medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Rapid Epidemiological Analysis of Comorbidities and Treatments as risk factors for COVID-19 in Scotland (REACT-SCOT): a population-based case-control study

Paul M McKeigue [1 3], Amanda Weir [3], Jen Bishop [3], Stuart J McGurnaghan [2], Sharon Kennedy [7], David McAllister [4 3], Chris Robertson [5 3], Rachael Wood [7], Nazir Lone [1], Janet Murray [3], Thomas M Caparrotta [2], Alison Smith-Palmer [3], David Goldberg [3], Jim McMenamin [3], Colin Ramsay [3], Sharon Hutchinson [6 3], Helen M Colhoun [2 3]

1 Usher Institute, College of Medicine and Veterinary Medicine, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, Scotland. PM - Professor of Genetic Epidemiology and Statistical Genetics. NL - Clinical Senior Lecturer in Critical Care
2 Institute of Genetics and Molecular Medicine, College of Medicine and Veterinary Medicine, University of Edinburgh, Western General Hospital Campus, Crewe Road, Edinburgh EH4 2XUC, Scotland. HC - Axa Chair in Medical Informatics and Epidemiology. TC - Sir George Alberti Doctoral Fellow in Pharmacoepidemiology.
3 Public Health Scotland, Meridian Court, 5 Cadogan Street, Glasgow G2 6QE
4 Institute of Health and Wellbeing, University of Glasgow, 1 Lilybank Gardens, Glasgow G12 8RZ. DM - Wellcome Trust Intermediate Clinical Fellow and Beit Fellow
5 Department of Mathematics and Statistics, University of Strathclyde, 16 Richmond Street, Glasgow G1 1XQ. CR - Professor of Public Health Epidemiology
6 School of Health and Life Sciences, Glasgow Caledonian University. SH - Professor of Epidemiology and Population Health
7 NHS Information Services Division (Public Health Scotland), Gyle Square, 1 South Gyle Crescent, Edinburgh, EH12 9EB. RW - Consultant in Maternal and Child Health.

On behalf of Public Health Scotland COVID-19 Health Protection Study Group [1]

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Abstract

Background– The objectives of this study were to identify risk factors for severe COVID-19 and to lay the basis for risk stratification based on demographic data and health records.

Methods – The design was a matched case-control study. Severe cases were all those with a positive nucleic acid test for SARS-CoV-2 in the national database who had entered a critical care unit or died within 28 days of the first positive test. Ten controls per case matched for sex, age and primary care practice were selected from the population register. All diagnostic codes from the past five years of hospitalisation records and all drug codes from prescriptions dispensed during the past nine months were extracted. Rate ratios for severe COVID-19 were estimated by conditional logistic regression.

Findings – In a logistic regression using the age-sex distribution of the national population, the odds ratios were 2.26 for a 10-year increase in age and 1.86 for male sex. In the case-control analysis, the strongest risk factor was residence in a care home, with rate ratio (95% CI) 14.9 (12.7, 17.5). Univariate rate ratios (95% CIs) for conditions listed by public health agencies as conferring high risk were 4.88 (3.26, 7.31) for Type 1 diabetes, 2.58 (2.30, 2.88) for Type 2 diabetes, 2.40 (2.14, 2.70) for ischemic heart disease, 3.90 (3.52, 4.32) for other heart disease, 3.10 (2.81, 3.42) for chronic lower respiratory tract disease, 12.1 (8.4, 17.4) for chronic kidney disease, 5.5 (4.8, 6.2) for neurological disease, 4.70 (2.90, 7.62) for chronic liver disease and 4.11 (2.72, 6.21) for immune deficiency or suppression.

72% of cases and 35% of controls had at least one listed condition (50% of cases and 9% of controls under age 40). Severe disease was associated with encashment of at least one prescription in the past nine months and with at least one hospital admission in the past five years [rate ratios 16.6 (13.3, 20.6)] and 5.6 (5.0, 6.2) respectively] even after adjusting for the listed conditions. In those without listed conditions significant associations with severe disease were seen across many hospital diagnoses and drug categories. Age and sex provided 1.81 bits of information for discrimination. A model based on demographic variables, listed conditions, hospital diagnoses and prescriptions provided an additional 1.5 bits (C-statistic 0.839).

Conclusions – Along with older age and male sex, severe COVID-19 is strongly associated with past medical history across all age groups. Many comorbidities beyond the risk conditions designated by public health agencies contribute to this. A risk classifier that uses all the information available in health records, rather than only a limited set of conditions, will more accurately discriminate between low-risk and high-risk individuals who may require shielding until the epidemic is over.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Background

Case series from many countries have suggested that in those with severe COVID-19 the prevalence of diabetes and cardiovascular disease is higher than expected. For example in a large UK series the commonest co-morbidities were cardiac disease, diabetes, chronic pulmonary disease and asthma [1]. However there are also anecdotal reports of apparently healthy young persons succumbing to disease [2].

Quantification of the risk associated with characteristics and co-morbidities has been limited by the lack of comparisons with the background population [3–5]. Two recent studies in the UK have included population comparators and have reported associations of in hospital test positive persons and COVID-19 death in hospital with co-morbidities including diabetes, asthma and heart disease [6,7]. These studies have focused on conditions presumptively listed by public health agencies as increasing risk for COVID-19 based on case series data.

Here we examine the frequency of sociodemographic factors and these listed conditions in all people with severe COVID-19 disease in Scotland compared to matched controls from the general population. In those without listed conditions we report a systematic examination of the hospitalisation record and prescribing history in severe COVID-19 cases compared to controls. The objectives were to identify risk factors for severe COVID-19 and to lay the basis for risk stratification based on a predictive model.

Methods

Case definition

The Electronic Communication of Surveillance in Scotland (ECOSS) database captures all virology testing in all NHS laboratories nationally. Individuals testing positive for nucleic acid for SARS-CoV-2 up to 30 April 2020 in ECOSS were ascertained. Using the Community Health Index (CHI) identifier contained in ECOSS (the CHI number is a unique identifier used in all care systems in Scotland) linkage to other datasets was carried out. Hospital admissions from the time of testing were obtained from the RAPID database a daily return of current hospitalisations each day. Admissions to critical care were obtained from the Scottish Intensive Care Society and Audit Group (SICSAG) database that covers admissions to critical care [comprising adult intensive care units (ICUs), high dependency units (HDUs) and combined ICU / HDU units] across Scotland and has returned a daily census of patients in critical care from the beginning of the COVID-19 epidemic. Death registrations up to 4 May 2020 were obtained from linkage to the National Register of Scotland.

Severe or fatal COVID-19 was defined by a record of entering critical care in the SICSAG database, or death within 28 days of a positive nucleic acid test, regardless of the cause of death given on the death certificate. By restricting the case definition to those cases that were fatal or received critical care, we ensured complete ascertainment of all test-positive cases that were severe enough to have been fatal without critical care, whatever selection policies may have determined admission to hospital or entry to critical care.

Matched controls

For each test-positive case, we ascertained ten matched controls of the same sex, one-year age band and registered with the same primary care practice who were alive on the date of the first test in the case using the Community Health Index (CHI) database. As this is an incidence density sampling design, it is possible and correct for an

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

individual to appear in the dataset more than once, initially as a control and subsequently as a case.

Demographic data

Residence in a care home was ascertained from the CHI database. Socioeconomic status was assigned as the Scottish Index of Multiple Deprivation (SIMD), an indicator based on postal code. Ethnicity was assigned based on applying a name classification algorithm (ONOMAP) [8] to the names in the CHI database. For 54% of cases and 28% of controls self-assigned ethnicity, based on the categories used in the Census, had been recorded in Scottish Morbidity Records (SMR). Cross-tabulation of 28011 records for which both name classification and SMR records of ethnicity were available showed that the ONOMAP algorithm had sensitivity of 93% and specificity of 99.57% for classifying South Asian ethnicity, but misclassified most of those who identified as African, Caribbean or Black.

Morbidity and drug prescribing

For all cases and controls, ICD-10 diagnostic codes were extracted from the last five years of hospital discharge records in the Scottish Morbidity Record (SMR01), excluding records of discharges less than 25 days before testing positive for SARS-CoV-2 and using all codes on the discharge. Diagnostic coding under ICD chapters 5 (Mental, Behavioural and Neurodevelopmental) and 15 (Pregnancy) is incomplete as most psychiatric and maternity unit returns are not captured in SMR01. British National Formulary (BNF) drug codes were extracted from the last year of encashed prescriptions, excluding those encashed less than 25 days before testing positive for SARS-CoV-2. The BNF groups drugs by 2-digit chapter codes. For this analysis prescription codes from chapters 14 and above, mostly for dressings and appliances but also including vaccines were grouped as "Other".

