1.09]. The risk of infection appears to be lower than that of females, with both the Negative-Positive and UKB-Positive transitions returning coefficients of 0.95 [0.95,0.96] and 0.94 [0.94,0.95], respectively. There is however a slight increase in the risk of testing negative (1.03 [1.02,1.03]). As for age differences, our model estimates that the effect of being one year older is an increased relative risk of 1.08 [1.06,1.09] for mortality and 1.06 [1.05,1.07] for hospitalisation. The risk of infection does appear to decrease with age. In conclusion, our results back up the now well-established notion that male sex and age are adverse predictors of COVID-19 health outcomes. It is interesting to note that these relationships are largely unaffected by confounder adjustment, which may be indicative of mostly biological causes.

4.1.5 Ethnic differences in testing prevalence

The final objective of this thesis was to model and analyse group differences in testing prevalence using an specially adapted methodology. This has been carried

out using Poisson regression count analysis, which allows for the obtention of relative risk ratios. These are to be interpreted in a similar manner as with the previous Cox coefficients, and they are also provided alongside their 95% confidence intervals, but in this scenario they are to be understood more like average differences in tests per person.

Compared with Whites, the fully adjusted Poisson coefficients return high risk ratios for all ethnic groups except Chinese (0.66 [0.51,0.86]). The most notable differences appear for Black (1.32 [1.23,1.43]) and Asian (1.26 [1.17,1.36]) patients. This implies that Chinese participants on average take about 34% less tests than Whites, and that Blacks and Asians take 32% and 26% more. These results back up the previous finding that Chinese individuals take less tests in general.

4.1.6 Socio-economic differences in testing prevalence

As with the Cox hazard ratios for socio-economic deprivation strata, we find a steady rise in testing prevalence for the more deprived quintiles. The coefficient for the most economically deprived with respect to the least is 1.44 [1.38,1.49]. This implies that the poorest section of the population takes on average more than 40% COVID-19 tests as the richest.

It is interesting to note that the Poisson risk ratios obtained do not seem to be altered very much by the level of confounder adjustment. This suggests that testing prevalence depends very strongly on socio-economic status, and not so much on ethnicity, sex or age.

Sex & age differences in testing prevalence. Our results point to an approximate increase of 10% in the amount of tests taken by males when compared with females (1.1 [1.07,1.12]). In the case of age, the estimation obtained is that, for every year older a patient is, testing prevalence increases by 2% (1.02 [1.02,1.02]). Again, these results do not vary depending on confounder adjustment.

4.1.7 Conclusions

Our study results reveal severely increased risk of infection for Asians, who also taken substantially more tests on average and still display a higher positivity rate than Whites. The same is observed for Black, Mixed and *'other'* ethnicity patients. Chinese individuals are the only group with reduced risk and less tests on average

than Whites.
In terms of hospitalisation, increased risk is found across all ethnic groups, and Blacks are the worst-off overall with up to twice the hazard for Whites, followed by Asians. The greatest difference is found for the scenario where an individual undergoes hospital admission on the same day they test positive.
Mortality results indicate increased risk for Asians, most notably, and Blacks. The other ethnicities display reduced risk when compared with Whites.

Regression coefficients for socio-economic status display a steadily proportional relationship between deprivation and relative hazard. Increased risk is found across all health outcomes, with same-day hospitalisation appearing up to three times as likely for the most deprived section of the population when compared with the least. Furthermore, there is severely increased risk of infection for the poorer quintiles, as well as many more tests taken on average. Mortality also increases with socio-economic deprivation, and the risk is estimated to be up to 10% higher for the bottom quintile.

Lastly, our results establish male sex and age as adverse predictors of hospitalisation and mortality. There is also indication of higher tests taken on average, though females appear to have a greater risk of infection.

4.2 Limitations, generalisability and strengths

There is a number of limitations, present by design in the study, that ought to be discussed as they are likely to affect the generalisabilty of the results obtained.

4.2.1 Socio-economic bias

There is considerable bias present in the population that makes up UK Biobank's study. It has been well documented that users who signed up for the program generally belong to noticeably privileged socio-economic backgrounds, when compared to the wider population. Figure 3.2 directly displays this phenomenon. The deprivation distribution is not heterogeneous however. We find that White and Chinese individuals in the cohort are notably more privileged than the other ethnic groups, and that Blacks are the worst-off overall. A 2017 study by the University of Oxford concluded that as a result of this socio-economic bias, UKB participants were non-representative of the wider population [51]. They found that they were less likely to be obese, smoke or drink alcohol on a daily basis, and that cancer was

much less prevalent than in the national average. Altogether, it has to be taken into consideration that participants of the study are generally healthier than the wider population, which may limit the result's generalisability.

