A nationwide study of 331 rare diseases among 58 million individuals: prevalence, demographics, and COVID-19 outcomes

Johan H Thygesen1, Huayu Zhang2, Hanane Issa1, Jinge Wu1, Tuankasfee Hama1, Ana Caterina Phiho Gomes1, Tudor Groza3, Sara Khalid4, Tom Lumbers1, Mevhibe Hocaoglu5, Kamlesh Khunti6, Rouven Priedon7, Amitava Banerjee1, Nikolas Pontikos8,9, Chris Tomlinson1, Ana Torralbo1, Paul Taylor1, Cathie Sudlow7,10,11, Spiros Denaxas1,7,10,12, Harry Hemingway1,7,12, Honghan Wu1,13

* First authors

Primary institutions:
1. Institute of Health Informatics, University College London, London, UK
2. Advanced Care Research Centre, Usher Institute, University of Edinburgh, Edinburgh, EH16 4TL, UK
3. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
4. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford, UK
5. Cicely Saunders Institute of Palliative Care, Policy & Rehabilitation, King’s College London, London, UK
6. College of Life Sciences, University of Leicester, Leicester, UK
7. Health Data Research UK, London, UK
8. UCL Institute of Ophthalmology, University College London, London EC1V 9EL, UK
9. Moorfields Eye Hospital NHS Foundation Trust, London EC1V 2PD, UK
10. British Heart Foundation Data Science Centre, London, UK
11. Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK
12. National Institute of Health Research University College London Hospitals Biomedical Research Centre (UCLH BRC), London, UK
13. Alan Turing Institute, London, UK

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.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

Background: The Global Burden of Disease study has provided key evidence to inform clinicians, researchers, and policy makers across common diseases, but no similar effort with single study design exists for hundreds of rare diseases. Consequently, many rare conditions lack population-level evidence including prevalence and clinical vulnerability. This has led to the absence of evidence-based care for rare diseases, prominently in the COVID-19 pandemic.

Method: This study used electronic health records (EHRs) of more than 58 million people in England, linking nine National Health Service datasets spanning healthcare settings for people alive on Jan 23, 2020. Starting with all rare diseases listed in Orphanet, we quality assured and filtered down to analyse 331 conditions with ICD-10 or SNOMED-CT mappings clinically validated in our dataset. We report 1) population prevalence, clinical and demographic details of rare diseases, and 2) investigate differences in mortality with SARs-CoV-2.

Findings: Among 58,162,316 individuals, we identified 894,396 with at least one rare disease. Prevalence data in Orphanet originates from various sources with varying degrees of precision. Here we present reproducible age and gender-adjusted estimates for all 331 rare diseases, including first estimates for 186 (56.2%) without any reported prevalence estimate in Orphanet. We identified 49 rare diseases significantly more frequent in females and 62 in males. Similarly we identified 47 rare diseases more frequent in Asian as compared to White ethnicity and 22 with higher Black to white ratios as compared to similar ratios in population controls. 37 rare diseases were overrepresented in the white population as compared to both Black and Asian ethnicities. In total, 7,965 of 894,396 (0.9%) of rare-disease patients died from COVID-19, as compared to 141,287 of 58,162,316 (0.2%) in the full study population. Eight rare diseases had significantly increased risks for COVID-19-related mortality in fully vaccinated individuals, with bullous pemphigoid (8.07[3.01-21.62]) being worst affected.

Interpretation: Our study highlights that National-scale EHRs provide a unique resource to estimate detailed prevalence, clinical and demographic data for rare diseases. Using COVID-19-related mortality analysis, we showed the power of large-scale EHRs in providing insights to inform public health decision-making for these often neglected patient populations.

Funding: British Heart Foundation Data Science Centre, led by Health Data Research UK.
Research in context

Evidence before the study: We have previously published the largest study looking at COVID-19 across rare diseases, but with a sample size of 158 COVID-19 infected rare disease patients and 125 unaffected relatives, from Genomics England, the power of that study was limited. We searched PubMed from database inception to Apr 21, 2023, for publications using the search terms “COVID-19” or “SARS-CoV-2” and “rare disease” or “ORPHANET”, without language restrictions. There are many studies examining the severity of COVID-19 in rare disease patients. However, to date, most studies have focused on a single or a few rare diseases associated with severity of COVID-19, and not taken a comprehensive rare disease wide approach. So far no studies have examined the impact of vaccination on mortality in rare disease patients. Moreover, the sample size used to examine rare diseases is limited in most studies. The largest study we identified included 168,680 individuals but only focused on autoimmune rheumatic disease.

Added value of this study: In this study we use national scale EHR data from England to report age and gender adjusted point prevalence for 331 rare diseases, with clinically-validated ICD-10 and/or SNOMED-CT code lists. Among these, 186 (56.2%) diseases did not have existing point prevalence data available in Orphanet. To our knowledge, this is the first time that rare diseases have been examined on a national scale, encompassing a population of over 58 million people. The large sample size provides sufficient statistical power to detect and describe enough carriers of even very rare conditions <1 case per million. Our analysis of COVID-related mortality has demonstrated the clinical relevance of national data for rare diseases. Specifically, we identified eight rare conditions that are associated with a significantly increased risk of mortality from COVID-19, even among fully vaccinated individuals.