We began by scoring a specific list of conditions that have been designated as risk conditions for COVID-19 by public health agencies [9]. A separate list of conditions designates "clinically extremely vulnerable" individuals who have been advised to shield themselves completely since early in the epidemic: this list includes solid organ transplant recipients, people receiving chemotherapy for cancer, and people with cystic fibrosis or leukaemia. We did not separately tabulate these conditions as we expected these individuals to be underrepresented among cases if shielding was adequate.

The eight listed conditions were scored based on diagnostic codes in any hospital discharge record during the last five years, or encashed prescription of a drug for which the only indications are in that group of diagnostic codes. The R script included as supplementary material contains the derivations of these variables from ICD-10 codes and BNF drug codes. Diagnosed cases of diabetes were identified through linkage to the national diabetes register (SCI-Diabetes), with a clinical classification of diabetes type as Type 1, Type 2 or Other/Unknown. Cases of diabetes diagnosed since the last update of the register were identified through discharge codes and drug codes, and assigned to the diabetes type Other/Unknown category.

Statistical methods

To estimate the relation of cumulative incidence and mortality from COVID-19 to age and sex, logistic regression models were fitted to the proportions of cases and non-cases in the Scottish population, using the estimated population of Scotland in mid-year 2019 which were available by one-year age group up to age 90 years. To allow for possible non-linearity of the relationship of the logit of risk to age, we also fitted generalized

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

additive models, implemented in the R function `gam::gam`, with default smoothing function.

For the case-control study, all estimates of associations with severe COVID-19 were based on conditional logistic regression, implemented as Cox regression in the R function `survival::clogit`. Among those cases and controls without any of the pre-defined conditions we then further examined associations of ICD-10 and BNF chapter with severe COVID-19. Where an exclusion criterion such as having a pre-defined condition was applied to cases this was also applied to controls as otherwise subsequent association estimates would be incorrect. Where the sample of cases and controls is restricted, this will generate strata that contain no cases but these strata will be ignored by the conditional logistic regression model as they do not contribute to the conditional likelihood. With incidence density sampling, the odds ratios in conditional logistic regression models are equivalent to rate ratios. Note that odds ratios in a matched case control study are based on the conditional likelihood and the unconditional odds ratios calculable from the frequencies of exposure in cases and controls will differ from these and should not be used [10]. Although matching on primary care practice will match to some extent for associated variables such as care home residence, socioeconomic disadvantage and prescribing practice, the effects of these variables are still estimated correctly by the conditional odds ratios but with less precision than in an unmatched study of the same size [10].

To construct risk prediction models, we used stepwise regression alternating between forward and backward steps to maximize the AIC, implemented in the R function `stats::step`. The performance of the risk prediction model in classifying cases versus non-cases of severe COVID-19 was examined by 4-fold cross-validation. We calculated the performance calculated over all test folds using the C-statistic but also using the "expected information for discrimination" Λ expressed in bits [11]. The use of bits (logarithms to base 2) to quantify information is standard in information theory: one bit can be defined as the quantity of information that halves the hypothesis space. Although readers may be unfamiliar with the expected information for discrimination Λ, it has several properties that make it more useful than the C-statistic for quantifying increments in the performance of a risk prediction model [11]. A key advantage of using Λ is that contributions of independent predictors can be added. Thus in this study we can add the predictive information from a logistic model of age and sex in the general population to the predictive information provided by other risk factors from the case-control study matched for age and sex.

Results

Incidence and mortality from severe COVID-19 in the Scottish population

Figure 1 shows the relationships of incidence and mortality rates to age for each sex separately. The relationship of mortality to age is almost exactly linear on a logit scale, and the lines for male and female mortality are almost parallel. In models that included age and sex as covariates, the odds ratio associated with a 10-year increase in age was 2.26 for all severe disease and 3.35 for fatal disease. The odds ratio associated with male sex was 1.86 for all severe disease and 1.87 for fatal disease. For severe cases as defined in this study, the sex differential is narrow up to about age 50 but widens between ages 50 and 70 years. Thus at younger ages the ratio of critical care admissions to total fatalities is higher in women than in men, but that at later ages the ratio of critical admissions to total fatalities is higher in men.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Risk factors

Sociodemographic factors

Table 1 shows univariate associations of demographic factors with severe disease. Residence in a care home was by far the strongest risk factor for severe disease. Higher risk of severe disease was also associated with socioeconomic deprivation. Associations with ethnicity are shown for the full dataset based on name classification and separately for the subset of cases and controls in whom ethnicity had been recorded in the Scottish Morbidity Record. With Whites as reference category, the rate ratio (95% CI) associated with South Asian ethnicity was 0.53 (0.37, 0.76) based on name classification and 0.81 (0.31, 2.10), based on the subset with SMR records. The numbers of cases in other non-White ethnic groups were too sparse to tabulate separately.

Factors derived from hospitalisation and prescribing records

Prevalence of the listed conditions in cases and controls by age band is shown in Table 2. 30 (50%) of the cases aged under 40 years had at least one listed condition, compared with only 53 (9%) of the controls. In those aged 75+ years 976 (80%) of the cases and 5172 (43%) of the controls had at least one listed condition. The proportion with at least one dispensed prescription was much higher in cases than in controls in each age group. Among those aged under 40 years, 50 (83%) of the cases and 305 (51%) of the controls had either a hospital admission in the last five years or a dispensed prescription in the last year.

Over all age groups, 1599 (72%) of severe cases and 7701 (35%) of controls had at least one of the listed conditions. As shown in Table 3, all the listed conditions were more frequent in cases than controls. The rate ratio associated with type 1 diabetes was higher than that for type 2 diabetes. The rate ratio was 2.40 (2.14, 2.70) for ischemic heart disease compared to 3.90 (3.52, 4.32) for the broad category "other heart disease". In multivariate analysis ischemic heart disease was not independently associated with severity whereas other heart disease remained strongly associated.

Supplementary Tables 8 to 10 examine these associations by age group, with the 0-39 and 40-59 year age bands combined. All listed conditions were associated with severe disease in each age band. In those aged under 60 years, the rate ratio was 9.8 (5.2, 18.4) for Type 1 diabetes and 5.4 (3.9, 7.5) for Type 2 diabetes. The multivariate analyses shown in Table 3 and 8 to 10 show that overall and within each age group dispensing of any prescription in the past year and any admission to hospital in the past five years were strongly and independently associated with severe disease even after adjusting for care home residence and listed conditions. Table 4 shows that in each age group the proportion of fatal cases who had not had either a hospital admission in the last five years or a dispensed prescription in the last year was very low.

Systematic analysis of diagnoses associated with severe disease

The association of severe COVID-19 with prior hospital admission was examined further by testing for association of hospitalisations at each ICD-10 chapter level with severe COVID-19, among those without any of the listed conditions. These results are shown in Table 5. In univariate analyses, almost all ICD-10 chapters, with the exception of Chapters 8 (ear) and Chapter 15 (pregnancy) were associated with increased risk of severe disease. Note that hospital diagnoses classified under the pregnancy chapter here are derived from admissions with pregnancy related medical conditions to non-obstetric units only, as obstetric returns are not in the SMR01 dataset. In a multivariate analysis the strongest associations were with diagnoses in ICD chapter 2 (neoplasms). Supplementary Table 11 extracts univariate associations with ICD-10 subchapters in

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

those without any listed conditions. This table is filtered to show only subchapters for which the univariate p-value is <0.001 and where there are at least 50 cases and controls with a diagnosis in this subchapter. This shows that many diagnoses are associated with markedly higher risk of severe COVID-19. Past hospital diagnoses of infections, pneumonia and acute respiratory diseases were strongly associated with severe COVID-19. Cardiovascular diagnoses associated with COVID-19 were not limited to heart disease but included also stroke and other circulatory disorders that are not designated as risk conditions.

Associations of prescribed drugs with severe disease

As shown in Table 3 and supplementary tables 8 to 10 , the strongest risk factor for severe disease, apart from residence in a care home, is the encashment of at least one prescription in the last year. The univariate rate ratio associated with this variable varies from 9.6 (6.9, 13.3) in those aged under 60 years to 40.3 (25.6, 63.3) in those aged 75 years and over. In a multivariate analysis adjusting for care home residence, any hospital admission and listed conditions, these rate ratios were reduced to 5.0 (3.5, 7.2) and 11.4 (7.1, 18.4) respectively. About one third of controls aged over 75 had not encashed a prescription in the previous year.

