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

Aldo Faisal (a.faisal@imperial.ac.uk)
Imperial College London https://orcid.org/0000-0003-0813-7207

Erwann Le Lannou
Imperial College London https://orcid.org/0000-0003-2236-1862

Benjamin Post
Imperial College London https://orcid.org/0000-0003-0337-0827

Shlomi Haar
Imperial College London https://orcid.org/0000-0003-2213-6585

Stephen Brett
Imperial College London https://orcid.org/0000-0003-4545-8413

Balasundaram Kadirvelu
Imperial College London

Article

Keywords: COVID-19, mortality prediction, comorbidities

Posted Date: April 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-374482/v1

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
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\(^6,8-10\), 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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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. 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. 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), 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 assess 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; 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. All participants were recruited between 2006 and 2010, and all consented to have their health followed. 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.

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. The previous hospital medical information of the study participants was obtained from linkage to the HES database (data
warehouse containing records of all admissions, emergency room attendances and outpatients appointments at NHS hospitals in England)\textsuperscript{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.

\textit{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\textsuperscript{45}).

\textit{Model Development}

The performance of machine learning methods is heavily dependent on the selection of features\textsuperscript{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\textsuperscript{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\textsuperscript{17} or listed by the Centers for Disease Control and Prevention (CDC) as defining ‘people at increased risk for severe illness’\textsuperscript{18}. This was further complemented with other emerging risk factors for severe outcomes of COVID-19 (such as raised blood pressure)\textsuperscript{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\textsuperscript{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. 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.

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, \( \beta_k \sim \text{Dirichlet}(\eta) \). Thus, each code \( c_{p,n} \) has a given probability \( \beta_k \) in each topic \( k \). Given the topics, LDA then generates for each patient \( p \) a distribution over topics \( \theta_p \sim \text{Dirichlet}(\alpha) \). 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 \( \theta \) and the code topic distribution \( \beta \).

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 $\theta$ and $\beta$ ($\alpha$ and $\eta$) 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\textsuperscript{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\textsuperscript{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\textsuperscript{55}. The models’ hyper-parameters\textsuperscript{56–59} were determined via grid search with AUC as the response as it yields a finer evaluation than commonly used error rates\textsuperscript{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.

\textit{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\textsuperscript{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**


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*


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


1022 (2003).