4.2.2 Age limitations

The other form of bias present in the sample comes from the fact that the cohort is of advanced age (all patients are at least fifty years old). This occurs as a result of UK Biobank's design, because this was their target population at the time of recruitment. Therefore, it must be taken into consideration that all the results for COVID-19 health outcomes presented concern the most vulnerable sectors of the population. Where this is likely to have the largest effect is in findings of hospitalisation and mortality. Generalisation to other age groups may be limited as a result. Nevertheless, serious course of the COVID-19 disease affects those patients above 60 especially. Therefore, the biased age of the cohort may actually provide valuable knowledge about hospitalisation and deaths.

An additional factor that possibly limits generalisability is the geographical limitation, as this study only covers individuals living in England. While results are certainly indicative of what one could expect from similar cohort studies in other Western countries like the United States, especially for findings on socio-economic, age and sex differences, proper analysis of each individual situation would be required before drawing any conclusions.
Furthermore, the presence of large aggregations of people in cities like London, Birmingham or Manchester, may condition results. It is now well-known that London has been a focal point of the COVID-19 pandemic, which may explain in part what has been observed in this study for the second least deprived quintile of the population with respect to the first.

The data and statistical methodology employed offer a number of great strengths too. UK Biobank provides the possibility of studying such a large and varied cohort and all of their population characteristics, which enables the obtention of powerful results like the ones shown above. It also offers the advantage of circumventing collider bias, as discussed in Introduction 1.1.4. Furthermore, there is a chronological advantage to having performed the study at this particular time (February to June 2021), as the observation window for COVID-19 health outcomes is greater now than it has ever been since the pandemic began. Cox and Poisson regression analysis also allow for confounder minimisation, which is necessary to properly analyse the effect of ethnicity and socio-economic deprivation on risk.

4.3 Interpretation & future work

Explaining the severity of the results observed for ethnic minorities, even when adjusting for socio-economic deprivation, age and sex, is complicated. Needless to say, it falls beyond the objectives and scope of this project, and much more bibliographic work and research would be required to draw any solid conclusions. Nevertheless, providing some speculation into the causes behind the conditions observed may attract further research into the topic.

There is a variety of factors that may be at play, and a correct explanation is likely to be based on a combination of them. Cultural norms may affect an individuals risk of infection greatly [52]. Differences in the number of people living in the same household, routine habits like eating together or not, and the overall approach to COVID-19 safety measures can vary between different cultural subgroups. It could also partly explain why Chinese patients have a much reduced risk of infection, since the SARS outbreak of 2002-2004 was a major epidemic that took its biggest toll on China. It is likely that it left its mark on the collective consciousness and that cultural norms with regards to epidemic safety reflect it in some way.

There is of course the possibility of institutional neglect and discrimination on the part of official institutions and society at large [4]. Furthermore, past history of racism and abuse may condition an individual's trust on the medical system [53], which could reflect adverse attitudes in regards to COVID-19 testing and early hospital admission for patients suffering from the disease.

Lastly, it is possible that genetics play a significant role in COVID-19 susceptibility and severity and that variability between ethnic groups partly explains the results obtained. This, like all other speculation provided in this section, requires careful research, in order to most effectively protect the population in future epidemics.

In regards to the last two points, our results can be interpreted to provide some indication into which factor has more relevance. We find that, once in the hospital, mortality does not vary greatly for ethnic minorities. This is heavily conditioned by the previous transitions, which reflect an increase in risk. All in all, this suggests that differences are not so much biological, but more to do with exposure (higher risk of infection) and with issues of health equity in early access to medical care. Furthermore, Black patients in the cohort have similar risk of infection to Whites, but much higher risk of direct diagnosis at the time of hospital admission. This could again be indicative of reduced access to health services.