To investigate this further, we partitioned the "Any prescription" variable into indicator variables for each chapter of the British National Formulary, in which drugs are grouped by broad indication, and restricted the analysis to those without one of the listed conditions. Table 6 shows these associations. In univariate analyses, prescriptions in almost all BNF chapters were associated with severe disease. In a multivariate analysis of all chapters, most of these associations were weaker. The BNF chapters with the strongest independent associations with severe disease were chapters 1 (gastrointestinal) and 2 (cardiovascular). Other chapters associated with severe disease were chapters 4 (central nervous system), 9 (nutrition and blood) and 14+ (other, mostly dressings and appliances).

Construction of a multivariate risk prediction model

To evaluate the contribution of the listed conditions to risk prediction, and the incremental contribution of other information in hospitalisation and prescription records after assigning these conditions, predictive models were constructed from three sets of variables: a baseline set consisting only of demographic variables, a set that included indicator variables for each listed condition, and an extended set that included demographic, variables, indicator variables for listed conditions and indicator variables for hospital diagnoses in each ICD-10 chapter and prescriptions in each BNF chapter.

For each variable set, a stepwise regression procedure was carried out using alternating forward-backward selection. The variables retained with each variable set are shown in Table 12. Coefficients for specific conditions here should not be interpreted as effect estimates, as global variables for any hospital diagnosis and any listed condition have been included in the model. The predictive performance of the model chosen by stepwise regression was estimated by 4-fold cross-validation. Observed and predicted case status were compared within each stratum over all test folds. Table 7 shows that using the extended set increased the C-statistic from 0.782 to 0.839 and the expected information for discrimination Λ from 0.89 bits to 1.5 bits.

This estimate of 1.5 bits for the information conditional on age and sex obtained from the matched case-control study can be added to the information for discrimination 1.81 bits obtained from the logistic regression on age and sex in the population using age and sex to estimate the total information for discrimination of a risk classifier that would be obtained in the population as 3.31 bits.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Figure 2 shows the distribution of the weight of evidence favouring case over control status from the model based on the extended variable set with a footnote explaining how Λ is derived. This shows, as expected for a multifactorial classifier, that the distributions are approximately Gaussian: there is no clear divide between high-risk and low-risk individuals of the same age and sex. Figure 3 shows the receiver operating characteristic curve with a footnote explaining its derivation from the distribution of the weights of evidence.

Discussion

Sociodemographic factors

This analysis confirms that risk for severe COVID-19 is associated with increasing age, male sex and socioeconomic deprivation. The slope of the relationship of severe disease (on the scale of log odds) to age is less steep than the slope of the relationship of fatal disease to age. Residence in a care home was associated with a 15-fold increased rate of severe COVID-19 in this age matched analysis, reduced to 7-fold by adjustment for listed conditions. This excess risk is likely to reflect both the spread of the epidemic in care homes and residual confounding by frailty.

Although the numbers of cases and controls of non-White ethnicity are small and the assignment of ethnicity is incomplete, the results give some indication of the likely upper bound of the absolute numbers of severe cases in non-White ethnic groups up to now. The only non-White ethnic group with any sizeable numbers is the South Asian category and we found no evidence of any elevation in risk in this group compared to Whites. Reports from England [7] found elevation in risks for some non-White groups. In the OpenSAFELY study risk ratios for fatal COVID-19 of 1.7 in those recorded as Black and and 1.6 in those recorded as Asian, in comparison with those recorded as White, persisted after adjustment for comorbidities and socioeconomic status. In a study of risk factors for hospitalized disease in the UK Biobank cohort, adjustment for health care worker status and other social variables attenuated but did not fully explain the elevated crude risk ratios associated with non-White ethnicity [6,12]. The relative socioeconomic position of ethnic groups in Scotland is different to that in England, so it is plausible that the relation of health status to ethnicity will also differ. For example in the 2011 Scottish Census 1.6% of the population reported South Asian ethnicity. Among the 1.0% who identified as Pakistani or Bangladeshi the proportion living in the most deprived neighbourhoods was not higher than the national average [13]. Future work may allow more complete assignment of ethnicity and disaggregation of broad categories based on continent of origin.

Co-morbidities

We have confirmed that the moderate risk conditions designated by the NHS and other agencies [9] are associated with increased risk of severe COVID-19. However the rate ratios associated with these conditions vary with age - for example the rate ratio associated with diabetes is higher at younger ages. The rate ratios of 4.9 for Type 1 diabetes and 2.6 for Type 2 diabetes are broadly similar to those reported in UK Biobank and in the OpenSAFELY studies. We confirm the higher risk with asthma and chronic lung disease and liver disease reported in these and earlier studies. Of note other heart disease is more strongly associated than ischaemic heart disease. This category includes conditions such as atrial fibrillation, cardiomyopathies and heart failure. Over all age groups, 72% of severe cases had at least one of these listed conditions. Among cases and controls without these conditions, not surprisingly, neoplasms were associated with severe COVID-19; we had omitted it from the pre-specified list as in the current

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

dataset we cannot separately identify those who are currently receiving chemotherapy or radiotherapy for whom shielding is advised. We have not attempted to estimate the risk associated with these conditions for which shielding is recommended, as the observed risk will depend on the adequacy of shielding rather than on the risk to those exposed to the epidemic. In patients without any listed conditions, further systematic evaluation of past hospitalisation history did not reveal a sparse set of underlying conditions; instead many diagnoses were associated with severe COVID-19.

Media reports of apparently healthy young people succumbing to severe COVID-19 have disseminated the message that all are at risk of disease whatever their age or health status. However we found that half of cases under 40 years had at least one of the listed conditions and among those who did not have one of these conditions, the proportions who had at least one prior hospitalisation or dispensed prescription were much higher in cases than in controls. In all age groups, very few of the fatal cases had not had either a hospital admission in the past five years or a dispensed prescription in the past year.

A striking finding of this study was the strong association of severe COVID-19 with having encashed at least one prescription in the past year, only partly explained by higher rates of prescribing among those with listed conditions. Partitioning of this association between BNF chapters, which represent broad indication-based drug classes, showed that the strongest association was with prescription of Chapter 1 drugs, prescribed for gastrointestinal conditions, which are not generally listed as risk factors for severe COVID-19. Also associated were those in the cardiovascular, nervous system and nutritional and blood chapters. Although it is likely that most associations of severe COVID-19 with drug prescribing are attributable to the indications for which these drugs were prescribed, or more diffuse frailty especially in older persons, causal effects of drugs or direct effects of polypharmacy on susceptibility cannot be ruled out. These associations are explored in an accompanying paper.

Relevance to policy

As lockdown restrictions are eased, there is general agreement that vulnerable individuals will require shielding, even if the restart of the epidemic can be slowed or suppressed by mass testing, contact tracing and isolation of those who test positive. The “stratify and shield” policy option [14], in which high-risk individuals comprising up to 15% of the population are shielded for a defined period while the epidemic is allowed to run relatively quickly in low-risk individuals until population-level immunity is attained, depends critically on informative risk discrimination. So too does the similarly named “segment and shield” option [15] which has the opposite objective of keeping transmissions low.

As awareness grows of how risk varies between individuals, individuals will seek information about their own level of risk. A key implication of our results is that risk of severe or fatal disease is multifactorial and that the rate ratio of 5.1 associated with a 20-year increase in age is far stronger than that associated with common diseases such as Type 2 diabetes and asthma that are listed as conditions associated with high risk. A corollary of this is that a crude classification based on assigning all persons with a listed condition to a group for whom shielding is recommended will have poor specificity, as one quarter of those aged 60-74 years in the population have at least one of the listed conditions we examined. It will also exclude many people at high risk because they have multiple risk factors each of small effect. The only way to optimize risk classification so as to ensure equity with respect to risk is to construct a classifier that uses all available information to assign a risk score. Our results show that this is possible in principle, though for this preliminary study we have not used the full repertoire of machine learning methods available for this type of problem. In Scotland it is technically possible to use existing electronic health records to calculate a risk score for every

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

individual in the population, though more work would be required to develop this as a basis for official advice and individual decisions.

Methodological strengths and weaknesses

Most reports of disease associations with COVID-19 have been case series. There have been few reports based on evaluating these associations in the population through cohort or case-control studies. With this matched case control design using incidence density sampling, we have been able to estimate rate ratios conditional on age and sex. An unpublished analysis from England explored the association of similar set of risk conditions with in-hospital COVID-19 deaths, but did not systematically evaluate the rest of the medical record including prescription records. Although we have records of encashment of prescriptions, we do not at present have access to other primary care data, which would contain additional information on morbidity and measurements such as body mass index. A strength of our study however is that hospital discharge diagnoses are coded to ICD-10 by trained coders, in contrast to the coding systems used in primary care databases that do not map to recognized disease classifications. Associations with ethnicity and other sociodemographic factors are not necessarily generalizable from Scotland to other populations.