**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 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

<table>
<thead>
<tr>
<th>Demographics</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Age in years</td>
</tr>
<tr>
<td>• Townsend deprivation index</td>
</tr>
<tr>
<td>• Ethnicity</td>
</tr>
<tr>
<td>Lifestyle</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>• Smoking status</td>
</tr>
<tr>
<td>o Current or previous smoker</td>
</tr>
<tr>
<td>o Never smoked at UK Biobank baseline</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>COVID-19-related death (n = 349)</th>
<th>Rest of the Cohort (n = 2150)</th>
<th>OR (95% CI)</th>
<th>Fully adjusted OR (95% CI)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Count (Percentage)</th>
<th>Median Age (95% CI)</th>
<th>Median Age (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 50-59</td>
<td>14 (4.01%)</td>
<td>619 (28.79%)</td>
<td>1</td>
</tr>
<tr>
<td>Age 60-69</td>
<td>55 (15.76%)</td>
<td>537 (24.98%)</td>
<td>4.53 (2.49 - 8.23)</td>
</tr>
<tr>
<td>Age 70-79</td>
<td>234 (67.05%)</td>
<td>894 (41.58%)</td>
<td>11.57 (6.68 - 20.04)</td>
</tr>
<tr>
<td>Age 80+</td>
<td>46 (13.18%)</td>
<td>100 (4.65%)</td>
<td>20.34 (10.78 - 38.36)</td>
</tr>
<tr>
<td>Sex</td>
<td>221 (63.32%)</td>
<td>1063 (49.44%)</td>
<td>1.77 (1.40 - 2.23)</td>
</tr>
<tr>
<td>Deprivation Q1</td>
<td>64 (18.34%)</td>
<td>389 (18.09%)</td>
<td>1</td>
</tr>
<tr>
<td>Deprivation Q2</td>
<td>69 (19.77%)</td>
<td>427 (19.86%)</td>
<td>0.98 (0.68 - 1.42)</td>
</tr>
<tr>
<td>Deprivation Q3</td>
<td>80 (22.92%)</td>
<td>567 (26.37%)</td>
<td>0.86 (0.60 - 1.22)</td>
</tr>
<tr>
<td>Depreviation Q4</td>
<td>136 (38.97%)</td>
<td>767 (35.67%)</td>
<td>1.08 (0.78 - 1.49)</td>
</tr>
<tr>
<td>Ethnicity (non-white)</td>
<td>35 (10.03%)</td>
<td>243 (11.30%)</td>
<td>0.87 (0.60 - 1.27)</td>
</tr>
<tr>
<td>Obesity</td>
<td>67 (19.20%)</td>
<td>279 (12.98%)</td>
<td>1.59 (1.19 - 2.14)</td>
</tr>
<tr>
<td>Smoking status</td>
<td>240 (68.77%)</td>
<td>1326 (61.67%)</td>
<td>1.37 (1.07 - 1.74)</td>
</tr>
<tr>
<td>Organ transplant</td>
<td>3 (0.86%)</td>
<td>19 (0.88%)</td>
<td>0.97 (0.29 - 3.30)</td>
</tr>
<tr>
<td>Cancer</td>
<td>97 (27.79%)</td>
<td>432 (20.09%)</td>
<td>1.53 (1.18 - 1.98)</td>
</tr>
<tr>
<td>Condition</td>
<td>Count</td>
<td>Percentage</td>
<td>Hazard Ratio (95% CI)</td>
</tr>
<tr>
<td>-------------------------------------------------------------</td>
<td>-------</td>
<td>------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Severe respiratory disease</td>
<td>69</td>
<td>19.77%</td>
<td>1.99 (1.48 - 2.67)</td>
</tr>
<tr>
<td>Rare disease or inborn errors of metabolism</td>
<td>6</td>
<td>1.72%</td>
<td>2.07 (0.82 - 5.26)</td>
</tr>
<tr>
<td>Congenital heart defects</td>
<td>4</td>
<td>1.15%</td>
<td>2.25 (0.71 - 7.12)</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>206</td>
<td>59.03%</td>
<td>2.63 (2.09 - 3.31)</td>
</tr>
<tr>
<td>Chronic, non-severe respiratory disease</td>
<td>52</td>
<td>14.90%</td>
<td>1.07 (0.78 - 1.47)</td>
</tr>
<tr>
<td>Chronic kidney disease</td>
<td>59</td>
<td>16.91%</td>
<td>2.10 (1.53 - 2.88)</td>
</tr>
<tr>
<td>Cardiac disease</td>
<td>71</td>
<td>20.34%</td>
<td>2.35 (1.74 - 3.16)</td>
</tr>
<tr>
<td>Liver disease</td>
<td>42</td>
<td>12.03%</td>
<td>1.53 (1.07 - 2.19)</td>
</tr>
<tr>
<td>Chronic neurological conditions</td>
<td>77</td>
<td>22.06%</td>
<td>2.55 (1.91 - 3.40)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>100</td>
<td>28.65%</td>
<td>2.04 (1.58 - 2.65)</td>
</tr>
<tr>
<td>Conditions or treatments that predispose to infection</td>
<td>50</td>
<td>14.33%</td>
<td>2.12 (1.51 - 2.98)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>238</td>
<td>68.19%</td>
<td>2.75 (2.16 - 3.50)</td>
</tr>
</tbody>
</table>
Table 2. Performance of all prediction models under consideration

<table>
<thead>
<tr>
<th>Predictor Set</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC-ROC</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.709</td>
</tr>
<tr>
<td>EHCF - Model</td>
<td>0.734</td>
</tr>
<tr>
<td>DCF - Model</td>
<td>0.730</td>
</tr>
</tbody>
</table>

Supplementary Material

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

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Certain infectious and parasitic diseases’</td>
</tr>
<tr>
<td>II</td>
<td>Neoplasms</td>
</tr>
<tr>
<td>III</td>
<td>Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism</td>
</tr>
<tr>
<td>IV</td>
<td>Endocrine, nutritional and metabolic diseases</td>
</tr>
<tr>
<td>V</td>
<td>Mental and behavioural disorders</td>
</tr>
</tbody>
</table>
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
<table>
<thead>
<tr>
<th>Agranulocytosis</th>
<th>Thrombocytopenia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonfamilial hypogammaglobulinaemia</td>
<td>Prophylactic immunotherapy</td>
</tr>
</tbody>
</table>

