Clustering of patient comorbidities within electronic medical records enables high-precision COVID-19 mortality prediction

Erwann Le Lannou [1], Benjamin Post MD [2, 3], Shlomi Haar [1, 4, 6], Stephen J. Brett MD [5], Balasundaram Kadirvelu [1,2] & A. Aldo Faisal [1, 2, 3, 6]

1: Brain & Behaviour Lab: Department of Bioengineering, Imperial College London, London, UK
2: Brain & Behaviour Lab: Department of Computing, Imperial College London, London, UK
3: UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, London, UK
4: Department of Brain Sciences, Imperial College London, London, UK
5: Department of Surgery and Cancer, Imperial College, London, UK
6: Behaviour Analytics Lab, Data Science Institute, London, UK

Acknowledgements
We are grateful to UK Biobank participants. This research has been conducted using the UK Biobank Resource under Application Number 21770. Infrastructure support for this research was provided by the Imperial NIHR Biomedical Research Centre. AAF acknowledges his UKRI Turing AI Fellowship (EP/V025449/1). The acknowledged parties and funders had no role in the design of the study.

Author Contributions
Conception and design: ELL, SH, AAF; data analysis: ELL, SH, BK, AAF; data interpretation: ELL, BP, SH, SJB, AAF; writing and revising the paper: ELL, BP, SH, SJB, BK, AAF

Conflict of Interest: The authors have declared that no competing interests exist.

Corresponding author: aldo.faisal@imperial.ac.uk (AAF)

Abstract:

We present an explainable AI framework to predict mortality after a positive COVID-19 diagnosis based solely on data routinely collected in electronic healthcare records (EHRs) obtained prior to diagnosis. We grounded our analysis on the ½ Million people UK Biobank and linked NHS COVID-19 records. We developed a method to capture the complexities and large variety of clinical codes present in EHRs, and we show that these have a larger impact on risk than all other patient data but age. We use a form of clustering for natural language processing of the clinical codes, specifically,

topic modelling by Latent Dirichlet Allocation (LDA), to generate a succinct digital fingerprint of a patient's full secondary care clinical history, i.e. their comorbidities and past interventions. These digital comorbidity fingerprints offer immediately interpretable clinical descriptions that are meaningful, e.g. grouping cardiovascular disorders with common risk factors but also novel groupings that are not obvious. The comorbidity fingerprints differ in both their breadth and depth from existing observational disease associations in the COVID-19 literature. Taking this data-driven approach allows us to avoid human-induction bias and confirmation bias during selection of what are important potential predictors of COVID-19 mortality. Together with age, these digital fingerprints are the single most important factor in our predictor. This holds the potential for improving individual risk profiling for clinical decisions and the identification of groups for public health interventions such as vaccine programmes. Combining our digital precondition fingerprints with demographic characteristics allow us to match or exceed the performance of existing state-of-the-art COVID-19 mortality predictors (EHCF) which have been developed through expert consensus. Our precondition fingerprinting and entire mortality prediction analytics pipeline are designed so as to be rapidly redeployable, e.g. for COVID-19 variants or other pre-existing diseases.

Introduction

The outbreak of the novel, severe and acute respiratory syndrome, Coronavirus 2 (SARS-CoV-2), and its associated disease COVID-19, in Wuhan China, has presented an important and urgent threat to global health since December 2019. Declared a pandemic by the World Health Organisation (WHO) in early March[1], this disease has, as of 8 March 2021, caused 2,582,528 deaths from 116,166,652 cumulative worldwide cases[2]. On the same day in the United Kingdom (UK) SARS-CoV-2 is reported as responsible for 124,419 deaths[2]. The COVID-19 outbreak has led to an increase in hospital admissions and has considerably increased the demand for both general hospital and critical care beds[3,4]. This increases the need for an objective, data-driven understanding of risk factors and accurate prediction of adverse outcomes based on pre-existing information. Such understanding may facilitate better individual clinical decisions, but potentially also increase the resilience of public health policy.

Emerging evidence throughout the pandemic has identified several risk factors associated with COVID-19 clinical severity[5] and fatality[6,7]. Most commonly reported risk factors are old age[8–10], male gender[11,12] and pre-existing underlying medical conditions[6]. In particular, the Chinese Centre for Disease Control and Prevention identified cardiovascular disease, hypertension, diabetes, respiratory disease, and numerous malignancies to be risk factors for COVID-19 fatality[12]. In the United States, the most commonly initially reported comorbidities were hypertension, obesity and diabetes[9]. Finally, a UK-wide study aiming to characterise the clinical features of patients with severe COVID-19 highlighted chronic cardiac disease, diabetes, and chronic pulmonary disease to be associated with a higher risk of clinical severity[13–15]. Better knowledge of an individual's risk can help better prioritise risk groups in the global vaccination effort as well as tailor our efforts for both prevention (e.g. targeted isolation or social distancing) and for treatment of confirmed cases such as immediate hospitalisation for people at greater risk. Many attempts have been made to provide risk prediction

models for COVID-19. Already in April 2020, Wynants et al. reviewed 145 prediction models[16]. Of the 145 described models, 23 models were aimed at estimating the risk of mortality. The review further identified that most studies had been performed on inpatients and required lab-testing (lymphocyte count, C-reactive protein, creatinine) and so were not aimed at community-based identification. In addition, the review concluded that the currently proposed models were 'poorly reported, at high risk of bias, and their reported performance is probably optimistic', with sample sizes rarely over 1000[16]. In order to define who is at high risk, governmental organisations have put forward defining criteria that aim to classify low- and high-risk groups of the population[17,18]. However, at present these criteria are developed by 'expert consensus' based on available bodies of evidence and so require time and effort and might be biased by the experts' priors. While the co-morbidities have been observed in affecting COVID-19 mortality[19] they have not been systematically analysed for building a principled integrated COVID-19 mortality risk predictor. Preconditions, especially if they appear grouped, so called multi-morbidities affect a substantial proportion of the world's population[20]. Estimates of prevalence vary depending on population examined and definitions used, in the UK >50% of people have at least two conditions by age 65, and >50% have 3+ conditions by age 70[21]. Hence, we took a more principled approach to incorporating patient specific information by systematically including comorbidities and their co-occurrence in our data-driven analysis to avoid human-induction bias, while carefully avoiding and signposting the potential for AI bias further below[22]. Our approach is enabled by the existence of large population datasets with linkage to registries such as death records, hospital admission and COVID-19 testing results and therefore represent a novel opportunity for automated clinical risk prediction model development on larger cohorts. Here, we used the UK Biobank[23,24], which since the 16 March 2020 has linked over 26,000 COVID-19 laboratory test results electronically to its 500,000 strong cohort of participants. The use of such linked datasets has an established track record for the development and evaluation of clinical risk models, including those for cardiovascular disease, cancer, and mortality[25–28].

In this paper, we utilise the longitudinal medical records from the UK Biobank to predict COVID-19-related death. We demonstrate that the data already routinely collected and stored in EHRs can be rapidly leveraged to address pressing questions related to the COVID-19 pandemic. Notably, we apply a Topic Modelling approach, Latent Dirichlet Allocation (LDA), on the entirety of the UK Biobank population to construct precise 'digital comorbidity fingerprints' (DCFs) from EHR data. Using the early data from the UK COVID-19 epidemic, we explored the utility of our approach by using the DCFs to train a Random Forrest Classifier to predict COVID-19-related death. We demonstrate the face validity of such a two-step approach to model development by comparing the prediction performance of our DCFs, combined with previously documented demographic data, to a model based on expert consensus led feature selection, and one based entirely on age. We further consider the relationship between our data-driven DCFs and clinician-driven features. The approach proposed here is designed to provide a rapid, intuitive and accurate forecasting of COVID-19-related death based on past medical records, which is of special importance for managing hospital resources, vaccination efforts and preventative policies.

Results

Characteristics of the study population

A total of 2,499 COVID-19 patients were included in this analysis (Figure 1; cohort description in Table 1). Overall, the median (IQR) age of patients at baseline was 70.4 (59.9 - 76.3, range: 50.4 - 83.7) years, and 1,284 patients (51.4 %) were male. Over the study period, COVID-19-related deaths were recorded in linked death registration data for 349 patients (14.0 % of the study population). The hospitalised patients encompassed 79% of the 349 death events. Figure 1, shows the weekly number of positive test results during the study period, from January 30 to September 9, along with the weekly number of COVID-19 related deaths in the same period. We can observe a peak in early April 2020 (week of the 8 April, 272 people tested positive) followed by a peak in COVID-19 related deaths a

week later (week of the 15 April 2020, 61 COVID-19 related deaths). The median age of the patients who died of a COVID-19 related death was 76.0 years (76.1 years for men and 75.4 years for women). The median follow-up time for all patients (time between diagnosis and either death or end of the study) was 113 days (10 days for COVID-19-related deaths and 123 days for patients still alive at the end of the study period). Figure 1 shows the overall probability of survival 28 days after diagnosis was over 0.88, dropping to 0.82; whereas for participants over 80 years old, the survival probability was 0.69 after 28 days falling to 0.64.

In order to better characterise the impact of pre-existing conditions on the study population, we used the categories from our Expert Hand-Crafted Features (EHCF) (see Box 1). The most common comorbidity in the population was hypertension (47.4%) followed by Cardiovascular disease (39.0%) and cancer (21.1%) (Table 1). The univariate and fully adjusted associations (calculated using logistic regression) between patient level characteristics (Table 1) and odds of COVID-19-related death are shown in Table 1 and Figure 2. Increasing age showed the strongest association with an increased likelihood of COVID-19-related death; with participants aged over 80 years being over 15 times (fully adjusted OR: 15.86, 95% CI: 7.85 - 32.04) more likely to die of a COVID-19-related death compared to 50–59-year-olds. Other significant characteristics that showed an increased likelihood of COVID-19-related death are male gender (fully adjusted OR: 1.54, 95% CI: 1.18 - 2.01), and a previous diagnosis of hypertension (fully adjusted OR: 1.54 (1.14 - 2.09)). These findings are consistent with previously reported associations[29–31].

Digital Comorbidity fingerprinting

Pre-existing conditions are coded as thousands of possible ICD-10 disease codes in patient records. Our methodology adapts unsupervised methods from Natural Language Processing (NLP) trained on the general population (i.e. the entire UK Biobank which includes over half a million patients) in order to cluster patients' conditions using only disease codes from a patient's past hospital visits. We use a form of Latent Dirichlet Allocation, a Bayesian clustering method that can naturally model the

categorical nature and inherent imprecision of disease codes. This clustering distils thousands of disease codes, and billions of possible combinations into a set of DCFs. Each patient's pre-existing conditions are thus summarised by a single number (between 0-100%) describing how strong a specific pre-existing condition fingerprint is present in their record. All pre-existing conditions and their myriad of combinations are summarised by a total of 30 DCFs.

We describe hereafter how a second interpretable machine learning algorithm is used to classify participants with suspected or confirmed COVID-19 into a high and low COVID-19-related death risk category based on their digital precondition fingerprint and their demographics.