This case-control study is limited to test-positive cases, excluding deaths with no record of a positive test where COVID-19 was mentioned on the death certificate as an underlying or contributing cause. Up to 13 May 2020 an additional 1200 such deaths had been reported by the National Records of Scotland. Future analyses of this study will include sensitivity analyses of the extent to which the results are changed by including these deaths, but linked data are not yet available to us. Apart from residential care home status, we do not expect most other risk factors to differ markedly between those who died from COVID-19 without being tested and those who died after testing.

Conclusion

This study confirms that risk of severe COVID-19 is associated with sociodemographic factors and with chronic conditions such as diabetes, asthma, circulatory disease and others. However the associations with pre-existing disease are not just with a small set of conditions that contribute to risk, but with many conditions as demonstrated by associations with past medical and prescribing history in relation to multiple physiological systems. As countries attempt to emerge from lockdown whist protecting vulnerable individuals, multivariate classifiers rather than crude rule-based approaches will be needed to define those most at risk of developing severe disease.

Declarations

Information governance

This study was conducted under approvals from the Privacy Advisory Committee ref 44/13 and Public Benefit Privacy Protection amendment 1617-0147. Datasets were de-identified before analysis.

Conflicts of interest

The authors declare no conflicts of interest.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Acknowledgements

We thank all staff in critical care units who submitted data to the SICSAG database, the Scottish Morbidity Record Data Team, the staff of the National Register of Scotland, the Public Health Scotland Terminology Services, the HPS COVID-19 Laboratory & Testing cell and the NHS Scotland Diagnostic Virology Laboratories, and Nicola Rowan (HPS) for coordinating this collaboration.

Public Health Scotland COVID-19 Health Protection Study Group

Alice Whettlock[1], Allan McLeod[1], Andrew Gasiorowski[1], Andrew Merrick[1], Andy McAuley[1], April Went[1], Calum Purdie[1], Colin Fischbacher[1], Colin Ramsay[1], David Bailey[1], David Henderson[1], Diogo Marques[1], Eisin McDonald[1], Genna Drennan[1], Graeme Gowans[1], Graeme Reid[1], Heather Murdoch[1], Jade Carruthers[1], Janet Fleming[1], Jade Carruthers[1], Joseph Jasperse[1], Josie Murray[1], Karen Heatlie[1], Lindsay Mathie[1], Lorraine Donaldson[1], Martin Paton[1], Martin Reid[1], Melissa Llano[1], Michelle Murphy-Hall[1], Paul Smith[1], Ros Hall[1], Ross Cameron[1], Susan Brownlie[1], Adam Gaffney[2], Aynsley Milne[2], Christopher Sullivan[2], Edward McArdle[2], Elaine Glass[2], Johanna Young[2], William Malcolm[2], Jodie McCoubrey [2]

1 Health Protection Scotland (Public Health Scotland), Meridian Court, 5 Cadogan Street, Glasgow G2 6QE.

2 NHS National Services Scotland, Meridian Court, 5 Cadogan Street, Glasgow G2 6QE.

Supplementary material

The R and Rmarkdown scripts used to generate this article, which include the code used to derive variables, will be made available with this manuscript.

References

1. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, et al. Features of 16,749 hospitalised UK patients with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol. medRxiv. 2020; 2020.04.23.20076042. doi:10.1101/2020.04.23.20076042

2. McGoogan C, Steafel E. Why are young, healthy people dying of coronavirus? The symptoms to look out for. The Telegraph. https://www.telegraph.co.uk/health-fitness/body/why-young-healthy-people-dying-coronavirus/; 2020.

3. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. Lancet. 2020;395: 1054–1062. doi:10.1016/S0140-6736(20)30566-3

4. Grasselli G, Zangrillo A, Zanella A, Antonelli M, Cabrini L, Castelli A, et al. Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. JAMA. 2020;323: 1574–1581. doi:10.1001/jama.2020.5394

5. Guan W-j, Ni Z-y, Hu Y, Liang W-h, Ou C-q, He J-x, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. New England Journal of Medicine. 2020. doi:10.1056/NEJMoa2002032

6. Niedzwiedz CL, O'Donnell CA, Jani BD, Demou E, Ho FK, Celis-Morales C, et al. Ethnic and socioeconomic differences in SARS-CoV2 infection in the UK Biobank cohort study. medRxiv. 2020; 2020.04.22.20075663. doi:10.1101/2020.04.22.20075663

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

7. "The OpenSAFELY Collaborative", Williamson E, Walker AJ, Bhaskaran KJ, Bacon S, Bates C, et al. OpenSAFELY: Factors associated with COVID-19-related hospital death in the linked electronic health records of 17 million adult NHS patients. medRxiv. 2020; 2020.05.06.20092999. doi:10.1101/2020.05.06.20092999

8. Lakha F, Gorman DR, Mateos P. Name analysis to classify populations by ethnicity in public health: Validation of Onomap in Scotland. Public Health. 2011;125: 688–696. doi:10.1016/j.puhe.2011.05.003

9. NHS. Who's at higher risk from coronavirus - Coronavirus (COVID-19). nhs.uk. https://www.nhs.uk/conditions/coronavirus-covid-19/people-at-higher-risk-from-coronavirus/whos-at-higher-risk-from-coronavirus/;

10. Breslow NE, Day NE, Halvorsen KT, Prentice RL, Sabai C. Estimation of multiple relative risk functions in matched case-control studies. Am J Epidemiol. 1978;108: 299–307. doi:10.1093/oxfordjournals.aje.a112623

11. McKeigue P. Quantifying performance of a diagnostic test as the expected information for discrimination: Relation to the C-statistic. Stat Methods Med Res. 2019;28: 1841–1851. doi:10.1177/0962280218776989

12. Ho FK, Celis-Morales CA, Gray SR, Katikireddi SV, Niedzwiedz CL, Hastie C, et al. Modifiable and non-modifiable risk factors for COVID-19: Results from UK Biobank. medRxiv. 2020; 2020.04.28.20083295. doi:10.1101/2020.04.28.20083295

13. Walsh D, Buchanan D, Douglas A, Erdman J, Fischbacher C, McCartney G, et al. Increasingly Diverse: The Changing Ethnic Profiles of Scotland and Glasgow and the Implications for Population Health. Appl Spatial Analysis. 2019;12: 983–1009. doi:10.1007/s12061-018-9281-7

14. McKeigue PM, Colhoun HM. Evaluation of "stratify and shield" as a policy option for ending the COVID-19 lockdown in the UK. medRxiv. 2020; 2020.04.25.20079913. doi:10.1101/2020.04.25.20079913

15. Bunnik BAD van, Morgan ALK, Bessell P, Calder-Gerver G, Zhang F, Haynes S, et al. Segmentation and shielding of the most vulnerable members of the population as elements of an exit strategy from COVID-19 lockdown. medRxiv. 2020; 2020.05.04.20090597. doi:10.1101/2020.05.04.20090597

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Figures

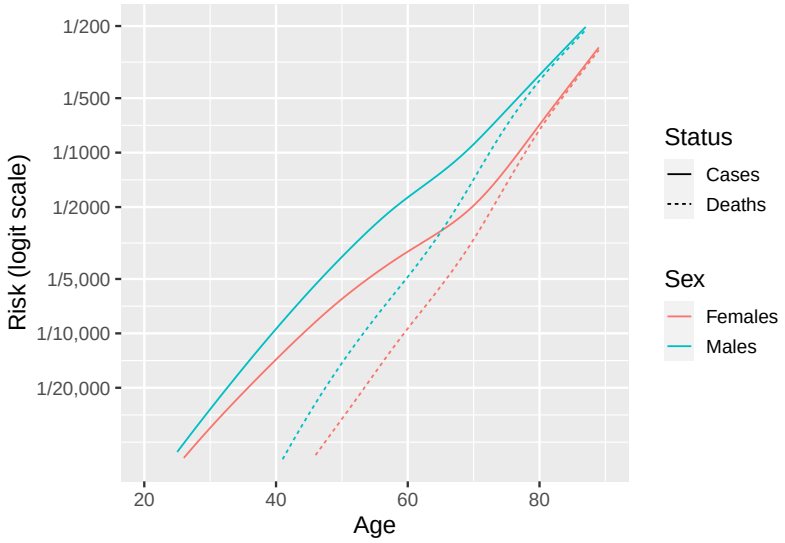

Fig 1. Incidence of severe and fatal COVID-19 in Scotland by age and sex: generalized additive models fitted to severe and fatal cases for males and females separately