**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                             |

<p>| <strong>DCF 12:</strong>                                      |</p>
<table>
<thead>
<tr>
<th>Low back pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nerve root and plexus compressions in intervertebral disk disorders</td>
</tr>
<tr>
<td>Spinal stenosis (Lumbar region)</td>
</tr>
<tr>
<td>Lumbar and other intervertebral disk disorders with radiculopathy</td>
</tr>
<tr>
<td>Other chemotherapy</td>
</tr>
<tr>
<td>Other specified intervertebral disk degeneration</td>
</tr>
<tr>
<td>Other spondylosis (Cervical region)</td>
</tr>
<tr>
<td>Dorsalgia, unspecified</td>
</tr>
<tr>
<td>Other specified intervertebral disk displacement</td>
</tr>
<tr>
<td>Palmar fascial fibromatosis [Dupuytren]</td>
</tr>
<tr>
<td>Low back pain (Lumbar region)</td>
</tr>
<tr>
<td>Other spondylosis (Lumbar region)</td>
</tr>
<tr>
<td>Personal history of diseases of the musculoskeletal system and connective tissue</td>
</tr>
<tr>
<td>Sciatica</td>
</tr>
<tr>
<td>Cervicalgia</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>DCF 13:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal history of malignant neoplasm of urinary tract</td>
</tr>
<tr>
<td>Excessive and frequent menstruation with regular cycle</td>
</tr>
<tr>
<td>Leiomyoma of uterus, unspecified</td>
</tr>
<tr>
<td>Postmenopausal bleeding</td>
</tr>
<tr>
<td>Malignant neoplasm of bladder</td>
</tr>
<tr>
<td>Calculus of kidney</td>
</tr>
<tr>
<td>Polyp of corpus uteri</td>
</tr>
<tr>
<td>Personal history of diseases of the genito-urinary system</td>
</tr>
<tr>
<td>Follow-up examination after surgery for malignant neoplasm</td>
</tr>
<tr>
<td>Other and unspecified ovarian cysts</td>
</tr>
<tr>
<td>Acquired absence of kidney</td>
</tr>
<tr>
<td>Polyp of cervix uteri</td>
</tr>
<tr>
<td>Incomplete uterovaginal prolapse</td>
</tr>
<tr>
<td>pelvic peritoneal adhesions</td>
</tr>
<tr>
<td>Excessive and frequent menstruation with irregular cycle</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>DCF 14:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Other chemotherapy</td>
</tr>
<tr>
<td>Raynaud's syndrome</td>
</tr>
<tr>
<td>Other interstitial pulmonary diseases with fibrosis</td>
</tr>
<tr>
<td>Sicca syndrome [Sjögren]</td>
</tr>
<tr>
<td>Polyneuropathy, unspecified</td>
</tr>
<tr>
<td>Systemic lupus erythematosus, unspecified</td>
</tr>
<tr>
<td>Interstitial cystitis (chronic)</td>
</tr>
<tr>
<td>Melanocytic naevi of other and unspecified parts of face</td>
</tr>
<tr>
<td>Sarcoïdosis, unspecified</td>
</tr>
<tr>
<td>Ankylosing spondylitis</td>
</tr>
<tr>
<td>Other inflammatory polyneuropathies</td>
</tr>
<tr>
<td>Other disorders of arteries</td>
</tr>
<tr>
<td>Other specified polyneuropathies</td>
</tr>
<tr>
<td>Other chronic pain</td>
</tr>
<tr>
<td>Wegener's granulomatosis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>DCF 15:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Other and unspecified abdominal pain</td>
</tr>
<tr>
<td>Nausea and vomiting</td>
</tr>
<tr>
<td>Headache</td>
</tr>
<tr>
<td>Non-infective gastro-enteritis and colitis, unspecified</td>
</tr>
<tr>
<td>Change in bowel habit</td>
</tr>
<tr>
<td>Constipation</td>
</tr>
<tr>
<td>Other chest pain</td>
</tr>
<tr>
<td>Pain localised to other parts of lower abdomen</td>
</tr>
</tbody>
</table>
### 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
- Arthritis, 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
<table>
<thead>
<tr>
<th><strong>Personal history of self-harm</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unknown and unspecified causes of morbidity</strong></td>
</tr>
<tr>
<td><strong>Poisoning by nonopioid analgesics, antipyretics and antirheumatics: 4-Aminophenol derivatives</strong></td>
</tr>
<tr>
<td><strong>Mental and behavioural disorders due to use of tobacco - harmful use</strong></td>
</tr>
<tr>
<td><strong>Personal history of other mental and behavioural disorders</strong></td>
</tr>
<tr>
<td><strong>Mental and behavioural disorders due to use of alcohol - withdrawal state</strong></td>