We initially fitted our DCF Model on the 402,902 participants who belonged to the UK Biobank cohort, were still alive in January 2020 and were part of the Hospital Episode Statistics (HES) sub-database. Latent Dirichlet Allocation was used to generate the topic distributions for all patients and ICD-10 code group. We tested multiple values of topics (k ranging from 2 to 100) and compared the resultant topic distribution with respect to the model coherence, C_v[32]. The best results were obtained for $k = 30$ topics, and Dirichlet priors $\alpha = 0.1$ and $\eta = 0.6$. We clinically reviewed disease codes grouped in each of the generated topics, i.e. the DCFs. We observe that our comorbidity fingerprinting does not group codes simply by clinical system (e.g. respiratory disorders or diseases of the circulatory system) as can be seen in Figure 3. Rather, the method groups past clinical codes in a more functional way (all precondition fingerprints are listed in the Supplementary Box 2). For example, $DCF\ 24$ groups codes of cardiovascular disorders together (such as angina and myocardial infarction), but also includes some of the most common risk factors for these diseases, such as lipidaemia or family history of cardiac disorders. $DCF\ 7$ groups codes for rhythm disorders (e.g. atrial fibrillation) with pulmonary embolisms. Precondition fingerprint $DCF\ 22$ groups diabetic disorders with occurrences of renal failure and obesity. Other fingerprints, such as $DCF\ 5$, $DCF\ 8$ and $DCF\ 19$ are mainly defined by neoplasm and surgical procedures (e.g. acquired absence of organs).

These results indicate that comorbidity fingerprinting is able to yield clinically relevant topics by grouping diseases that often appear simultaneously. Some of the clusters are harder to interpret than others and would need to be further investigated. The DCFs are further described in Figure 4, where we demonstrate the ability of the DCFs to differentiate between patients that died of COVID-19 and the overall UK Biobank cohort. Figure 4 further outlines the DCFs ability to encompass notions of gender, age and BMI. It is, nonetheless, clear that comorbidity fingerprinting does generate coherent and interpretable results when applied to disease codes.

Prediction accuracy

In order to evaluate our DCF approach to feature engineering, we used it as an input to a supervised classifier of COVID-19 mortality and assessed its discrimination capabilities when compared to a model built using the EHCF. The prediction accuracy for the two models under consideration are shown in Table 2 and Figure 5. We use a model built from age alone as a baseline for performance evaluation (area under the receiver operating curve (AUC-ROC): 0.709 95% CI: 0.692 - 0.727). Both the EHCF Model, (AUC-ROC: 0.734, 95% CI: 0.704 - 0.764) and the DCF Model, (AUC-ROC: 0.730, 95% CI: 0.700 -0.760) achieved similar discrimination scores. A Wilcoxon's rank-sum test was performed to assess if our DCF Model showed better discrimination capabilities than the baseline. Although there was a small absolute difference between these two models (a difference of 0.021), the *p*-value was significant at 0.0409.

Model interpretability and impact

Comparing the EHCF Model to our DCF Model showed overlap in the top contributing factors for predicting COVID-19-related death (Figure 6). Age is the feature that plays the most predominant role in both models. Nine of the other top ten ranked predictor variables (diabetes, renal disease,

hypertension, respiratory diseases, sex, predisposition to infection and cardiovascular diseases) were identified in both models. Indeed, the EHCF Model highlights the importance of cardiovascular diseases, diabetes, sex, cardiac diseases, severe respiratory disease, chronic kidney disease, conditions that predispose to infection and cancer. These are mirrored in $DCF\ 24$ (cardiovascular diseases), fingerprint $DCF\ 22$, (severe diabetes and obesity), fingerprint $DCF\ 4$ (chronic pulmonary disease and tobacco usage), $DCF\ 18$ (chronic kidney failure and hypertension), $DCF\ 0$ (haematological disorders and immunosuppression). $DCF\ 8$ in the DCF Model reflects some preconditions of the female sex (acting as an indicator for the sex variable, see Figure 6) and some associated diseases by concentrating on female-specific cancer codes such as breast cancer and more generic ICD-10 cancer codes such as chemotherapy sessions. Chronic neurological conditions are important predictors in the EHCF Model but this was not directly identified as important in the DCF Model. Instead, the DCF Model puts emphasis on gastrointestinal cancers and disorders of the GI tract through the importance of $DCF\ 19$, this is a feature not covered in the EHCF Model. Thus, we obtained a different set of features in our comorbidity fingerprinting (DCF), as it encompassed significantly more comorbidities per topic, than feature sets assembled by-hand (EHCF).

Discussion

Our findings demonstrate how taking a strictly data driven approach using disease codes directly from EHRs can be rapidly leveraged for predicting outcomes and studying disease patterns. The model developed in this study required, in addition to the previously available data (prior to January 2020), only diagnosis information and cause of death, to predict who, with a positive COVID-19 test, was likely to die. Even though our model does not rely on symptoms, laboratory values or images (at the time of diagnosis or during the illness) to predict mortality, we are able to achieve comparably high discrimination results (ROC-AUC of 0.703). Other published COVID-19 prognosis models have AUC-ROC's ranging from 0.68 to 0.90[33–35]. The use of physiological data at the point of admission further makes these models difficult to implement in a community setting or use a specific predefined subset of comorbidities[36].

In order to evaluate the capabilities of our topic model-based approach to clinical dimensionality reduction, we compared it to current medical risk prediction approaches [19,35]. We opted to develop a set of expert hand-crafted features, built using the CDC and NHS COVID-19 shielding list combined with current research. Our method showed comparable discrimination capabilities, (ROC-AUC of 0.730) to the current expert models (ROC-AUC of 0.734). The significance of our AI-based methodology is highlighted by the ability of our model to encompass the entire past medical history of the participants and, in the context of a new and unknown disease, not be restricted by potentially biased prior belief. In this way, the automated, data-driven approach we present here is able to avoid human-induction bias in the classifier's input features and permits the detection of novel insights in a new disease in a very short amount of time and with minimal human supervision. Furthermore, by ensuring a fully automated process, the methodology we present here is easily scalable to other healthcare systems and generalisable to future events. Furthermore, the DCFs we present here are not specifically tied to COVID-19 risk prediction and can thus be immediately reused to summarise risk predictions for COVID-19 variants or other diseases.

Our study provides results regarding key findings in the underpinning structure of comorbidities and validates the usefulness of these associations with their predictability of COVID-19 mortality. This is of particular importance given the high UK prevalence (18.5 million individuals) of underlying conditions that increase the risk of severe COVID-19[37]. Our DCFs offer a rich and deep set of features, that can succinctly summarise a patient's comorbidity profile using topic modelling. This approach allows for the use of a high-performance AI algorithm whilst remaining clinically interpretable and intuitive. Some obvious common causes of comorbidity clusters are well known (e.g. cardiovascular disease due to smoking or obesity[38]), advances in the science of multi-morbidities are challenging because diseases may have linked or independent causes, combine heterogeneously, are treated differently, and thus affect care pathways and medication strategies. Structuring the myriad of combinations of comorbidities, such as through our comorbidity fingerprinting, underpins any

measures to slow the accumulation of conditions and to optimise their treatment. We have shown that AI methods can help in identifying multimorbidity clusters without human-induction bias, while at the same time ensure that findings remain clinically useful.

We further demonstrate that the DCFs generated using LDA correspond well and show significant variable importance overlap when compared with the EHCF Model and previously identified risk factors from published epidemiological studies. For example, our model supports the previous epidemiological findings that cardiovascular disorders and hypertension increase risk of COVID-19-related death [12,13,19]. Several features identified in the model warrant further evaluation as they were not specifically seen in the initial COVID-19 epidemiological studies. DCF 19 'GI Disorders' was one of the most important features in our model, however many of the epidemiological studies do not report this as a significant risk factor [13,31].

There are a few limitations to our study. Firstly, our data set while vast, may also reflect the inherent bias of the UK Biobank[39], which has been discussed in detail elsewhere[39–41]; notably, the demographic reflect a "healthy volunteer" bias, with individuals being generally older, from more educated, less deprived socioeconomic backgrounds, and with significant under-representation of ethnic minorities compared to the UK population. Secondly, testing, treatments, and outcomes of COVID-19 have continuously improved during the study period, thus possibly having a confounding effect on the results. Moreover, due to the limited availability of testing kits for COVID-19, priority was initially offered to those considered at a higher clinical risk, thus potentially leading to an overestimation of severe outcomes in the database. However, these COVID-19 specific data bias factors only affect the mortality prediction but not the structure of the DCFs. This study further relied on retrospective secondary care EHRs and the model is therefore currently blind to conditions entirely managed in primary care, such conditions include many less severe cases of diabetes, asthma and hypertension. At the time of the study, data from primary care was not available. Future work will be needed to

incorporate data from General Practices into the development of the DCFs. Balancing the strengths and limitations, we consider our derivation population to be relevant for the initial exploration of COVID-19 mortality in an older and more 'at risk' population using a machine leaning approach. Further research and external validation of the predictive models and DCFs in this study will be required before applying this work in a more general context, specifically to better asses the risk in a younger and fully diverse population.

In conclusion, we developed a novel machine learning based approach to predict COVID-19 mortality in a community cohort only from past EHR data. We have demonstrated the feasibility of a comorbidity clustering approach to avoid human-induction bias and succinctly summarise billions of possible comorbidity combinations in COVID-19 prognosis modelling and thus rapidly leverage vast EHR dataset for outcome prediction. We find, that age and qualitative co-morbidity information are very powerful mortality predictors, making other demographic or physiological data pale in comparison. This implies, that in rapid decision situations or decisions with limited information will benefit from our solution as this enables healthcare professionals to form a rapid picture from information that patients and their relatives are familiar with and which do not change rapidly, unlike data such as blood pressure. Thus, our model may enable early stratification of key clinical risk groups thus permitting earlier intervention and may help better prioritise risk groups in the global vaccination effort.

Crucially, however, our DCF framework can be applied to any other form of novel disease as the multi-morbidity features we discovered span a space that is disease agnostic.

Methods

Study design and participants

We conducted a prospective open cohort study using the Hospital Episode Statistics for England (HES) database, COVID-19 test results and death data linked through the UK Biobank (see `Data Source'). The cohort study began on 30 January 2020, which was chosen as it corresponds to the day of the first laboratory-confirmed case of COVID-19 in the UK[42]; and ended on 9 September 2020. The cohort study examines risk among the population of patients who are positive for COVID-19. Therefore, we only included patients from the UK Biobank if they had a positive SARS-COV-2 test result or a COVID-19 diagnosis code recorded in their clinical notes (ICD-10 code U071 or U072).