When it comes to the drastic socio-economic differences in risk observed, the most direct cause is likely to be greater workplace exposure. We find a considerable increase in the number of tests taken by the poorer quintiles, which probably indicates that their professions require face to face interaction, and therefore regular testing. It is to be expected that those in the least deprived quintiles found less trouble adapting to COVID-19 restrictions, as their jobs could be more easily adapted to online work from home. Besides an increased risk of infection, poorer individuals also face greater risk of hospitalisation and mortality due to the severe adverse effect of socio-economic deprivation on health outcomes in general.

In any case, the conclusion that should be drawn from the study's results is that medical institutions may need to put more emphasis on protecting ethnic minorities from viral epidemics of this kind in the future, as they appear to be at greater risk. The same is clearly true for socio-economically deprived individuals, who have also been under much increased risk throughout the COVID-19 pandemic.

4.3.1 Future work

All of the results and notions discussed above are incomplete without proper adjustment of the residual confounding caused by comorbidities. The greatest future effort will go towards expanding our dataset and building behaviour and comorbidity population characteristics into it. Relevant diseases when it comes to COVID-19 hazard are leukemia, diabetes, morbid obesity and a range of others that will be accounted for in the Cox and Poisson regression coefficients. Smoking, alcohol consumption and body-mass index also reveal important health habits, and they ought to be included too.

Although non-linear dependencies between hazard and age were experimented with, they were deemed a very minor improvement to the model. We found that for all transitions the relationship was best fitted linearly or very close to linear, so it was deemed unnecessary to implement it for the time being. This occurs due to the particular age distribution of the cohort. Nevertheless, more research may be directed at this topic in the future to assess if implementing it would be lead to more accurate results.

Throughout the month of July 2021, the team will continue perfecting and updating the model, especially as more data is incorporated into our sources and it becomes possible to advance the right-censoring date. All codes developed for this project will be made publicly accessible through a GitHub repository once it is finalised. Ultimately, we hope to publish this thesis' findings in a scientific journal.

Bibliography

[1] Kristian G Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, and Robert F Garry. "The proximal origin of SARS-CoV-2". In: *Nature medicine* 26.4 (2020), pp. 450–452 (cit. on p. 1).

[2] Patrick J Lillie, Anda Samson, Ang Li, et al. "Novel coronavirus disease (Covid-19): the first two patients in the UK with person to person transmission". In: *The Journal of infection* 80.5 (2020), p. 578 (cit. on p. 1).

[4] Delan Devakumar, Geordan Shannon, Sunil S Bhopal, and Ibrahim Abubakar. "Racism and discrimination in COVID-19 responses". In: *The Lancet* 395.10231 (2020), p. 1194 (cit. on pp. 1, 2, 49).

[5] Rachel E Jordan, Peymane Adab, and KK32217618 Cheng. *Covid-19: risk factors for severe disease and death*. 2020 (cit. on p. 1).

[6] Sonja S Hutchins, Kevin Fiscella, Robert S Levine, Danielle C Ompad, and Marian McDonald. "Protection of racial/ethnic minority populations during an influenza pandemic". In: *American journal of public health* 99.S2 (2009), S261–S270 (cit. on p. 1).

[7] Kamlesh Khunti, Awadhesh Kumar Singh, Manish Pareek, and Wasim Hanif. *Is ethnicity linked to incidence or outcomes of covid-19?* 2020 (cit. on p. 2).

[8] Kamaldeep Bhui. "Ethnic inequalities in health: The interplay of racism and COVID-19 in syndemics". In: *EClinicalMedicine* 36 (2021) (cit. on p. 2).

[9] Annette Flanagin, Tracy Frey, Stacy L Christiansen, and Howard Bauchner. "The reporting of race and ethnicity in medical and science journals: comments invited". In: *JAMA* 325.11 (2021), pp. 1049–1052 (cit. on p. 2).

[11] Cathie Sudlow, John Gallacher, Naomi Allen, et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *Plos med* 12.3 (2015), e1001779 (cit. on pp. 2, 9).

[12] Claire L Niedzwiedz, Catherine A O'Donnell, Bhautesh Dinesh Jani, et al. "Ethnic and socioeconomic differences in SARS-CoV-2 infection: prospective cohort study using UK Biobank". In: *BMC medicine* 18 (2020), pp. 1–14 (cit. on p. 2).

[13] Albert Prats-Uribe, Roger Paredes, and Daniel Prieto-Alhambra. "Ethnicity, comorbidity, socioeconomic status, and their associations with COVID-19 infection in England: a cohort analysis of UK Biobank data". In: *medRxiv* (2020) (cit. on p. 2).