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

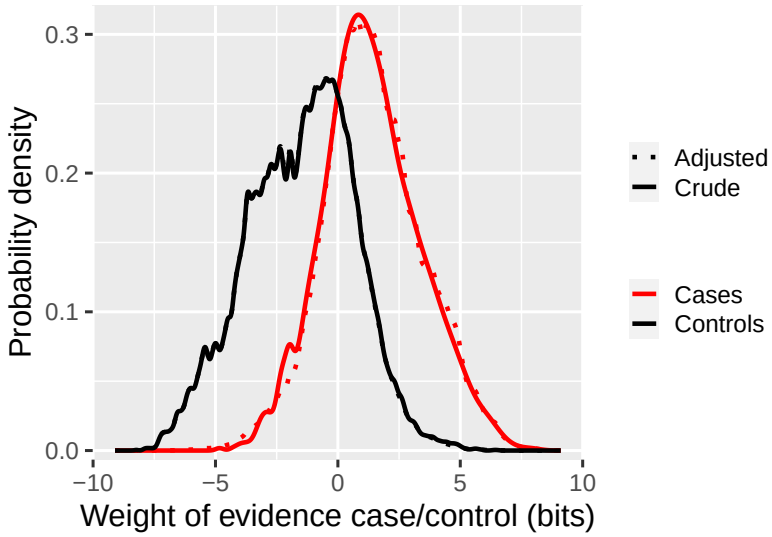

Fig 2. Cross-validation of model chosen by stepwise regression using extended variable set: class-conditional distributions of weight of evidence

Footnote for Figure 2

For each individual, the risk prediction model outputs the posterior probability of being a case, which can also be expressed as the posterior odds. Dividing the posterior odds by the prior odds gives the likelihood ratio favouring case over non-case status for an individual. The weight of evidence W is the logarithm of this ratio. The distributions of W in cases and controls in the test data are plotted in Figure 2. For a classifier, the further apart these curves are, the better the predictive performance. The expected information for discrimination Λ is the average of the mean of the distribution of W in cases and minus 1 times the mean of the distribution of W in controls. The distributions have been adjusted by taking a weighted average to make them mathematically consistent [11].

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

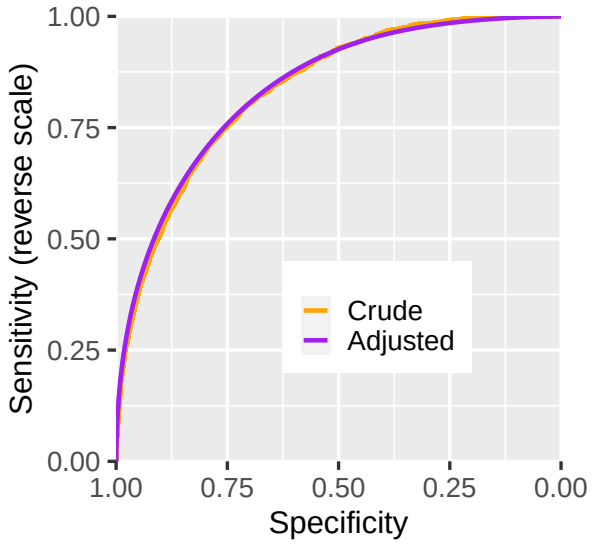

Fig 3. Cross-validation of model chosen by stepwise regression using extended variable set: receiver operating characteristic curve

Footnote for Figure 3

The crude receiver operator characteristic (ROC) curve is computed by calculating at each value of the risk score the sensitivity and specificity of a classifier that uses this value as the threshold for classifying cases and non-cases. The C-statistic is the area under this curve, computed as the probability of correctly classifying a case/noncase pair using the score, evaluated over all possible such pairs in the dataset.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Tables

Table 1. Univariate associations of severe disease with demographic factors

	Controls (22089)	Cases (2222)	Rate ratio (95% CI)	p-value
Ethnicity based on name classification				
White	21081 (96%)	2169 (98%)		
South Asian	614 (3%)	35 (2%)	0.53 (0.37, 0.76)	4×10^{-4}
Other	303 (1%)	10 (0%)	0.31 (0.17, 0.59)	3×10^{-4}
SIMD quintile				
1 - most deprived	5323 (25%)	610 (28%)		
2	4513 (21%)	534 (24%)	1.00 (0.88, 1.14)	1
3	3858 (18%)	392 (18%)	0.83 (0.71, 0.96)	0.01
4	3784 (18%)	376 (17%)	0.79 (0.68, 0.92)	0.002
5 - least deprived	4049 (19%)	305 (14%)	0.56 (0.48, 0.67)	1×10^{-11}
Care home	801 (4%)	516 (23%)	14.9 (12.7, 17.5)	7×10^{-240}
Ethnicity based on Scottish Morbidity Record				
White	5903 (99%)	1172 (99%)		
South Asian	42 (1%)	9 (1%)	0.81 (0.31, 2.10)	0.7
Black	15 (0%)	2 (0%)	0.34 (0.04, 2.84)	0.3
Other	27 (0%)	5 (0%)	0.88 (0.27, 2.84)	0.8

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Table 2. Frequencies of risk factors in cases and controls, by age group

	0-39 years		40-59 years		60-74 years		75+ years	
	Controls (600)	Cases (60)	Controls (3519)	Cases (352)	Controls (5848)	Cases (585)	Controls (12122)	Cases (1225)
Care home	0 (0%)	1 (2%)	4 (0%)	9 (3%)	38 (1%)	50 (9%)	759 (6%)	456 (37%)
Any prescription	274 (46%)	50 (83%)	1793 (51%)	318 (90%)	3906 (67%)	556 (95%)	8170 (67%)	1204 (98%)
Any admission	133 (22%)	30 (50%)	869 (25%)	209 (59%)	2207 (38%)	442 (76%)	5872 (48%)	1034 (84%)
Any listed condition	53 (9%)	30 (50%)	616 (18%)	194 (55%)	1860 (32%)	399 (68%)	5172 (43%)	976 (80%)
No listed condition, other diagnosis	102 (17%)	9 (15%)	520 (15%)	62 (18%)	909 (16%)	95 (16%)	1641 (14%)	144 (12%)
Diagnosis or prescription	305 (51%)	50 (83%)	1899 (54%)	333 (95%)	4026 (69%)	569 (97%)	8291 (68%)	1218 (99%)
Type 1 diabetes	4 (1%)	3 (5%)	19 (1%)	15 (4%)	31 (1%)	6 (1%)	27 (0%)	10 (1%)
Type 2 diabetes	3 (0%)	2 (3%)	151 (4%)	63 (18%)	648 (11%)	148 (25%)	1515 (12%)	276 (23%)
Other/unknown type	1 (0%)	4 (7%)	18 (1%)	9 (3%)	36 (1%)	5 (1%)	62 (1%)	20 (2%)
Ischaemic heart disease	1 (0%)	1 (2%)	87 (2%)	28 (8%)	463 (8%)	113 (19%)	1651 (14%)	311 (25%)
Other heart disease	1 (0%)	7 (12%)	88 (3%)	60 (17%)	555 (9%)	171 (29%)	2565 (21%)	579 (47%)
Asthma or chronic airway disease	40 (7%)	23 (38%)	344 (10%)	91 (26%)	793 (14%)	209 (36%)	1915 (16%)	407 (33%)
Chronic kidney disease or transplant recipient	0 (0%)	0 (0%)	4 (0%)	17 (5%)	11 (0%)	18 (3%)	42 (0%)	30 (2%)
Neurological (except epilepsy) or dementia	3 (0%)	7 (12%)	39 (1%)	31 (9%)	144 (2%)	80 (14%)	963 (8%)	361 (29%)
Liver disease	1 (0%)	0 (0%)	13 (0%)	9 (3%)	19 (0%)	9 (2%)	22 (0%)	7 (1%)
Immune deficiency or suppression	4 (1%)	2 (3%)	14 (0%)	11 (3%)	25 (0%)	10 (2%)	35 (0%)	9 (1%)

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 3. Associations of severe disease with listed conditions over all age groups