Data source and ethical approval

The UK Biobank is a large prospective population cohort of 502,505 participants aged 40-70 years at baseline, prepared to travel to 1 of 22 assessment centres in England, Scotland, and Wales[23]. All participants were recruited between 2006 and 2010, and all consented to have their health followed[23]. Baseline assessments include nurse-led interviews surrounding socio-demographics, lifestyle, physiological measurements, and medical history. Health outcomes were sourced from linkages to electronic medical records. UK Biobank obtained approval from the North West Multi-Centre Research Ethics Committee (MREC), and the Community Health Index Advisory Group (CHIAG). All participants provided written informed consent prior to enrolment in the study. The UK Biobank protocol is available online[43].

In the context of the COVID-19 pandemic, the UK Biobank has implemented a rapid and dynamic linkage between the laboratory results for COVID-19, stored on Public Health England's Second-Generation Surveillance System, and the UK Biobank participants[24]. The previous hospital medical information of the study participants was obtained from linkage to the HES database (a data

warehouse containing records of all admissions, emergency room attendances and outpatients appointments at NHS hospitals in England)[44].

Access to anonymised data for the UK Biobank cohort was granted by the UK Biobank Access Management Team (application number 21770). Ethical approval was granted by the national research ethics committee (REC 16/NW/0274) for the overall UK Biobank cohort.

Outcome

The primary outcome measure was COVID-19-related death in patients with a COVID-19 infection. Confirmation of a COVID-19-related death COVID-19 death was defined as 1) a COVID-19-related death recorded in ONS death certificate data (International Classification of Disease, 10th edition, ICD-10 codes U071 or U072). 2) a death occurring within 28 days of a laboratory-confirmed COVID-19 infection (as per UK government guidelines[45]).

Model Development

The performance of machine learning methods is heavily dependent on the selection of features[46]. For that reason, most of the effort in an analytic model is spent pre-processing, merging, customizing, and cleaning datasets. This task is made all the more complicated in EHRs as the different predictors may number the tens of thousands (over 11,700 are present in the UK Biobank). Traditional modelling approaches deal with this complexity by choosing a limited number of variables and creating custom features[47]. However, this is a largely manual and often labour-intensive task. Another common

approach is to use unsupervised dimensionality reduction approaches, this is often far faster but may yield results that are less accurate or interpretable.

Hand-Crafted features: Predictor variables

The patient-level characteristics included in the development of the EHCF Model were selected from the existing health conditions listed on original population-level risk stratification method as exercised in UK[17] or listed by the Centers for Disease Control and Prevention (CDC) as defining 'people at increased risk for severe illness'[18]. This was further complemented with other emerging risk factors for severe outcomes of COVID-19 (such as raised blood pressure)[9,12,13,19] . The final predictor variables chosen are summarised in Box 1.

Demographic and lifestyle variables were also included, namely: age, sex, Townsend deprivation index[48], ethnicity, and smoking status. Age groups were categorised as 50-60, 60-70, 70-80 and 80+ years. Deprivation was measured using the Townsend Deprivation Index (grouped into quartiles based on the entire UK Biobank distribution, greater scores imply a greater degree of deprivation). Ethnicity was grouped as white and non-white (thus collapsing the UK Biobank categories mixed, asian, black or chinese), due to the otherwise small number of participants in each of the specific categories. Smoking status was determined from data from the initial UK Biobank assessment and grouped into participants that smoke or have smoked (current or previous at baseline) and participants that, at baseline, were never smokers.

The demographic variables considered as potential factors affecting COVID-19 risk and were determined using the answers given at UK Biobank initial assessment centre. Age was determined on

30 January 2020. Information on all other covariate variables was determined using the ICD-10 codes in the linked HES database.

Automated features: topic modelling

Topic modelling was originally developed as a tool to model collections of discrete data with particular application in text modelling and Natural Language Processing (NLP). Topic modelling can discover abstract 'topics' within a collection of documents, where a topic consists of a collection of words that frequently occur together. Here we opted to train topic models using Latent Dirichlet Allocation (LDA), an unsupervised and interpretable method for topic modelling[49,50]. LDA is a Bayesian probabilistic model that works by taking as input a corpus of clinical codes and represents each document as a finite mixture of an underlying set of fixed topics, with each topic characterised by its distribution over words[49,51].

Here, LDA is used to represent a patient p as a mixture over the collection of K "topics". Each topic k defines a multinomial distribution over a finite vocabulary of ICD-10 diagnosis codes and is assumed to have been drawn from a Dirichlet, $\beta_k \sim D\text{irichlet}(\eta)$. Thus, each code $c_{p,n}$ has a given probability β_k in each topic k. Given the topics, LDA then generates for each patient p a distribution over topics $\theta_p \sim D\text{irichlet}(\alpha)$. In other words, each topic is determined as a distribution over closely related codes and each patient is represented as a distribution over multiple topics. For simplicity, we assumed symmetric priors on the patient topic distribution θ and the code topic distribution β.

For each participant of the UK Biobank, we considered a full longitudinal visit history to hospital, including all admissions, emergency room attendances and outpatient appointments. For each recorded hospital episode, we extracted a list of diagnosis codes, using the ICD-10. This generated a document

for each unique patient in the UK Biobank, representing their full hospital visit history as a series of ‘sentences` (one for each visit, so if a patient visit the hospital 10 times for asthma it is represented 10 times in their record.), wherein each ‘word` was a unique ICD-10 diagnosis code. Thus, this way of accounting naturally weighs conditions that require more frequent visits to the hospital.

The LDA topic model was trained using as input the corpus of documents (i.e. sentences of ICD-10 diagnosis codes) for the entirety of the UK Biobank population that is not lost to follow up, still alive on 30 January 2020 and is part of the HES database (‘Population part of the HES database’, Fig. 1). This was chosen with the aim of generating more clinically robust and disease agnostic topics (i.e. the topics formed are not specific to any prior comorbidity). From this corpus, we used a grid-searching method to determine the optimal number K of topics, and the prior value for θ and β (α and η) when optimising with respect to model coherence, C_v in[32] as this method has been shown to achieve the highest correlation with human topic ranking data[32,52]. This model was implemented using the gensim library in the Python programming language[53].

Development of a random forest model

Two predictive classifiers were trained in this study, EHCF Model and DCF Model. The first, EHCF Model, uses as input the ‘manually’ determined participant characteristics or ‘Hand-Crafted Features’ (see Hand-Crafted features: Predictor variables). These were grouped into the aforementioned categories and each participant was set a binary value indicating whether or not their past medical records (extracted from the HES database) showed the occurrence of at least one of the conditions for each category. The DCF Model, on the other hand, uses the patients’ distribution over the previously derived topics of an LDA topic model as input to a classifier. The LDA model was further complemented with the demographic and lifestyle variables (e.g. age).

The models described above were trained to predict the same binary output; the patients that survived were assigned to class 0 and those that died to class 1.

In this study, we opted for use of a random forest (RF) classifier[54], to determine COVID-19 associated mortality. RF is a high-performance ML algorithm that fits several decision tree classifiers on various sub-samples of the dataset and, subsequently, averages these tree predictors[54]. This allows for a reduction in over-fitting and improvement in accuracy. We implemented all RF models using the Scikit-lean library in python programming language[55]. The models' hyper-parameters[56–59] were determined via grid search with AUC as the response as it yields a finer evaluation than commonly used error rates[58].

The overview of the pipeline used for DCF Model is illustrated in Figure 7. Figure 7 describes the learning of topics from the data and shows how we use the distribution of each patient over each of these topics as input to a RF classifier.

Variable ranking

In order to better understand the predictor variables that may lead to increased risk of a COVID-19-related death and their interactions, we use an impurity-based approach to rank the contribution of the different variables in the prediction made by the model. This resulting variable importance score reflects the relative impact that each variable has on the prediction issued by the model.

We chose to use a RF algorithm for this analysis because it is a non-parametric algorithm that can recognise complex patterns and automatically capture nonlinear and interaction effects without specifying these a priori[60].

Statistical analysis

Study participant numbers are depicted in the flow chart (Figure 1). For each of the 'Hand-Crafted' patient characteristics (see Hand-Crafted features: Predictor variables), a logistic regression model was fitted with the COVID-19-related death as the outcome in order to determine a univariate (not-adjusted) Odds Ratio (OR). All patients' characteristics were subsequently included in a single multivariable logistic regression to determine the adjusted OR. All OR from the univariable and multivariable logistic regression models are reported with 95% confidence intervals. For the purpose of this study, the logistic regression model was not considered for its predictive power but rather for its ability to simply and accurately describe the available data.

The HES data were considered complete with no missing data for our population. Participants with missing smoking information were assumed to be non-smokers. Missing deprivation score was imputed using the median value of the entire UK Biobank cohort. The participants with missing ethnicity were dropped from the study as no reliable way could be determined to input this field.

We subsequently developed RF based models in order to assess discrimination capability and compare variable importance. In order to avoid over-fitting, we evaluated the prediction accuracy of these models using 10-fold stratified cross-validation and calculated the area under the AUC-ROC as a measure of model discrimination. In every cross-validation fold, a training sample (90% of the participants) was used to derive the models, and then a hold-out sample (10% of the participants) was

used for performance evaluation. We report the mean AUC-ROC and the 95% confidence intervals for all models. In all RF models, age and deprivation score are considered as continuous variables.

Data Availability

The UK Biobank cohort data that supports the findings of this study is available to researchers as approved by the Biobank Access Management Team.

References

1. WHO. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (2020).

2. WHO. Coronavirus disease (COVID-19): Weekly Epidemiological Update. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (2020).

3. Arabi, Y. M., Murthy, S. & Webb, S. COVID-19: a novel coronavirus and a novel challenge for critical care. *Intensive Care Med.* 1–4 (2020).

4. Grasselli, G., Pesenti, A. & Cecconi, M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response. *Jama* **323**, 1545–1546 (2020).

5. Jordan, R. E., Adab, P. & Cheng, K. K. Covid-19: risk factors for severe disease and death. (2020).

6. Onder, G., Rezza, G. & Brusaferro, S. Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *Jama* **323**, 1775–1776 (2020).

7. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* (2020).

8. Garg, S. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—COVID-NET, 14 States, March 1--30, 2020. *MMWR. Morb. Mortal. Wkly. Rep.* **69**, (2020).

9. Richardson, S. *et al.* Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *Jama* (2020).

10. Guan, W. *et al.* Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* **382**, 1708–1720 (2020).

11. Jin, J.-M. *et al.* Gender differences in patients with COVID-19: Focus on severity and mortality. *Front. Public Heal.* **8**, 152 (2020).

12. Deng, G., Yin, M., Chen, X. & Zeng, F. Clinical determinants for fatality of 44,672 patients with COVID-19. *Crit. Care* **24**, 1–3 (2020).

13. Docherty, A. B. *et al.* Features of 16,749 hospitalised UK patients with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol. *medRxiv* (2020).

14. House, N., Holborn, H. & Wc, L. ICNARC report on COVID-19 in critical care. *Publ. online* **26**, 24 (2020).

15. Patel, B. V *et al.* Natural history, trajectory, and management of mechanically ventilated COVID-19 patients in the United Kingdom. *medRxiv* (2020).

16. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *bmj* **369**, (2020).

17. Digital, N. Covid-19—high risk shielded patient list identification methodology. https://digital.nhs.uk/coronavirus/shielded-patient-list/methodology.

18. CDC. Covid-19: People at Increased Risk. https://www.cdc.gov/coronavirus/2019-ncov/need-

extra-precautions/people-with-medical-conditions.html (2020).

19. Clift, A. K. *et al.* Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *bmj* **371**, (2020).

20. The Academy of Medical Sciences. Multimorbidity: a priority for global health research. *Https://Acmedsci.Ac.Uk/Policy/Policy-Projects/Multiple-Morbidities-As-a-Global-Health-Challenge* (2015).

21. Barnett, K. *et al.* Epidemiology of multimorbidity and implications for health care, research, and medical education: A cross-sectional study. *Lancet* (2012) doi:10.1016/S0140-6736(12)60240-2.

22. Rich, A. S. & Gureckis, T. M. Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence* (2019) doi:10.1038/s42256-019-0038-z.

23. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med* **12**, e1001779 (2015).

24. Armstrong, J. *et al.* Dynamic linkage of COVID-19 test results between Public Health England's second generation surveillance system and UK biobank. *Microb. genomics* **6**, e000397 (2020).

25. Alaa, A. M., Bolton, T., Angelantonio, E. Di, Rudd, J. H. F. & van der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS One* (2019) doi:10.1371/journal.pone.0213653.

26. Weng, S. F., Vaz, L., Qureshi, N. & Kai, J. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS One* (2019) doi:10.1371/journal.pone.0214365.

27. Hippisley-Cox, J. & Coupland, C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: Prospective cohort study. *BMJ*

Open (2015) doi:10.1136/bmjopen-2015-007825.

28. Clift, A. K. *et al.* Development and validation of risk scores for all-cause mortality for a smartphone-based 'general health score' application: a prospective cohort study using the UK Biobank (Preprint). *JMIR mHealth uHealth* (2020) doi:10.2196/25655.

29. Jain, V. & Yuan, J.-M. Systematic review and meta-analysis of predictive symptoms and comorbidities for severe COVID-19 infection. *medRxiv* (2020).

30. Chen, T. *et al.* Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. *Bmj* **368**, (2020).

31. Wu, Z. & McGoogan, J. M. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *Jama* **323**, 1239–1242 (2020).

32. Röder, M., Both, A. & Hinneburg, A. Exploring the space of topic coherence measures. in *Proceedings of the eighth ACM international conference on Web search and data mining* 399–408 (2015).

33. Carr, E. *et al.* Supplementing the National Early Warning Score (NEWS2) for anticipating early deterioration among patients with COVID-19 infection. *medRxiv* (2020).

34. Zhang, H. *et al.* Risk prediction for poor outcome and death in hospital in-patients with COVID-19: derivation in Wuhan, China and external validation in London, UK. (2020).

35. Knight, S. R. *et al.* Risk stratification of patients admitted to hospital in the United Kingdom with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of a multivariable prediction model for mortality. *Br. Med. J.* (2020).

36. Williams, R. D. *et al.* Seek COVER: Development and validation of a personalized risk calculator for COVID-19 outcomes in an international network. *medRxiv* (2020).

37. Walker, J. L. *et al.* UK prevalence of underlying conditions which increase the risk of severe

COVID-19 disease: a point prevalence study using electronic health records. *BMC Public Health* **21**, 484 (2021).

38. Vetrano, D. L. *et al.* Twelve-year clinical trajectories of multimorbidity in a population of older adults. *Nat. Commun.* (2020) doi:10.1038/s41467-020-16780-x.

39. Fry, A. *et al.* Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

40. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* **12**, e0174944 (2017).

41. Ganna, A. & Ingelsson, E. 5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study. *Lancet* **386**, 533–540 (2015).

42. Gov.UK. Coronavirus (COVID-19) cases in the UK. https://web.archive.org/web/20200502045059/https://coronavirus.data.gov.uk/ (2020).

43. Palmer, L. J. UK Biobank: bank on it. *Lancet* **369**, 1980–1982 (2007).

44. Adamska, L. *et al.* Challenges of linking to routine healthcare records in UK Biobank. *Trials* **16**, 1 (2015).

45. GOV.UK. New UK-wide methodology agreed to record COVID-19 deaths. (2020).

46. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).

47. Goldstein, B. A., Navar, A. M., Pencina, M. J. & Ioannidis, J. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J. Am. Med. Informatics Assoc.* **24**, 198–208 (2017).

48. Black, D. *Inequalities in health: the Black report*. (Penguin Books, 1982).

49. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–

1022 (2003).

50. Blei, D. M. Probabilistic topic models. *Commun. ACM* **55**, 77–84 (2012).

51. Griffiths, T. L. & Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci.* **101**, 5228–5235 (2004).

52. Syed, S. & Spruit, M. Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. in *2017 IEEE International conference on data science and advanced analytics (DSAA)* 165–174 (2017).

53. Rehurek, R. & Sojka, P. Software Framework for Topic Modelling with Large Corpora. *Proc. Lr. 2010 Work. New Challenges NLP Fram.* (2010).

54. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

55. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).

56. Huang, B. F. F. & Boutros, P. C. The parameter sensitivity of random forests. *BMC Bioinformatics* **17**, 331 (2016).

57. Probst, P., Wright, M. N. & Boulesteix, A.-L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **9**, e1301 (2019).

58. Probst, P. & Boulesteix, A.-L. To tune or not to tune the number of trees in random forest. *J. Mach. Learn. Res.* **18**, 6673–6690 (2017).

59. Boulesteix, A.-L., Janitza, S., Kruppa, J. & König, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2**, 493–507 (2012).

60. Wright, M. N., Ziegler, A. & König, I. R. Do little interactions get lost in dark random forests? *BMC Bioinformatics* **17**, 145 (2016).

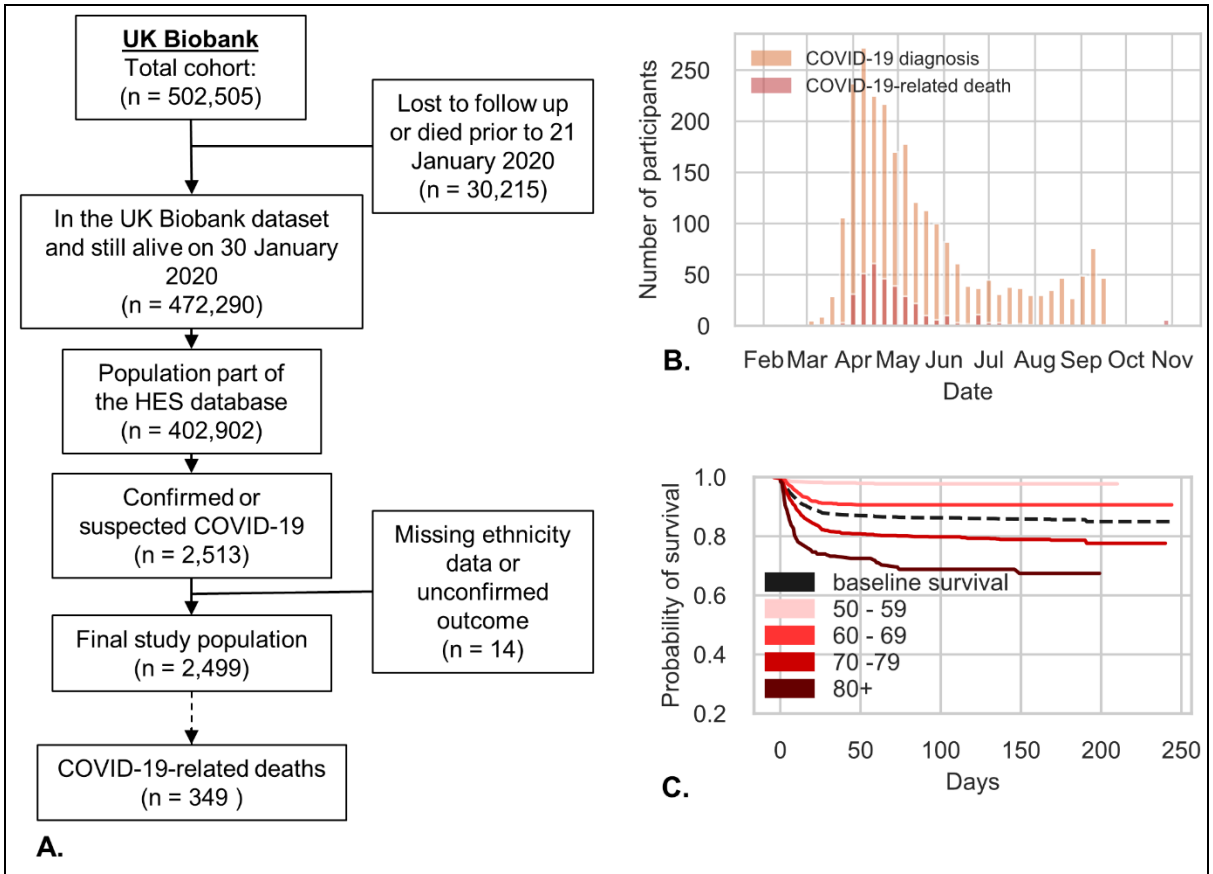

Figure 1. Cohort description. (A) Flowchart diagram of the study cohort. (B) Distribution of COVID-19 diagnosis and death. (C) Kaplan-Meier plots for COVID-19-related death per age group

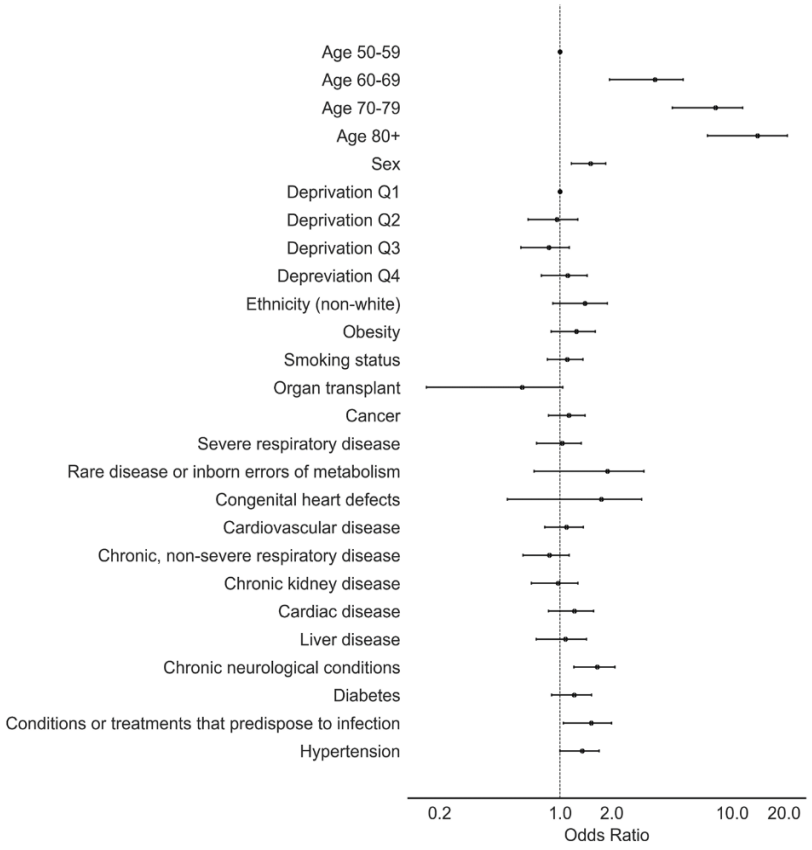

Figure 2. Estimated Odds Ratio for each patient characteristic from a multivariable logistic regression