[14] Gareth J Griffith, Tim T Morris, Matthew J Tudball, et al. "Collider bias undermines our understanding of COVID-19 disease risk and severity". In: *Nature communications* 11.1 (2020), pp. 1–12 (cit. on p. 3).

[15] Olivier Vandenberg, Delphine Martiny, Olivier Rochas, Alex van Belkum, and Zisis Kozlakidis. “Considerations for diagnostic COVID-19 tests”. In: *Nature Reviews Microbiology* 19.3 (2021), pp. 171–183 (cit. on p. 3).

[16] Steven Riley, Haowei Wang, Oliver Eales, et al. “REACT-1 round 9 final report: Continued but slowing decline of prevalence of SARS-CoV-2 during national lockdown in England in February 2021”. In: *MedRxiv* (2021) (cit. on pp. 3, 10).

[17] Tyler J VanderWeele and Ilya Shpitser. “On the definition of a confounder”. In: *Annals of statistics* 41.1 (2013), p. 196 (cit. on p. 4).

[18] Ross L Prentice, John D Kalbfleisch, Arthur V Peterson Jr, et al. “The analysis of failure times in the presence of competing risks”. In: *Biometrics* (1978), pp. 541–554 (cit. on p. 5).

[19] Hein Putter, Marta Fiocco, and Ronald B Geskus. “Tutorial in biostatistics: competing risks and multi-state models”. In: *Statistics in medicine* 26.11 (2007), pp. 2389–2430 (cit. on pp. 6, 21).

[20] Per Kragh Andersen and Niels Keiding. “Multi-state models for event history analysis”. In: *Statistical methods in medical research* 11.2 (2002), pp. 91–115 (cit. on p. 6).

[21] Morten Frydenberg. “The chain graph Markov property”. In: *Scandinavian Journal of Statistics* (1990), pp. 333–353 (cit. on p. 6).

[22] Syrene A Miller and Jane L Forrest. “Enhancing your practice through evidence-based decision making: PICO, learning how to ask good questions”. In: *Journal of Evidence Based Dental Practice* 1.2 (2001), pp. 136–141 (cit. on p. 7).

[23] Jan P Vandenbroucke, Erik Von Elm, Douglas G Altman, et al. “Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration”. In: *PLoS Med* 4.10 (2007), e297 (cit. on p. 8).

[24] Thirumalaisamy P Velavan and Christian G Meyer. “The COVID-19 epidemic”. In: *Tropical medicine & international health* 25.3 (2020), p. 278 (cit. on p. 9).

[26] Jacob Armstrong, Justine K Rudkin, Naomi Allen, et al. “Dynamic linkage of covid-19 test results between public health england’s second generation surveillance system and UK Biobank”. In: *Microbial genomics* 6.7 (2020) (cit. on p. 10).

[28] David McLennan, Stefan Noble, Michael Noble, et al. “The English Indices of Deprivation 2019: technical report”. In: (2019) (cit. on pp. 11, 14).

[31] Hude Quan, Vijaya Sundararajan, Patricia Halfon, et al. “Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data”. In: *Medical care* (2005), pp. 1130–1139 (cit. on p. 16).

[32] David R Cox. “Regression models and life-tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202 (cit. on p. 18).

[33] Kenneth R Hess. “Graphical methods for assessing violations of the proportional hazards assumption in Cox regression”. In: *Statistics in medicine* 14.15 (1995), pp. 1707–1723 (cit. on p. 19).

[34]Stefany Coxe, Stephen G West, and Leona S Aiken. “The analysis of count data: A gentle introduction to Poisson regression and its alternatives”. In: *Journal of personality assessment* 91.2 (2009), pp. 121–136 (cit. on p. 21).

[35]Jay M Ver Hoef and Peter L Boveng. “Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?” In: *Ecology* 88.11 (2007), pp. 2766–2772 (cit. on p. 22).

[36]Edward L Kaplan and Paul Meier. “Nonparametric estimation from incomplete observations”. In: *Journal of the American statistical association* 53.282 (1958), pp. 457–481 (cit. on p. 22).

[41]Hein Putter. “Tutorial in biostatistics: Competing risks and multi-state models Analyses using the mstate package”. In: *Companion file for the mstate package* (2011) (cit. on p. 24).