</tr>
<tr>
<td><strong>Schizophrenia</strong></td>
</tr>
<tr>
<td><strong>Mental and behavioural disorders due to use of alcohol - acute intoxication</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>DCF 23:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Non-insulin-dependent diabetes mellitus</strong></td>
</tr>
<tr>
<td><strong>apnoea</strong></td>
</tr>
<tr>
<td><strong>Insulin-dependent diabetes mellitus</strong></td>
</tr>
<tr>
<td><strong>Obesity, unspecified</strong></td>
</tr>
<tr>
<td><strong>Diabetic retinopathy</strong></td>
</tr>
<tr>
<td><strong>Unspecified diabetes mellitus</strong></td>
</tr>
<tr>
<td><strong>Ophthalmic complications of non-insulin-dependent diabetes mellitus</strong></td>
</tr>
<tr>
<td><strong>Ulcer of lower limb</strong></td>
</tr>
<tr>
<td><strong>Other obesity</strong></td>
</tr>
<tr>
<td><strong>Cellulitis of other parts of limb</strong></td>
</tr>
<tr>
<td><strong>Acute renal failure</strong></td>
</tr>
<tr>
<td><strong>Hypoglycaemia</strong></td>
</tr>
<tr>
<td><strong>Chronic renal failure</strong></td>
</tr>
<tr>
<td><strong>9 Iron deficiency anaemia, unspecified</strong></td>
</tr>
<tr>
<td><strong>Neurological complications of non-insulin-dependent diabetes mellitus</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>DCF 24:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Extracorporeal dialysis</strong></td>
</tr>
<tr>
<td><strong>Chronic kidney disease, stage 5</strong></td>
</tr>
<tr>
<td><strong>Chronic renal failure, unspecified</strong></td>
</tr>
<tr>
<td><strong>End-stage renal disease</strong></td>
</tr>
<tr>
<td><strong>Polymyalgia rheumatica</strong></td>
</tr>
<tr>
<td><strong>Kidney transplant status</strong></td>
</tr>
<tr>
<td><strong>Dependence on renal dialysis</strong></td>
</tr>
<tr>
<td><strong>Hypertensive renal disease with renal failure</strong></td>
</tr>
<tr>
<td><strong>Other and unspecified cirrhosis of liver</strong></td>
</tr>
<tr>
<td><strong>Chronic kidney disease, stage 4</strong></td>
</tr>
<tr>
<td><strong>Portal hypertension</strong></td>
</tr>
<tr>
<td><strong>Alcoholic cirrhosis of liver</strong></td>
</tr>
<tr>
<td><strong>Unspecified renal failure</strong></td>
</tr>
<tr>
<td><strong>Oesophageal varices without bleeding</strong></td>
</tr>
<tr>
<td><strong>Ascites</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>DCF 25:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Angina pectoris</strong></td>
</tr>
<tr>
<td><strong>Atherosclerotic heart disease</strong></td>
</tr>
<tr>
<td><strong>Chronic IHD</strong></td>
</tr>
<tr>
<td><strong>Old myocardial infarction</strong></td>
</tr>
<tr>
<td><strong>Presence of coronary angioplasty implant and graft</strong></td>
</tr>
<tr>
<td><strong>Personal history of long-term use of other medicaments</strong></td>
</tr>
<tr>
<td><strong>Pure hypercholesterolaemia</strong></td>
</tr>
<tr>
<td><strong>Presence of aortocoronary bypass graft</strong></td>
</tr>
<tr>
<td><strong>Family history of IHD</strong></td>
</tr>
<tr>
<td><strong>Other forms of chronic ischaemic heart disease</strong></td>
</tr>
<tr>
<td><strong>Personal history of psychoactive substance abuse</strong></td>
</tr>
<tr>
<td><strong>Chest pain</strong></td>
</tr>
<tr>
<td><strong>Essential (primary) hypertension</strong></td>
</tr>
<tr>
<td><strong>Unstable angina</strong></td>
</tr>
<tr>
<td><strong>Hyperlipidaemia, unspecified</strong></td>
</tr>
</tbody>
</table>
**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
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.
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).

DCF 24: “Cardiovascular”
- 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

DCF 22: “Diabetes”
- Non-insulin-dependent diabetes mellitus
- Sleep apnoea
- Insulin-dependent diabetes mellitus
- Obesity
- Diabetic retinopathy
- Unspecified diabetes mellitus
- Ophthalmic complications of diabetes mellitus
- Ulcer of lower limb
- Other obesity
- Cellulitis of other parts of limb
- Acute renal failure
- Hypoglycaemia
- Chronic renal failure

Figure 3
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.