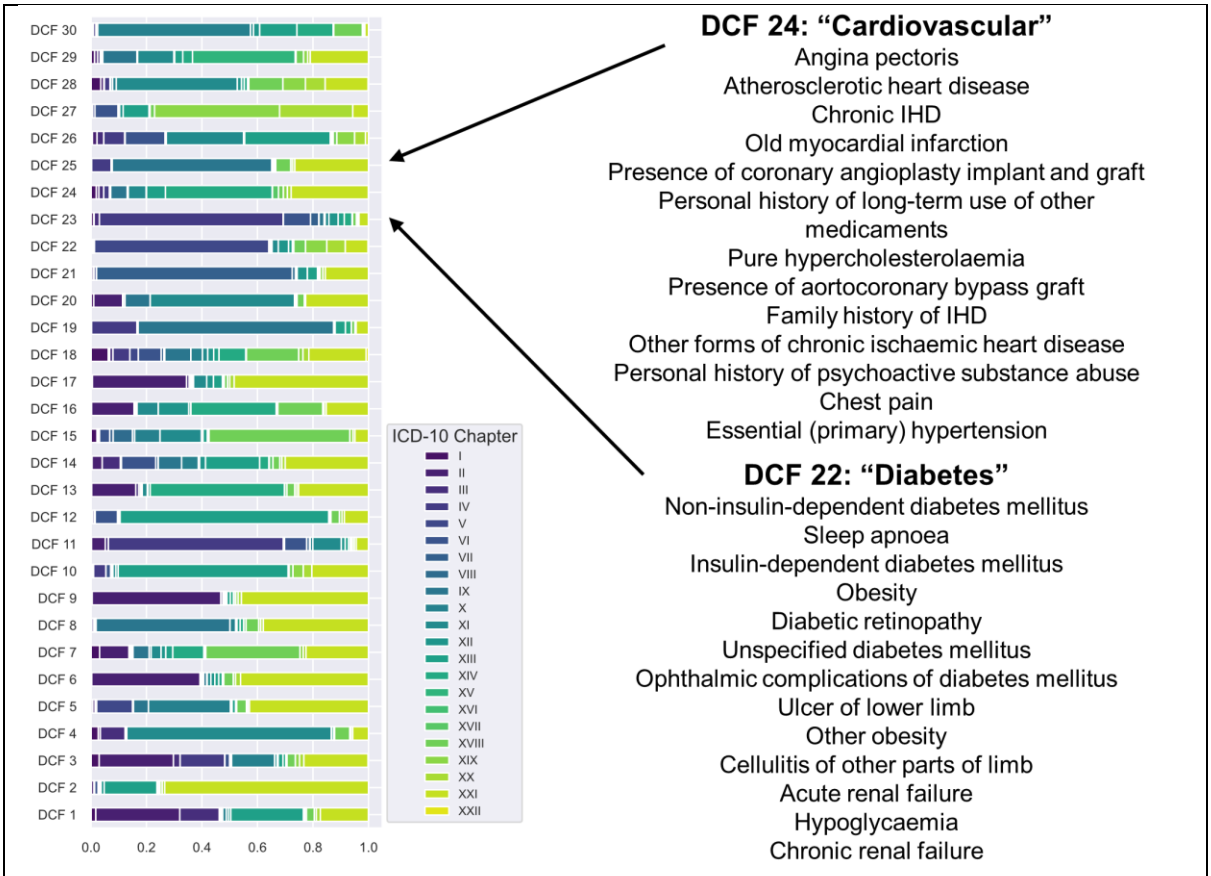

Figure 3. Topic description. We fit a 30-topic LDA model to 420,000 patients from the UK Biobank. On the left is the distribution of the 22 ICD-10 'Chapters' (Supplementary Box 1) over the 30 topics. On the right are the top 13 ICD-10 codes from topics 22 and 24 (in rank order).

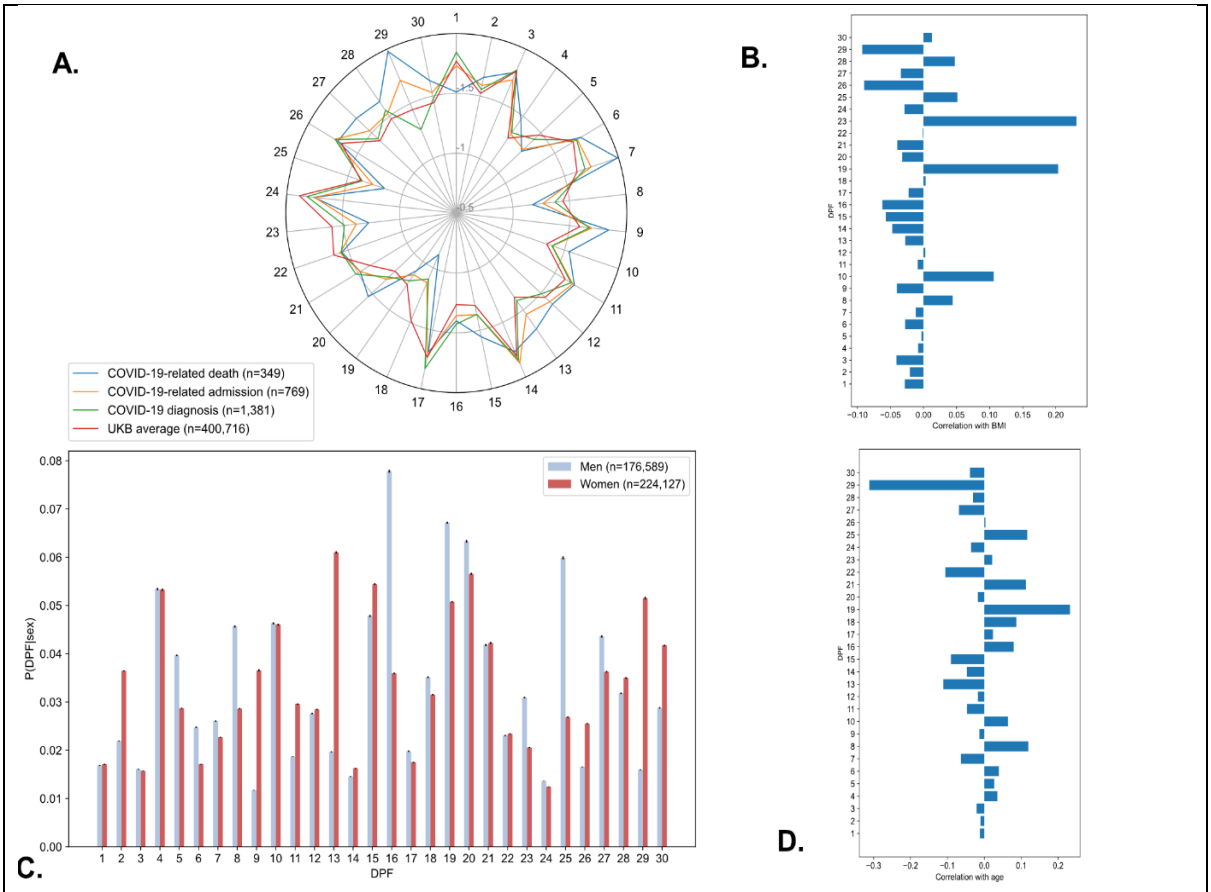

Figure 4. Summary of the Digital Precondition Fingerprints. (A) Details the average probability of belonging to each DPF (log 10 scale) by COVID-19 outcome. (B) Shows the Pearson correlation between each of our DCFs and the patient's BMI. (C) Shows the average probability of belonging to each DCF by gender and (D) shows the Pearson correlation between the each DCF and the patient's age.

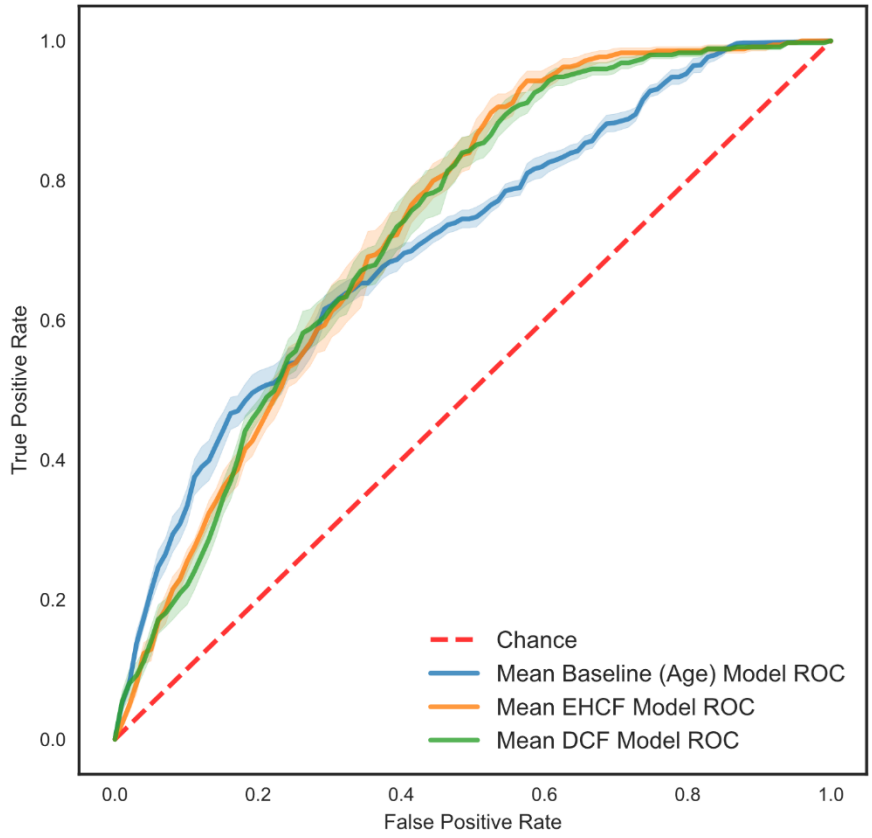

Figure 5. Mean receiver operating characteristic curves derived from predicting COVID-19-related death in the 10-Fold Cross Validation (standard error of the mean) for all models.