[51]Anna Fry, Thomas J Littlejohns, Cathie Sudlow, et al. “Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population”. In: *American journal of epidemiology* 186.9 (2017), pp. 1026–1034 (cit. on p. 47).

[52]Máté Kapitány-Fövény and Mihály Sulyok. “Social markers of a pandemic: modeling the association between cultural norms and COVID-19 spread data”. In: *Humanities and Social Sciences Communications* 7.1 (2020), pp. 1–9 (cit. on p. 49).

[53]Jason Schnittker and Mehul Bhatt. “The role of income and race/ethnicity in experiences with medical care in the United States and United Kingdom”. In: *International Journal of Health Services* 38.4 (2008), pp. 671–695 (cit. on p. 49).

Webpages

[@3]World Health Organisation. *Coronavirus disease (COVID-19) pandemic*. 2021. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (cit. on p. 1).

[@10]*UK Biobank*. 2021. URL: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank (cit. on pp. 2, 9).

[@25]*NHS Hospital Episode Statistics (HES)*. 2021. URL: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics (cit. on pp. 9, 10).

[@27]*Demographic data for coronavirus (COVID-19) testing (England): 28 May to 26 August*. 2020. URL: https://www.gov.uk/government/publications/demographic-data-for-coronavirus-testing-england-28-may-to-26-august/demographic-data-for-coronavirus-covid-19-testing-england-28-may-to-26-august (cit. on p. 10).

[@29]*Ethnicity Harmonised Standard*. 2020. URL: https://gss.civilservice.gov.uk/policy-store/ethnicity/#presentation-england-and-wales (cit. on p. 12).

[@30]*English indices of deprivation*. 2020. URL: https://www.gov.uk/government/collections/english-indices-of-deprivation (cit. on p. 14).

[@37]R Core Team et al. *R: A language and environment for statistical computing*. 2013. URL: https://cran.r-project.org/ (cit. on p. 23).

[@38]Hadley Wickham, Mara Averick, Jennifer Bryan, et al. *Welcome to the tidyverse*. 2019. URL: https://CRAN.R-project.org/package=tidyverse (cit. on p. 24).

[@39]J Allaire. *RStudio: integrated development environment for R*. 2012. URL: https://www.rstudio.com/ (cit. on p. 24).

[@40]Liesbeth C. de Wreede, Marta Fiocco, and Hein Putter. *mstate: An R Package for the Analysis of Competing Risks and Multi-State Models*. 2011. URL: https://www.jstatsoft.org/v38/i07/ (cit. on p. 24).

[@42]Terry M Therneau. *survival: A Package for Survival Analysis in R*. R package version 3.2-10. 2021. URL: https://CRAN.R-project.org/package=survival (cit. on p. 24).

[@43]Marcel Wiesweg. *survivalAnalysis: High-Level Interface for Survival Analysis and Associated Plots*. R package version 0.2.0. 2021. URL: https://CRAN.R-project.org/package=survivalAnalysis (cit. on p. 24).

[@44]Original by Gareth Ambler and modified by Axel Benner. *mfp: Multivariable Fractional Polynomials*. R package version 1.5.2. 2015. URL: https://CRAN.R-project.org/package=mfp (cit. on p. 24).

[@45]Garrett Grolemund and Hadley Wickham. *Dates and Times Made Easy with lubridate*. 2011. URL: https://www.jstatsoft.org/v40/i03/ (cit. on p. 24).

[@46]Nicholas Cooper. *reader: Suite of Functions to Flexibly Read Data from Files*. R package version 1.0.6. 2017. URL: https://CRAN.R-project.org/package=reader (cit. on p. 24).

[@47]Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*. R package version 1.14.0. 2021. URL: https://CRAN.R-project.org/package=data.table (cit. on p. 24).

[@48]Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.4. 2021. URL: https://CRAN.R-project.org/package=kableExtra (cit. on p. 24).

[@49]Alboukadel Kassambara, Marcin Kosinski, and Przemyslaw Biecek. *survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.9. 2021. URL: https://CRAN.R-project.org/package=survminer (cit. on p. 24).

[@50]Tyler W. Rinker and Dason Kurkiewicz. *pacman: Package Management for R*. version 0.5.0. 2018. URL: http://github.com/trinker/pacman (cit. on p. 24).

List of Figures

List of Tables