			Univariate		Multivariate	
	Controls (22089)	Cases (2222)	Rate ratio (95% CI)	p-value	Rate ratio (95% CI)	p-value
Care home	801 (4%)	516 (23%)	14.9 (12.7, 17.5)	7×10^{-240}	7.4 (6.2, 8.9)	1×10^{-105}
Any prescription	14143 (64%)	2128 (96%)	16.6 (13.3, 20.6)	3×10^{-141}	6.3 (5.0, 8.0)	1×10^{-54}
Any admission	9081 (41%)	1715 (77%)	5.6 (5.0, 6.2)	9×10^{-213}	1.92 (1.68, 2.18)	4×10^{-23}
Type 1 diabetes	81 (0%)	34 (2%)	4.88 (3.26, 7.31)	1×10^{-14}	2.20 (1.39, 3.50)	8×10^{-4}
Type 2 diabetes	2317 (10%)	489 (22%)	2.58 (2.30, 2.88)	1×10^{-60}	1.57 (1.38, 1.78)	2×10^{-12}
Other/unknown type	117 (1%)	38 (2%)	3.89 (2.69, 5.63)	6×10^{-13}	2.01 (1.33, 3.03)	0.001
Ischaemic heart disease	2202 (10%)	453 (20%)	2.40 (2.14, 2.70)	1×10^{-49}	1.08 (0.94, 1.23)	0.3
Other heart disease	3209 (15%)	817 (37%)	3.90 (3.52, 4.32)	3×10^{-151}	1.42 (1.25, 1.60)	4×10^{-8}
Asthma or chronic airway disease	3092 (14%)	730 (33%)	3.10 (2.81, 3.42)	2×10^{-112}	1.59 (1.43, 1.78)	1×10^{-16}
Chronic kidney disease or transplant recipient	57 (0%)	65 (3%)	12.1 (8.4, 17.4)	1×10^{-40}	4.98 (3.30, 7.50)	2×10^{-14}
Neurological (except epilepsy) or dementia	1149 (5%)	479 (22%)	5.5 (4.8, 6.2)	3×10^{-158}	1.84 (1.59, 2.13)	9×10^{-16}
Liver disease	55 (0%)	25 (1%)	4.70 (2.90, 7.62)	3×10^{-10}	1.65 (0.96, 2.85)	0.07
Immune deficiency or suppression	78 (0%)	32 (1%)	4.11 (2.72, 6.21)	2×10^{-11}	1.17 (0.71, 1.93)	0.5

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 4. Proportions of fatal cases and matched controls without and with a dispensed prescription or hospital diagnosis, by age group

	Controls	Fatal cases
Age <60		
No scrip or diagnosis	1941 (44%)	3 (3%)
Scrip or diagnosis	2483 (56%)	104 (97%)
Age 60-74		
No scrip or diagnosis	1830 (30%)	8 (2%)
Scrip or diagnosis	4256 (70%)	339 (98%)
Age 75+		
No scrip or diagnosis	3832 (31%)	6 (1%)
Scrip or diagnosis	8366 (69%)	1143 (99%)

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 5. Associations of severe disease with hospital diagnoses in last 5 years, in those without any listed condition

			Univariate		Multivariate	
	Controls (14388)	Cases (623)	Rate ratio (95% CI)	p-value	Rate ratio (95% CI)	p-value
Ch.1 infectious	245 (2%)	48 (8%)	5.0 (3.4, 7.4)	6×10^{-16}	1.54 (0.96, 2.48)	0.07
Ch.2 Neoplasms	543 (4%)	67 (11%)	3.56 (2.61, 4.86)	1×10^{-15}	2.25 (1.55, 3.27)	2×10^{-5}
Ch.3 blood	165 (1%)	23 (4%)	4.89 (2.79, 8.58)	3×10^{-8}	1.42 (0.74, 2.73)	0.3
Ch.4 Endocrine	155 (1%)	27 (4%)	4.38 (2.64, 7.26)	1×10^{-8}	1.36 (0.75, 2.46)	0.3
Ch.5 Mental	215 (1%)	54 (9%)	5.6 (3.8, 8.3)	2×10^{-18}	1.72 (1.06, 2.79)	0.03
Ch.6 nervous	130 (1%)	16 (3%)	3.61 (1.89, 6.87)	1×10^{-4}	2.38 (1.16, 4.88)	0.02
Ch.7 eye	714 (5%)	55 (9%)	1.99 (1.43, 2.77)	5×10^{-5}	1.65 (1.14, 2.38)	0.008
Ch.8 ear	50 (0%)	4 (1%)	2.46 (0.79, 7.68)	0.1	1.62 (0.47, 5.59)	0.4
Ch.9 circulatory	226 (2%)	34 (5%)	3.51 (2.27, 5.45)	2×10^{-8}	1.14 (0.67, 1.94)	0.6
Ch.10 respiratory	164 (1%)	41 (7%)	5.8 (3.8, 8.9)	9×10^{-16}	2.04 (1.21, 3.44)	0.008
Ch.11 digestive	981 (7%)	105 (17%)	2.83 (2.20, 3.64)	6×10^{-16}	1.47 (1.09, 1.98)	0.01
Ch.12 skin	170 (1%)	18 (3%)	2.10 (1.23, 3.57)	0.006	0.69 (0.36, 1.32)	0.3
Ch.13 musculoskeletal	601 (4%)	77 (12%)	3.34 (2.49, 4.48)	9×10^{-16}	1.65 (1.17, 2.34)	0.004
Ch.14 genitourinary	550 (4%)	81 (13%)	3.87 (2.88, 5.20)	2×10^{-19}	1.47 (1.01, 2.14)	0.04
Ch.15 Pregnancy	22 (0%)	0 (0%)	0.00 (0.00, Inf)	1	0.00 (0.00, Inf)	1
Ch.17 Congenital	21 (0%)	4 (1%)	6.1 (1.6, 22.9)	0.007	2.39 (0.56, 10.28)	0.2
Ch.18 Symptoms	853 (6%)	123 (20%)	3.98 (3.12, 5.08)	2×10^{-28}	1.56 (1.13, 2.15)	0.007
Ch.19 Injury	434 (3%)	81 (13%)	4.36 (3.22, 5.91)	2×10^{-21}	1.18 (0.46, 3.03)	0.7
Ch.20 External	494 (3%)	93 (15%)	4.71 (3.53, 6.29)	5×10^{-26}	1.79 (0.72, 4.42)	0.2
Ch.21 Health factors	872 (6%)	99 (16%)	3.03 (2.34, 3.92)	2×10^{-17}	0.94 (0.67, 1.32)	0.7

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Table 6. Associations of severe disease with prescribed drugs in those without any listed condition

			Univariate		Multivariate	
	Controls (14388)	Cases (623)	Rate ratio (95% CI)	p-value	Rate ratio (95% CI)	p-value
BNF 1 Gastro	2952 (21%)	330 (53%)	5.1 (4.2, 6.1)	2×10^{-66}	2.23 (1.78, 2.80)	5×10^{-12}
BNF 2 Cardiovascular	3861 (27%)	334 (54%)	3.92 (3.24, 4.74)	4×10^{-45}	2.07 (1.66, 2.58)	8×10^{-11}
BNF 3 Respiratory	556 (4%)	68 (11%)	3.15 (2.33, 4.26)	1×10^{-13}	1.34 (0.94, 1.91)	0.1
BNF 4 Nervous	3234 (22%)	342 (55%)	4.58 (3.81, 5.50)	1×10^{-58}	1.82 (1.44, 2.30)	5×10^{-7}
BNF 5 Infections	1666 (12%)	182 (29%)	3.15 (2.58, 3.86)	6×10^{-29}	1.32 (1.04, 1.68)	0.02
BNF 6 Endocrine	1289 (9%)	122 (20%)	2.80 (2.20, 3.55)	4×10^{-17}	1.19 (0.89, 1.59)	0.2
BNF 7 Obstetrics	1056 (7%)	79 (13%)	1.90 (1.45, 2.50)	4×10^{-6}	0.73 (0.53, 1.00)	0.05
BNF 8 Malignant	153 (1%)	16 (3%)	2.17 (1.22, 3.88)	0.009	0.85 (0.44, 1.66)	0.6
BNF 9 Nutrition	1327 (9%)	168 (27%)	4.37 (3.48, 5.49)	4×10^{-37}	1.71 (1.31, 2.23)	9×10^{-5}
BNF 10 Musculoskeletal	1731 (12%)	176 (28%)	2.89 (2.36, 3.54)	8×10^{-25}	1.08 (0.85, 1.37)	0.5
BNF 11 Eye	906 (6%)	60 (10%)	1.57 (1.16, 2.13)	0.004	0.72 (0.51, 1.02)	0.06
BNF 12 Ear	818 (6%)	59 (9%)	1.71 (1.27, 2.31)	5×10^{-4}	0.65 (0.46, 0.92)	0.01
BNF 13 Skin	1643 (11%)	167 (27%)	3.17 (2.57, 3.91)	3×10^{-27}	1.32 (1.02, 1.71)	0.03
BNF 14 Other	1687 (12%)	202 (32%)	4.13 (3.36, 5.08)	4×10^{-41}	1.91 (1.49, 2.46)	4×10^{-7}

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 7. Prediction of severe COVID-19: cross-validation of models chosen by stepwise regression