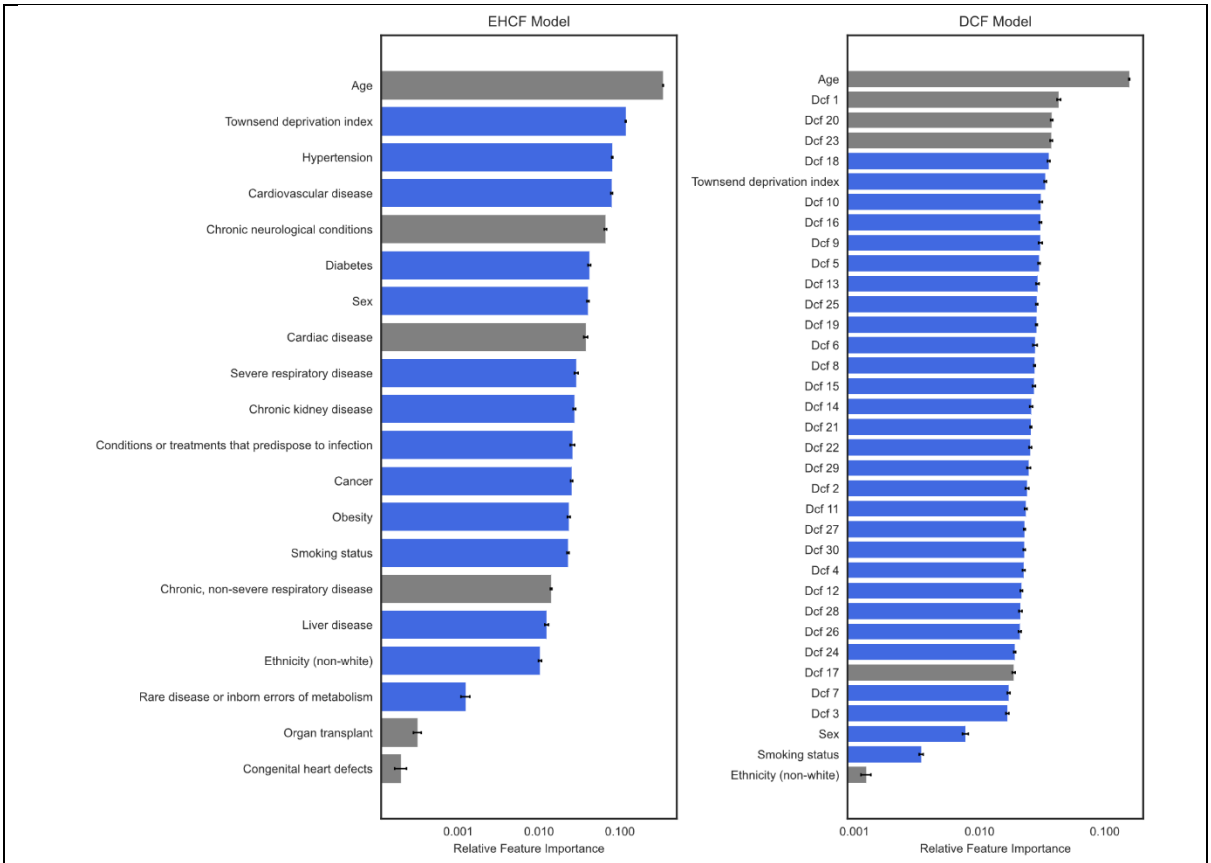

Figure 6. Comparison of relative feature importance (note logarithmic scale) showing the contribution to the predictions of the EHCF and our DCF Model. In grey the demographic and lifestyle features which were the same in both models and in blue the model specific features. Note, how in the DCF Model aside from age the co-morbidity and multi-morbidity make up the key 15 features.

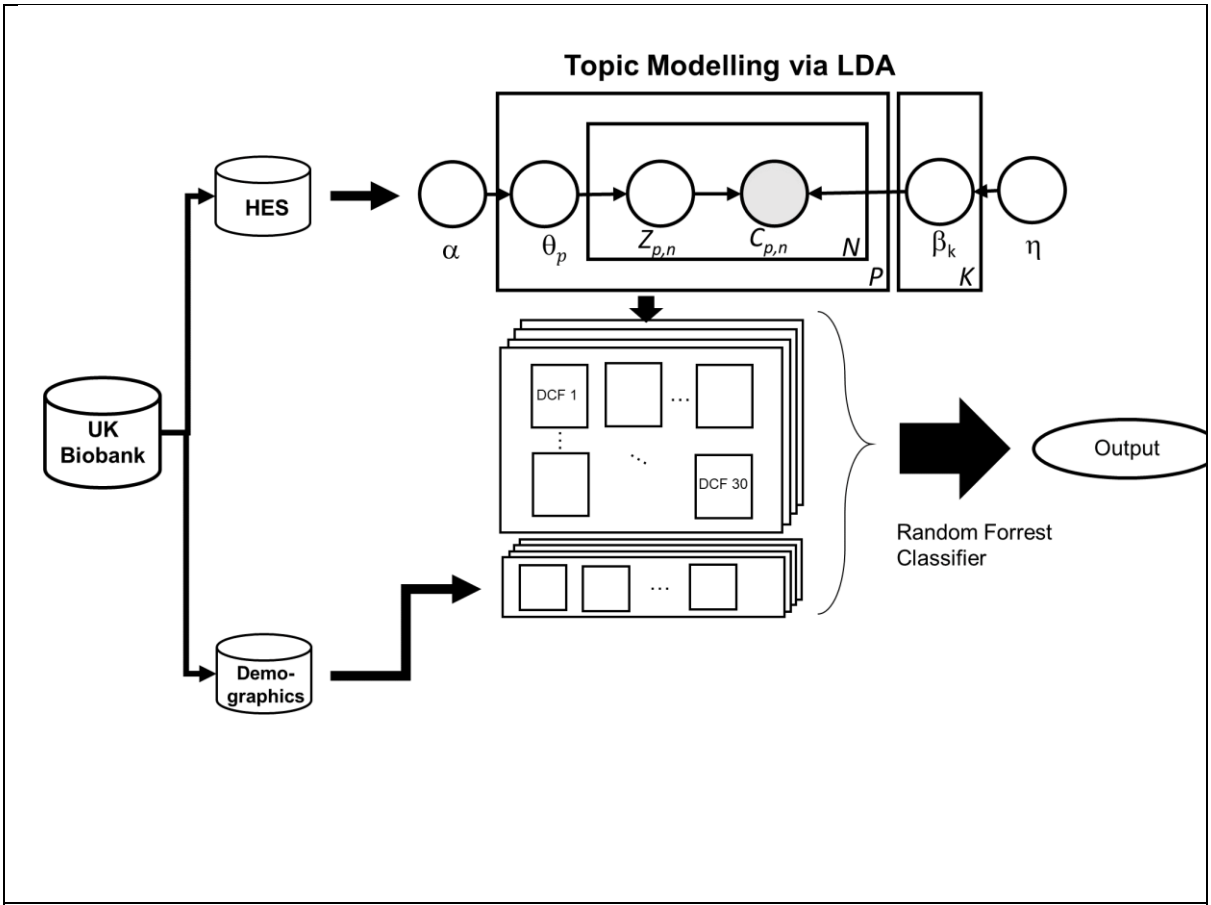

Figure 7. Illustration of the overall study design for the DCF Model. First, we separate the UK Biobank into HES and demographic data, we generate the features (using LDA) from the discrete input of the HES database. We subsequently use the distribution of each patient over these features as input to a supervised random forest classifier to predict COVID-19 fatality.

Box 1. Hand-Crafted Predictor variables

Demographics
• Age in years • Townsend deprivation index • Ethnicity

- White
- Non-white

Lifestyle

- Smoking status
  - Current or previous smoker
  - Never smoked at UK Biobank baseline

Conditions associated with high risk (clinically extremely vulnerable) on current shielded patient list as per NHS guidance

- Solid organ transplant
- Cancers
- Severe respiratory disease
  - Chronic obstructive pulmonary disease (COPD, emphysema and chronic bronchitis)
  - Status Asthmaticus
  - Cystic fibrosis, interstitial lung diseases, sarcoidosis, pulmonary hypertension
- Rare disease or inborn errors of metabolism
  - Sickle-cell disorders
  - Immunodeficiency
  - Severe metabolic disorders (e.g amino-acid metabolism disorders)
  - Congenital malformations (e.g Phakomatoses)
- Congenital heart defects

Conditions associated with an increased risk for severe COVID-19 as per CDC guidance (not already considered)

- Cardiovascular disease
  - Conduction disorders (e.g atrial fibrillation)

- Past cardiovascular event (e.g myocardial infarction, stroke, angina)
- Peripheral vascular disease
- Endocarditis

- HIV

Conditions associated with a moderate risk (clinically vulnerable) as per NHS guidance

- Chronic, non-severe respiratory disease
  - Asthma
  - Hypersensitivity pneumonitis
- Chronic kidney disease
  - CKD >= stage 3
  - End stage renal failure involving dialysis
  - End stage renal failure involving a transplant
- Cardiac disease
  - Heart failure
  - Valve disorders

- Liver disease
  - Chronic hepatitis
  - Alcoholic liver disease
  - Toxic liver disease
  - Other causes of liver cirrhosis
- Chronic neurological conditions
  - Motor neurone disease
  - Parkinson's disease
  - Multiple sclerosis

- Myasthenia gravis
- Cerebral palsy
- Epilepsy
- Down's syndrome
- Dementia (vascular or Alzheimer's)

- Diabetes
  - Type 1
  - Type 2
- Immunocompromised states (including conditions or treatments that predispose to infection (e.g. steroid treatment)
  - Rheumatoid arthritis
  - Systemic lupus erythematosus
  - Psoriasis
  - Ankylosing spondylitis
  - Connective tissue diseases (e.g. Ehlers-Danlos)
  - Vasculitis (e.g. giant cell arteritis)

Other emerging risk factors for severe outcomes of COVID-19

- Hypertension

Table 1. Descriptive characteristics of the UK Biobank cohort by COVID-19 fatality and Odds ratios and 95 % confidence intervals for COVID-19-related death

	COVID-19-related death (n = 349)	**Rest of the Cohort (n = 2150)**	**OR (95% CI)**	**Fully adjusted OR (95% CI)**