	Cases / controls	Crude C-statistic	Adjusted C-statistic	Crude Λ (bits)	Adjusted Λ (bits)	Test log-likelihood (nats)
Demographic only	2109 / 20417	0.697	0.696	0.52	0.46	0.0
Demographic + listed conditions	2109 / 20417	0.793	0.782	0.96	0.89	482.2
Extended variable set	2109 / 20417	0.836	0.839	1.44	1.50	912.2

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Supplementary tables

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Table 8. Associations of severe disease with listed conditions in those aged less than 60

	Controls (4119)	Cases (412)	Univariate		Multivariate	
			Rate ratio (95% CI)	p-value	Rate ratio (95% CI)	p-value
Care home	4 (0%)	10 (2%)	25.0 (7.8, 79.7)	5×10^{-8}	6.6 (1.7, 25.1)	0.006
Any prescription	2067 (50%)	368 (89%)	9.6 (6.9, 13.3)	6×10^{-41}	5.0 (3.5, 7.2)	1×10^{-18}
Any admission	1002 (24%)	239 (58%)	4.54 (3.66, 5.63)	2×10^{-43}	1.62 (1.25, 2.10)	3×10^{-4}
Type 1 diabetes	23 (1%)	18 (4%)	9.8 (5.2, 18.4)	1×10^{-12}	4.29 (2.04, 8.99)	1×10^{-4}
Type 2 diabetes	154 (4%)	65 (16%)	5.4 (3.9, 7.5)	6×10^{-25}	2.84 (1.97, 4.09)	2×10^{-8}
Other/unknown type	19 (0%)	13 (3%)	7.4 (3.6, 15.0)	3×10^{-8}	4.28 (1.96, 9.37)	3×10^{-4}
Ischaemic heart disease	88 (2%)	29 (7%)	3.66 (2.34, 5.72)	1×10^{-8}	0.76 (0.43, 1.33)	0.3
Other heart disease	89 (2%)	67 (16%)	9.2 (6.5, 13.1)	1×10^{-35}	2.81 (1.81, 4.36)	4×10^{-6}
Asthma or chronic airway disease	384 (9%)	114 (28%)	3.84 (3.00, 4.90)	8×10^{-27}	1.81 (1.37, 2.40)	4×10^{-5}
Chronic kidney disease or transplant recipient	4 (0%)	17 (4%)	42.5 (14.3, 126.3)	2×10^{-11}	22.3 (5.1, 98.5)	4×10^{-5}
Neurological (except epilepsy) or dementia	42 (1%)	38 (9%)	10.1 (6.4, 16.0)	1×10^{-22}	3.67 (2.09, 6.44)	6×10^{-6}
Liver disease	14 (0%)	9 (2%)	7.3 (3.0, 17.7)	1×10^{-5}	1.61 (0.54, 4.75)	0.4
Immune deficiency or suppression	18 (0%)	13 (3%)	7.2 (3.5, 14.7)	6×10^{-8}	0.35 (0.10, 1.27)	0.1

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 9. Associations of severe disease with listed conditions in those aged 60-74 years

			Univariate		Multivariate	
	Controls (5848)	Cases (585)	Rate ratio (95% CI)	p-value	Rate ratio (95% CI)	p-value
Care home	38 (1%)	50 (9%)	19.0 (11.6, 31.3)	5×10^{-31}	9.2 (5.3, 16.0)	2×10^{-15}
Any prescription	3906 (67%)	556 (95%)	10.8 (7.3, 16.0)	3×10^{-33}	4.59 (3.02, 6.98)	9×10^{-13}
Any admission	2207 (38%)	442 (76%)	5.3 (4.4, 6.5)	1×10^{-60}	2.22 (1.76, 2.81)	2×10^{-11}
Type 1 diabetes	31 (1%)	6 (1%)	2.38 (0.99, 5.76)	0.05	0.80 (0.27, 2.33)	0.7
Type 2 diabetes	648 (11%)	148 (25%)	2.77 (2.26, 3.41)	3×10^{-22}	1.67 (1.33, 2.09)	9×10^{-6}
Other/unknown type	36 (1%)	5 (1%)	1.73 (0.67, 4.43)	0.3	0.63 (0.22, 1.85)	0.4
Ischaemic heart disease	463 (8%)	113 (19%)	2.87 (2.28, 3.62)	4×10^{-19}	1.20 (0.92, 1.57)	0.2
Other heart disease	555 (9%)	171 (29%)	4.15 (3.38, 5.09)	4×10^{-42}	1.44 (1.13, 1.85)	0.004
Asthma or chronic airway disease	793 (14%)	209 (36%)	3.68 (3.05, 4.45)	2×10^{-41}	1.84 (1.49, 2.26)	1×10^{-8}
Chronic kidney disease or transplant recipient	11 (0%)	18 (3%)	17.6 (8.1, 38.1)	4×10^{-13}	8.2 (3.5, 19.2)	2×10^{-6}
Neurological (except epilepsy) or dementia	144 (2%)	80 (14%)	6.4 (4.8, 8.6)	4×10^{-35}	2.54 (1.80, 3.58)	1×10^{-7}
Liver disease	19 (0%)	9 (2%)	4.84 (2.17, 10.80)	1×10^{-4}	1.73 (0.67, 4.44)	0.3
Immune deficiency or suppression	25 (0%)	10 (2%)	4.00 (1.92, 8.33)	2×10^{-4}	0.86 (0.34, 2.20)	0.8

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Table 10. Associations of severe disease with listed conditions in those aged 75 years and over

			Univariate		Multivariate	
	Controls (12122)	Cases (1225)	Rate ratio (95% CI)	p-value	Rate ratio (95% CI)	p-value
Care home	759 (6%)	456 (37%)	14.3 (12.0, 17.0)	2×10^{-204}	7.3 (6.0, 8.9)	1×10^{-90}
Any prescription	8170 (67%)	1204 (98%)	40.3 (25.6, 63.3)	8×10^{-58}	11.4 (7.1, 18.4)	1×10^{-23}
Any admission	5872 (48%)	1034 (84%)	6.4 (5.5, 7.6)	9×10^{-109}	1.72 (1.41, 2.10)	7×10^{-8}
Type 1 diabetes	27 (0%)	10 (1%)	4.13 (2.00, 8.55)	1×10^{-4}	2.20 (0.96, 5.05)	0.06
Type 2 diabetes	1515 (12%)	276 (23%)	2.14 (1.85, 2.48)	8×10^{-24}	1.33 (1.13, 1.57)	6×10^{-4}
Other/unknown type	62 (1%)	20 (2%)	3.82 (2.29, 6.37)	3×10^{-7}	2.23 (1.25, 3.98)	0.007
Ischaemic heart disease	1651 (14%)	311 (25%)	2.18 (1.89, 2.50)	1×10^{-27}	1.07 (0.91, 1.26)	0.4
Other heart disease	2565 (21%)	579 (47%)	3.45 (3.05, 3.90)	2×10^{-86}	1.32 (1.13, 1.54)	4×10^{-4}
Asthma or chronic airway disease	1915 (16%)	407 (33%)	2.71 (2.38, 3.08)	1×10^{-50}	1.43 (1.24, 1.66)	2×10^{-6}
Chronic kidney disease or transplant recipient	42 (0%)	30 (2%)	7.5 (4.6, 12.1)	2×10^{-16}	3.47 (2.05, 5.90)	4×10^{-6}
Neurological (except epilepsy) or dementia	963 (8%)	361 (29%)	4.97 (4.31, 5.74)	7×10^{-107}	1.60 (1.34, 1.90)	1×10^{-7}
Liver disease	22 (0%)	7 (1%)	3.19 (1.35, 7.54)	0.008	1.37 (0.52, 3.57)	0.5
Immune deficiency or suppression	35 (0%)	9 (1%)	2.57 (1.23, 5.36)	0.01	1.66 (0.73, 3.76)	0.2

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Table 11. Univariate associations of severe disease with hospital diagnoses by ICD subchapters in those without any listed conditions: rows retained are those with p < 0.001 and at least 50 cases and controls