Age 50-59	14 (4.01%)	619 (28.79%)	1	1
Age 60-69	55 (15.76%)	537 (24.98%)	4.53 (2.49 - 8.23)	3.56 (1.93 - 6.54)
Age 70-79	234 (67.05%)	894 (41.58%)	11.57 (6.68 - 20.04)	8.00 (4.49 - 14.25)
Age 80+	46 (13.18%)	100 (4.65%)	20.34 (10.78 - 38.36)	14.03 (7.18 - 27.44)
Sex	221 (63.32%)	1063 (49.44%)	1.77 (1.40 - 2.23)	1.50 (1.16 - 1.94)
Deprivation Q1	64 (18.34%)	389 (18.09%)	1	1
Deprivation Q2	69 (19.77%)	427 (19.86%)	0.98 (0.68 - 1.42)	0.96 (0.65 - 1.41)
Deprivation Q3	80 (22.92%)	567 (26.37%)	0.86 (0.60 - 1.22)	0.86 (0.59 - 1.25)
Depreviation Q4	136 (38.97%)	767 (35.67%)	1.08 (0.78 - 1.49)	1.11 (0.78 - 1.57)
Ethnicity (non-white)	35 (10.03%)	243 (11.30%)	0.87 (0.60 - 1.27)	1.40 (0.91 - 2.14)
Obesity	67 (19.20%)	279 (12.98%)	1.59 (1.19 - 2.14)	1.24 (0.89 - 1.74)
Smoking status	240 (68.77%)	1326 (61.67%)	1.37 (1.07 - 1.74)	1.10 (0.84 - 1.44)
Organ transplant	3 (0.86%)	19 (0.88%)	0.97 (0.29 - 3.30)	0.60 (0.17 - 2.17)
Cancer	97 (27.79%)	432 (20.09%)	1.53 (1.18 - 1.98)	1.13 (0.86 - 1.48)

Severe respiratory disease	69 (19.77%)	237 (11.02%)	1.99 (1.48 - 2.67)	1.03 (0.73 - 1.45)
Rare disease or inborn errors of metabolism	6 (1.72%)	18 (0.84%)	2.07 (0.82 - 5.26)	1.89 (0.71 - 5.04)
Congenital heart defects	4 (1.15%)	11 (0.51%)	2.25 (0.71 - 7.12)	1.74 (0.49 - 6.09)
Cardiovascular disease	206 (59.03%)	761 (35.40%)	2.63 (2.09 - 3.31)	1.09 (0.81 - 1.46)
Chronic, non-severe respiratory disease	52 (14.90%)	302 (14.05%)	1.07 (0.78 - 1.47)	0.87 (0.61 - 1.24)
Chronic kidney disease	59 (16.91%)	190 (8.84%)	2.10 (1.53 - 2.88)	0.97 (0.68 - 1.40)
Cardiac disease	71 (20.34%)	211 (9.81%)	2.35 (1.74 - 3.16)	1.21 (0.86 - 1.71)
Liver disease	42 (12.03%)	176 (8.19%)	1.53 (1.07 - 2.19)	1.07 (0.73 - 1.59)
Chronic neurological conditions	77 (22.06%)	215 (10.00%)	2.55 (1.91 - 3.40)	1.64 (1.20 - 2.24)
Diabetes	100 (28.65%)	353 (16.42%)	2.04 (1.58 - 2.65)	1.21 (0.89 - 1.63)
Conditions or treatments that predispose to infection	50 (14.33%)	157 (7.30%)	2.12 (1.51 - 2.98)	1.52 (1.05 - 2.20)
Hypertension	238 (68.19%)	942 (43.81%)	2.75 (2.16 - 3.50)	1.34 (1.00 - 1.81)

Table 2. Performance of all prediction models under consideration

Predictor Set	**Results**	
	AUC-ROC	**95% CI**
Baseline	0.709	0.692 - 0.727
EHCF - Model	0.734	0.704 - 0.764
DCF - Model	0.730	0.700 -0.760

Supplementary Material

Supplementary Box 1. Chapter number and description for each of the 22 chapter's making up the ICD-10 clinical codes

I: Certain infectious and parasitic diseases'

II: Neoplasms

III: Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

IV: Endocrine, nutritional and metabolic diseases

V: Mental and behavioural disorders

VI: Diseases of the nervous system

VII: Diseases of the eye and adnexia

VIII: Diseases of the ear and mastoid process

IX: Diseases of the circulatory system

X: Diseases of the respiratory system

XI: K00-K95', 'Diseases of the digestive system'

XII: Diseases of the skin and subcutaneous tissue

XIII: Diseases of the musculoskeletal system and connective tissue

XIV: Diseases of the genitourinary system

XV: Pregnancy, childbirth and the puerperium

XVI: Certain conditions originating in the perinatal period

XVII: Congenital malformations, deformations and chromosomal abnormalities

XVIII: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

XIX: Injury, poisoning and certain other consequences of external causes

XX: External causes of morbidity and mortality

XXI: Factors influencing health status and contact with health services

XXII: Codes for special purposes

Supplementary Box 2. Top 15 disease codes (in rank order) from all 30 topics of the 30-topic LDA model used in the DCF Model

DCF 1:
Rheumatoid arthritis
Chemotherapy session for neoplasm
Rheumatoid arthritis, unspecified (Site unspecified)
Chronic lymphocytic leukaemia
Diffuse non-Hodgkin's lymphoma - Large cell (diffuse)
Personal history of malignant neoplasms of lymphoid, haematopoietic and related tissues
Non-Hodgkin's lymphoma, unspecified type
Rheumatoid arthritis, unspecified (Multiple sites)
Polycythaemia vera
Personal history of chemotherapy for neoplastic disease
Follicular non-Hodgkin's lymphoma, unspecified

Agranulocytosis
Thrombocytopenia
Nonfamilial hypogammaglobulinaemia
Prophylactic immunotherapy

DCF 2:

Personal history of allergy to penicillin
Personal history of allergy to other drugs, medicaments and biological substances
Personal history of allergy to analgesic agent
Personal history of allergy to other antibiotic agents
Personal history of allergy, other than to drugs and biological substances
Hallux valgus (acquired)
Special screening examination for diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
Personal history of allergy to narcotic agent
Arthrosis (Ankle and foot)
Acquired deformities of fingers and toes - hammer toe(s) (acquired)
Acquired deformities of fingers and toes - deformities of toe(s) (acquired)
Hallux rigidus
Personal history of allergy to sulphonamides
Personal history of diseases of the genito-urinary system
Lesion of plantar nerve

DCF 3:

Disorders of iron metabolism
Multiple myeloma
Bronchiectasis
Acute myeloid leukaemia
Transplanted organ and tissue status
Personal history of infectious and parasitic diseases
Personal history of chemotherapy for neoplastic disease
Acquired absence of lung [part of]
Personal history of malignant neoplasm of trachea, bronchus and lung
Malignant neoplasm of bronchus and lung
Malignant neoplasm of bronchus and lung - upper lobe, bronchus or lung
Unspecified acute lower respiratory infection
Abnormal findings on diagnostic imaging of lung
Intrathoracic lymph nodes
Personal history of diseases of the respiratory system

DCF 4:

Diaphragmatic hernia without obstruction or gangrene
Gastro-oesophageal reflux disease without oesophagitis
Gastritis, unspecified
Gastro-oesophageal reflux disease with oesophagitis
Iron deficiency anaemia, unspecified
Oesophagitis
Barrett's oesophagus
Anaemia, unspecified
Personal history of diseases of the digestive system
Dyspepsia
Polyp of stomach and duodenum
Duodenitis
Ulcer of oesophagus
Other gastritis
Dysphagia

DCF 5:

Personal history of psychoactive substance abuse

Mental and behavioural disorders due to use of tobacco - Harmful use
Chronic obstructive pulmonary disease, unspecified
Problems related to lifestyle - Tobacco use
Problems related to lifestyle - Alcohol use
Lobar pneumonia, unspecified
Chronic obstructive pulmonary disease with acute lower respiratory infection
Peripheral vascular disease, unspecified
Emphysema, unspecified
Pneumonia, unspecified
Pleural effusion, not elsewhere classified
Abdominal aortic aneurysm, without mention of rupture
Unspecified acute lower respiratory infection
Pulmonary collapse
Chronic obstructive pulmonary disease with acute exacerbation, unspecified

DCF 6:

Malignant neoplasm of prostate
Acquired absence of genital organ(s)
Personal history of malignant neoplasm of genital organs
Personal history of irradiation
Malignant neoplasm of ovary
Malignant neoplasm of corpus uteri - endometrium
Family history of malignant neoplasm of genital organs
Other specified abnormal findings of blood chemistry
Fitting and adjustment of urinary device
Personal history of chemotherapy for neoplastic disease
Intrapelvic lymph nodes
Myelodysplastic syndrome, unspecified
Secondary malignant neoplasm of bone and bone marrow
Naevus, nonneoplastic
Other hypertrophic disorders of skin

DCF 7:

Chest pain, unspecified
Radiotherapy session
Calculus of ureter
Internal haemorrhoids with other complications
Benign lipomatous neoplasm of skin and subcutaneous tissue of trunk
Dyspnoea
Viral infection, unspecified
Unspecified renal colic
Benign lipomatous neoplasm of skin and subcutaneous tissue of limbs
Observation for suspected myocardial infarction
Lymphoedema, not elsewhere classified
Bilateral inguinal hernia, without obstruction or gangrene
R07 Pain in throat and chest - R07.2 Precordial pain
R04 Haemorrhage from respiratory passages - R04.0 Epistaxis
Cardiac murmur, unspecified

DCF 8:

Personal history of long-term (current) use of anticoagulants
Atrial fibrillation and flutter
Personal history of diseases of the circulatory system
Atrial fibrillation and atrial flutter, unspecified
Presence of cardiac pacemaker
Congestive heart failure
Paroxysmal atrial fibrillation
Pulmonary embolism without mention of acute cor pulmonale
Aortic (valve) stenosis

Left ventricular failure
Cardiomegaly
Mitral (valve) insufficiency
Supraventricular tachycardia
Phlebitis and thrombophlebitis of other deep vessels of lower extremities
Left bundle-branch block, unspecified

DCF 9:
Malignant neoplasm of breast
Chemotherapy session for neoplasm
Personal history of malignant neoplasm of breast
Axillary and upper limb lymph nodes
Acquired absence of breast(s)
Malignant neoplasm of breast - upper-outer quadrant of breast
Secondary malignant neoplasm of bone and bone marrow
Family history of malignant neoplasm of breast
Personal history of irradiation
Personal history of chemotherapy for neoplastic disease
Follow-up care involving plastic surgery of breast
Lymph nodes of head, face and neck
Malignant neoplasm of breast - upper-inner quadrant of breast
Secondary malignant neoplasm of lung
Carcinoma in situ of breast - intraductal carcinoma in situ

DCF 10:
Presence of orthopaedic joint implants
Gonarthrosis, unspecified
Coxarthrosis, unspecified
Arthrosis, unspecified
Obesity, unspecified
Polyarthrosis, unspecified
Surgical operation with implant of artificial internal device
Arthritis, unspecified
Other primary gonarthrosis
Arthrosis, unspecified (Site unspecified)
Derangement of meniscus due to old tear or injury
Carpal tunnel syndrome
Other primary coxarthrosis
Mechanical complication of internal joint prosthesis
Derangement of meniscus due to old tear or injury

DCF 11:
Hypothyroidism, unspecified
Multiple sclerosis
Dental caries, unspecified
Postprocedural hypothyroidism
Thyrotoxicosis, unspecified
Retained dental root
Special screening examination for other specified diseases and disorders
Hypopituitarism
Thyrotoxicosis with diffuse goitre
Vitamin B12 deficiency anaemia due to intrinsic factor deficiency
Benign neoplasm of other and unspecified endocrine glands - pituitary gland
Benign neoplasm of breast
Malignant neoplasm of thyroid gland
Primary adrenocortical insufficiency
Other benign neoplasms of skin - skin of eyelid, including canthus

DCF 12:

Low back pain
Nerve root and plexus compressions in intervertebral disk disorders
Spinal stenosis (Lumbar region)
Lumbar and other intervertebral disk disorders with radiculopathy
Other chemotherapy
Other specified intervertebral disk degeneration
Other spondylosis (Cervical region)
Dorsalgia, unspecified
Other specified intervertebral disk displacement
Palmar fascial fibromatosis [Dupuytren]
Low back pain (Lumbar region)
Other spondylosis (Lumbar region)
Personal history of diseases of the musculoskeletal system and connective tissue
Sciatica
Cervicalgia

DCF 13:

Personal history of malignant neoplasm of urinary tract
Excessive and frequent menstruation with regular cycle
Leiomyoma of uterus, unspecified
Postmenopausal bleeding
Malignant neoplasm of bladder
Calculus of kidney
Polyp of corpus uteri
Personal history of diseases of the genito-urinary system
Follow-up examination after surgery for malignant neoplasm
Other and unspecified ovarian cysts
Acquired absence of kidney
Polyp of cervix uteri
Incomplete uterovaginal prolapse
pelvic peritoneal adhesions
Excessive and frequent menstruation with irregular cycle

DCF 14:

Other chemotherapy
Raynaud's syndrome
Other interstitial pulmonary diseases with fibrosis
Sicca syndrome [Sjogren]
Polyneuropathy, unspecified
Systemic lupus erythematosus, unspecified
Interstitial cystitis (chronic)
Melanocytic naevi of other and unspecified parts of face
Sarcoidosis, unspecified
Ankylosing spondylitis
Other inflammatory polyneuropathies
Other disorders of arteries
Other specified polyneuropathies
Other chronic pain
Wegener's granulomatosis

DCF 15:

Other and unspecified abdominal pain
Nausea and vomiting
Headache
Non-infective gastro-enteritis and colitis, unspecified
Change in bowel habit
Constipation
Other chest pain
Pain localised to other parts of lower abdomen

Syncope and collapse
Migraine, unspecified
Dizziness and giddiness
Abnormal weight loss
Pain localised to upper abdomen
Hearing loss, unspecified
Gastro-oesophageal reflux disease without oesophagitis

DCF 16:

Hyperplasia of prostate
Unilateral or unspecified inguinal hernia, without obstruction or gangrene
Personal history of malignant neoplasms of other organs and systems
Unspecified haematuria
Other malignant neoplasms of skin and unspecified parts of face
Other specified disorders of bladder
Retention of urine
Urethral stricture, unspecified
Personal history of diseases of the genito-urinary system
Disorder of skin and subcutaneous tissue, unspecified
Trichilemmal cyst
Polyuria
Urinary tract infection, site not specified
Seborrhoeic keratosis
Other and unspecified symptoms and signs involving the urinary system

DCF 17:

Personal history of malignant neoplasm of digestive organs
Acquired absence of other parts of digestive tract
Chemotherapy session for neoplasm
Malignant neoplasm of rectum
Intestinal bypass and anastomosis status
Intra-abdominal lymph nodes
Malignant neoplasm of colon - sigmoid colon
Ileostomy status
Secondary malignant neoplasm of liver
0 Follow-up examination after surgery for malignant neoplasm
5 Arthropathic psoriasis
Personal history of chemotherapy for neoplastic disease
Colostomy status
Malignant neoplasm of colon - colon
Malignant neoplasm of colon - caecum

DCF 18:

Personal history of diseases of the circulatory system
Urinary tract infection, site not specified
Personal history of diseases of the nervous system and sense organs
Acute renal failure, unspecified
Tendency to fall, not elsewhere classified
Physical therapy
Parkinson's disease
Hypo-osmolality and hyponatraemia
Other and unspecified abnormalities of gait and mobility
Constipation
Occupational therapy and vocational rehabilitation, not elsewhere classified
Escherichia coli [E. coli] as the cause of diseases
Hemiplegia, unspecified
Volume depletion
Disorientation, unspecified

DCF 19:
I10 Essential (primary) hypertension
hypercholesterolaemia
Personal history of long-term (current) use of other medicaments
3 Chronic kidney disease, stage 3
Gout, unspecified
Obesity, unspecified
Personal history of diseases of the nervous system and sense organs
Hyperlipidaemia, unspecified
Gout, unspecified (Site unspecified)
Other specified abnormal findings of blood chemistry
Arthrosis, unspecified
Arthritis, unspecified
Gastro-oesophageal reflux disease without oesophagitis
Procedure not carried out because of contraindication
Personal history of psychoactive substance abuse

DCF 20:
Diverticular disease of large intestine without perforation or abscess
Polyp of colon
Diverticular disease of intestine, part unspecified, without perforation or abscess
Family history of malignant neoplasm of digestive organs
Unspecified haemorrhoids without complication
Personal history of diseases of the digestive system
Special screening examination for neoplasm of intestinal tract
Haemorrhage of anus and rectum
Personal history of other neoplasms
Rectal polyp
Sigmoid colon
Gastro-intestinal haemorrhage, unspecified
Follow-up examination after surgery for other conditions
Haemorrhoids, unspecified
Change in bowel habit

DCF 21:
Cataract, unspecified
Senile nuclear cataract
Presence of intraocular lens
Glaucoma
Degeneration of macula and posterior pole
Psoriasis, unspecified
Sterilisation
Senile cataract
Senile incipient cataract
6 Personal history of diseases of the nervous system and sense organs
Primary open-angle glaucoma
Myopia
Family history of eye and ear disorders
After-cataract
Arthritis, unspecified

DCF 22:
Depressive episode
Anxiety disorder
Mental and behavioural disorders due to use of alcohol - harmful use
Cellulitis of other parts of limb
Mental and behavioural disorders due to use of alcohol - dependence syndrome
Bipolar affective disorder, unspecified
Mixed anxiety and depressive disorder

Personal history of self-harm
Unknown and unspecified causes of morbidity
Poisoning by nonopioid analgesics, antipyretics and antirheumatics: 4-Aminophenol derivatives
Mental and behavioural disorders due to use of tobacco - harmful use
Personal history of other mental and behavioural disorders
Mental and behavioural disorders due to use of alcohol - withdrawal state
Schizophrenia
Mental and behavioural disorders due to use of alcohol - acute intoxication

DCF 23:

Non-insulin-dependent diabetes mellitus
apnoea
Insulin-dependent diabetes mellitus
Obesity, unspecified
Diabetic retinopathy
Unspecified diabetes mellitus
Ophthalmic complications of non-insulin-dependent diabetes mellitus
Ulcer of lower limb
Other obesity
Cellulitis of other parts of limb
Acute renal failure
Hypoglycaemia
Chronic renal failure
9 Iron deficiency anaemia, unspecified
Neurological complications of non-insulin-dependent diabetes mellitus

DCF 24:

Extracorporeal dialysis
Chronic kidney disease, stage 5
Chronic renal failure, unspecified
End-stage renal disease
Polymyalgia rheumatica
Kidney transplant status
Dependence on renal dialysis
Hypertensive renal disease with renal failure
Other and unspecified cirrhosis of liver
Chronic kidney disease, stage 4
Portal hypertension
Alcoholic cirrhosis of liver
Unspecified renal failure
Oesophageal varices without bleeding
Ascites

DCF 25:

Angina pectoris
Atherosclerotic heart disease
Chronic IHD
Old myocardial infarction
Presence of coronary angioplasty implant and graft
Personal history of long-term use of other medicaments
Pure hypercholesterolaemia
Presence of aortocoronary bypass graft
Family history of IHD
Other forms of chronic ischaemic heart disease
Personal history of psychoactive substance abuse
Chest pain
Essential (primary) hypertension
Unstable angina
Hyperlipidaemia, unspecified

DCF 26:
Osteoporosis, unspecified
Epilepsy, unspecified
Irritable bowel syndrome without diarrhoea
Crohn's disease, unspecified
Osteoporosis, unspecified (Site unspecified)
Coeliac disease
Fracture of lower end of radius (closed)
Vitamin D deficiency, unspecified
Other specified disorders of bone density and structure
Arthrosis, unspecified
Deficiency of other specified B group vitamins
Fracture of neck of femur (closed)
Crohn's disease of large intestine
Crohn's disease of small intestine
Primary hyperparathyroidism

DCF 27:
Carpal tunnel syndrome
Fibromyalgia
Exposure to unspecified factor - unspecified place
Follow-up care involving removal of fracture plate and other internal fixation device
Unspecified fall
Fractures of other parts of lower leg (closed)
wound of finger(s) without damage to nail
Unspecified injury of head
Fall on and from stairs and steps at home
Fall on same level from slipping, tripping and stumbling at home
Fracture of upper end of humerus (closed)
Postviral fatigue syndrome
Open wound of other parts of head
Fall on same level from slipping, tripping and stumbling, unspecified place
Multiple fractures of ribs (closed)

DCF 28:
Personal history of diseases of the digestive system
Calculus of gallbladder without cholecystitis
Acquired absence of other parts of digestive tract
Calculus of gallbladder with other cholecystitis
following a procedure, not elsewhere classified
Removal of other organ (partial) (total)
Pain localised to upper abdomen
Abnormal results of liver function studies
Other surgical procedures
Peritoneal adhesions
Fatty (change of) liver, not elsewhere classified
Unspecified abdominal pain
Haemorrhage and haematoma complicating a procedure, not elsewhere classified
Calculus of bile duct without cholangitis or cholecystitis
Ventral hernia without obstruction or gangrene

DCF 29:
Single live birth
Varicose veins of lower extremities without ulcer or inflammation
Ulcerative colitis, unspecified
degree perineal laceration during delivery
First degree perineal laceration during delivery

Other disorders of synovium and tendon: ganglion
Other specified pregnancy-related conditions
Spontaneous vertex delivery
Labour and delivery complicated by foetal heart rate anomaly
Missed abortion
Maternal care for other specified foetal problems
Prolonged second stage (of labour)
Maternal care due to uterine scar from previous surgery
Prolonged pregnancy
Other and unspecified abdominal pain

DCF 30:

Asthma, unspecified
Unknown and unspecified causes of morbidity
Stress incontinence
Female genital prolapse: cystocele
Impingement syndrome of shoulder
Female genital prolapse: rectocele
Rotator cuff syndrome
Dermatitis, unspecified
Arthrosis, unspecified (Shoulder region)
Unspecified acute lower respiratory infection
Predominantly allergic asthma
Adhesive capsulitis of shoulder
Obesity
Gastro-oesophageal reflux disease without oesophagitis
Other shoulder lesions