	Controls (14388)	Cases (623)	Rate ratio (95% CI)	p-value
Intestinal Infectious Diseases	94 (1%)	17 (3%)	4.19 (2.20, 7.96)	1×10^{-5}
Bacterial And Viral Infectious Agents	77 (1%)	13 (2%)	3.74 (1.86, 7.52)	2×10^{-4}
Melanoma And Other Malignant Neoplasms Of Skin	117 (1%)	12 (2%)	3.11 (1.52, 6.36)	0.002
Malignant Neoplasms Of Breast	58 (0%)	1 (0%)	0.39 (0.05, 2.96)	0.4
Malignant Neoplasms Of Male Genital Organs	58 (0%)	5 (1%)	2.08 (0.75, 5.75)	0.2
Malignant Neuroendocrine Tumors	60 (0%)	11 (2%)	5.1 (2.2, 11.4)	1×10^{-4}
Benign Neoplasms, Except Benign Neuroendocrine Tumors	182 (1%)	23 (4%)	3.33 (2.00, 5.52)	3×10^{-6}
Nutritional Anemias	79 (1%)	5 (1%)	2.03 (0.73, 5.64)	0.2
Aplastic And Other Anemias And Other Bone Marrow Failure Syndromes	73 (1%)	11 (2%)	8.2 (3.3, 20.6)	7×10^{-6}
Metabolic Disorders	101 (1%)	20 (3%)	4.50 (2.49, 8.14)	7×10^{-7}
Mental Disorders Due To Known Physiological Conditions	74 (1%)	33 (5%)	11.1 (6.0, 20.9)	5×10^{-14}
Mental And Behavioral Disorders Due To Psychoactive Substance Use	109 (1%)	13 (2%)	2.32 (1.22, 4.41)	0.01
Episodic And Paroxysmal Disorders	55 (0%)	7 (1%)	3.45 (1.25, 9.51)	0.02
Nerve, Nerve Root And Plexus Disorders	75 (1%)	9 (1%)	3.63 (1.59, 8.27)	0.002
Disorders Of Eyelid, Lacrimal System And Orbit	46 (0%)	6 (1%)	2.43 (0.91, 6.48)	0.07
Disorders Of Lens	616 (4%)	47 (8%)	1.93 (1.35, 2.75)	3×10^{-4}
Disorders Of Choroid And Retina	72 (1%)	4 (1%)	1.73 (0.57, 5.26)	0.3
Glaucoma	65 (0%)	4 (1%)	1.86 (0.61, 5.71)	0.3
Cerebrovascular Diseases	78 (1%)	13 (2%)	4.07 (1.97, 8.38)	1×10^{-4}
Diseases Of Veins, Lymphatic Vessels And Lymph Nodes, Not Elsewhere Classified	76 (1%)	9 (1%)	2.15 (0.99, 4.67)	0.05
Other And Unspecified Disorders Of The Circulatory System	49 (0%)	11 (2%)	4.78 (2.15, 10.64)	1×10^{-4}
Influenza And Pneumonia	62 (0%)	13 (2%)	5.5 (2.6, 11.8)	9×10^{-6}
Other Acute Lower Respiratory Infections	65 (0%)	23 (4%)	9.7 (5.0, 18.8)	1×10^{-11}
Diseases Of Esophagus, Stomach And Duodenum	313 (2%)	30 (5%)	2.15 (1.39, 3.32)	5×10^{-4}
Hernia	247 (2%)	22 (4%)	2.34 (1.42, 3.87)	9×10^{-4}

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Noninfective Enteritis And Colitis	58 (0%)	9 (1%)	3.94 (1.68, 9.27)	0.002
Other Diseases Of Intestines	468 (3%)	54 (9%)	3.21 (2.27, 4.54)	4×10^{-11}
Disorders Of Gallbladder, Biliary Tract And Pancreas	110 (1%)	13 (2%)	2.12 (1.09, 4.13)	0.03
Other Diseases Of The Digestive System	83 (1%)	12 (2%)	3.48 (1.72, 7.04)	5×10^{-4}
Infections Of The Skin And Subcutaneous Tissue	64 (0%)	8 (1%)	2.31 (1.05, 5.10)	0.04
Other Disorders Of The Skin And Subcutaneous Tissue	46 (0%)	10 (2%)	4.82 (2.12, 10.94)	2×10^{-4}
Inflammatory Polyarthropathies	47 (0%)	11 (2%)	5.6 (2.5, 12.7)	3×10^{-5}
Osteoarthritis	263 (2%)	29 (5%)	3.02 (1.90, 4.80)	3×10^{-6}
Other Joint Disorders	97 (1%)	10 (2%)	2.40 (1.14, 5.04)	0.02
Other Dorsopathies	85 (1%)	7 (1%)	2.05 (0.85, 4.94)	0.1
Other Soft Tissue Disorders	115 (1%)	16 (3%)	3.09 (1.68, 5.66)	3×10^{-4}
Disorders Of Bone Density And Structure	55 (0%)	10 (2%)	6.0 (2.2, 16.4)	6×10^{-4}
Acute Kidney Failure And Chronic Kidney Disease	117 (1%)	29 (5%)	7.2 (4.0, 12.8)	2×10^{-11}
Urolithiasis	44 (0%)	6 (1%)	2.69 (1.03, 7.05)	0.04
Other Diseases Of The Urinary System	241 (2%)	42 (7%)	4.47 (2.96, 6.76)	1×10^{-12}
Diseases Of Male Genital Organs	159 (1%)	14 (2%)	2.17 (1.17, 4.02)	0.01
Noninflammatory Disorders Of Female Genital Tract	85 (1%)	7 (1%)	1.75 (0.75, 4.09)	0.2
Symptoms And Signs Involving The Circulatory And Respiratory Systems	173 (1%)	15 (2%)	2.01 (1.10, 3.69)	0.02
Symptoms And Signs Involving The Digestive System And Abdomen	277 (2%)	33 (5%)	2.46 (1.62, 3.74)	2×10^{-5}
Symptoms And Signs Involving The Nervous And Musculoskeletal Systems	120 (1%)	38 (6%)	11.0 (6.2, 19.3)	1×10^{-16}
Symptoms And Signs Involving The Genitourinary System	182 (1%)	25 (4%)	3.22 (1.94, 5.36)	7×10^{-6}
Symptoms And Signs Involving Cognition, Perception, Emotional State And Behavior	81 (1%)	21 (3%)	4.99 (2.75, 9.07)	1×10^{-7}
General Symptoms And Signs	212 (1%)	37 (6%)	4.00 (2.60, 6.14)	3×10^{-10}
Abnormal Findings On Diagnostic Imaging And In Function Studies, Without Diagnosis	64 (0%)	16 (3%)	6.6 (3.2, 13.5)	2×10^{-7}
Injuries To The Head	99 (1%)	19 (3%)	3.79 (2.09, 6.89)	1×10^{-5}
Injuries To The Wrist, Hand And Fingers	52 (0%)	5 (1%)	2.04 (0.74, 5.60)	0.2
Injuries To The Hip And Thigh	67 (0%)	22 (4%)	8.3 (4.3, 16.1)	5×10^{-10}
Injuries To The Knee And Lower Leg	44 (0%)	9 (1%)	3.85 (1.65, 9.01)	0.002

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Complications Of Surgical And Medical Care, Not Elsewhere Classified	94 (1%)	14 (2%)	3.69 (1.86, 7.33)	2×10^{-4}
Slipping, Tripping, Stumbling And Falls	243 (2%)	52 (8%)	5.1 (3.5, 7.6)	1×10^{-16}
Misadventures To Patients During Surgical And Medical Care	42 (0%)	13 (2%)	10.7 (4.2, 27.1)	6×10^{-7}
Surgical And Other Medical Procedures As The Cause Of Abnormal Reaction Of The Patient, Or Of Later Complication, Without Mention Of Misadventure At The Time Of The Procedure	115 (1%)	16 (3%)	3.79 (2.00, 7.21)	5×10^{-5}

medRxiv preprint doi: https://doi.org/10.1101/2020.05.28.20115394; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 12. Stepwise regression: variables retained in model for severe disease

	log rate ratio	p-value
Care home resident	1.89	2×10^{-89}
SIMD - quintile 1 as reference		
SIMD.quintile 2	0.00	1
SIMD.quintile 3	-0.08	0.4
SIMD.quintile 4	-0.21	0.02
SIMD.quintile 5 - least deprived	-0.33	0.001
Diabetes - non-diabetic as reference		
Other diabetes type	0.52	0.1
Type 1	0.45	0.07
Type 2	0.21	0.004
Other heart disease	0.25	1×10^{-4}
Asthma or chronic airway disease	0.27	6×10^{-6}
Chronic kidney disease or transplant recipient	0.57	1×10^{-8}
Neurological (except epilepsy) or dementia	0.42	1×10^{-8}
Liver disease	0.52	0.07
Any admission	0.52	2×10^{-13}
Any prescription	1.22	4×10^{-19}
BNF 1 Gastro	0.25	7×10^{-5}
BNF 2 Cardiovascular	0.11	0.1
BNF 4 Nervous	0.28	3×10^{-5}
BNF 5 Infections	0.16	0.005
BNF 6 Endocrine	0.29	3×10^{-6}
BNF 7 Obstetrics	-0.17	0.01
BNF 9 Nutrition	0.39	4×10^{-11}
BNF 14 Other	0.25	2×10^{-5}