Implication of all the available evidence: These findings provide robust reproducible prevalence, gender, and ethnicity estimates for disease that may often have been under prioritised, and where such information in most cases was not previously available. Our COVID-19 mortality findings highlight the need for targeted policy and support addressing the high level of vulnerability of these patients to COVID-19.
Introduction

A key conclusion of the largest study to date examining rare-disease prevalence, by Wakap et al. 2020, was that further registry research and rare-disease codification was needed to improve current prevalence estimates. In their study they examined 6,172 RDs listed in the Orphanet database, an international initiative for rare disease research, and estimated that they affect 3.5 - 5.9% of the world population (equal to 263-446 million or about 1 in 17 people) at any point in time\(^1\). However, they also pointed out that many prevalence estimates in Orphanet are either unknown, based on a small sample size, or expert opinion, this is likely to be imprecise.

RDs, many of which are debilitating and life-threatening, comprise a spectrum of heterogenous phenotypes sharing an often protracted diagnosis, scarce effective treatments, and broadly low quality of life\(^2,3\). Their low prevalence renders them inherently hard to study, with little incentive to invest in research on diagnostics and treatment for individual RDs. RD research has long been plagued by numerous challenges including: scarce data and reliance on case and family reports for single diseases, geographical variations, lack of definition consensus, use of heterogeneous epidemiological approaches for case ascertainment, and incomplete codification in healthcare systems\(^1\). Consequently, many RDs have unknown or imprecise prevalence estimates, preventing data-driven healthcare prioritisation which may be detrimental in events such as the COVID-19 pandemic, and hampering systematic assessment and comparison across healthcare systems.

Previous research has measured the indirect impact of COVID-19 on the life and care of RD patients, highlighting the exacerbated difficulties they face such as healthcare service disruption, difficulty of access due to social distancing and safety concerns, declining mental health due to prolonged diagnosis, treatment interference, and subsequently deteriorating symptoms\(^4\)-\(^7\). Direct quantification of the risk of severe COVID-19 outcomes in RD populations remains scant. Rutter et al. examined the increased COVID-19 related death rates among patients with rare autoimmune rheumatic diseases in England using electronic health records (EHRs)\(^8\). In a population-wide study conducted in Hong Kong, Chung et al. described an increase in risk of COVID-19 related death among hospitalised patients with RD\(^9\). Similar results were found with respect to the first COVID-19 wave in the UK in our previous study, using data from Genomics England\(^10\).

Risk factors and clinical characteristics continue to be studied to identify those at high risk of severe COVID-19 outcomes, with a growing emphasis on using whole population data for more precise insights\(^11\)-\(^14\). In many countries including the United Kingdom (UK), high risk groups such as people aged 60 and above, those who are immunocompromised, or with comorbidities associated with increased risk of adverse outcomes have been prioritised for vaccination. However, many patients with RDs were not considered extremely vulnerable due to the lack of reliable data on their risk of adverse COVID-19 outcomes. The UK shielded patient list (those with specified underlying conditions considered to make them clinically extremely vulnerable to the development of severe COVID-19 infections) was created in a data-driven way and refined using a risk prediction tool to consider additional risk groups\(^15,16\). While a few rare respiratory conditions (bronchiectasis, cystic fibrosis, alveolitis) and neurological conditions (myasthenia gravis, Huntington’s chorea, multiple sclerosis, motor neurone disease) were included in the model, the majority of RDs were unaccounted. This highlights the need to explore the adverse COVID-19 outcomes among RD patients in a large population cohort, and the necessity for improved and effective codification to capture and easily identify RDs in EHRs. In this context, population scale data is crucial to define and refine RD epidemiological estimates and the risk of severe COVID-19 outcomes in this group. This would be useful not only for risk management and vaccine prioritisation in the COVID-19 pandemic but also for the future health of these patients in general.

Using comprehensive and longitudinal patient health data for more than 58 million people in England, this study aims to 1) report reproducible population prevalence of RDs from EHRs and the demographic characteristics of
those individuals affected, 2) determine whether people affected by certain RDs have differences in mortality from COVID-19 infection during a time period spanning multiple variants and vaccine roll out.

**Methods**

**Data sources and Study population.**

This cohort study used nine linked datasets from the English National Health Service (NHS) available within the NHS Digital Trusted Research Environment (TRE), accessed through the CVD-COVID-UK / COVID-IMPACT Consortium. The datasets used were: primary care data from the General Practice Extraction Service Extract for Pandemic Planning and Research (GDPPR), COVID-19 testing data from the Public Health England Second Generation Surveillance System (SGSS), COVID-19 hospital admission data from Secondary Uses Service (SUS), Hospital Episode Statistics for admitted patient care (HES-APC), adult critical care (HES-CC), and outpatients (HES-OP) and the COVID-19 Hospitalisations in England Surveillance System (CHESS); COVID-19 vaccination status; and mortality information from the Office for National Statistics (ONS) Civil Registration of Deaths. A detailed description of these datasets has been published previously by Wood et al. 2021.

The study start date was 23rd January 2020, the date of the first recorded COVID-19 case in the UK, and the end date was 30th November 2021. We included individuals that were: a) alive at the start of the study, b) registered with a General Practitioner (GP) in England (minimum one patient record in GDPPR) between 1st of January 1990 (first year with good data coverage) and prior to 23rd January 2020, c) associated with a valid person pseudo-identifier enabling data linkage.

**Identification of participants with rare-diseases**

Full details on identification of rare-disease participants can be found in supplementary methods, a brief description follows below. Orphanet, an extensive online resource for RDs, was used to identify and define RDs, data was downloaded on the 6th of May 2022. We adopted a stepwise approach to identify RDs that could be accurately mapped to the disease codes available in our data sources. First, we extracted all Orphanet diseases with mappings to ICD-10 or SNOMED-CT, the clinical terminologies used in our data sources. Second, we applied the disease classification filter available on Orphanet to filter diseases to those included in ‘disorder types’: disease, morphological anomaly, malformation syndrome, or clinical syndrome (n = 5,864). Third, we excluded all diseases for which mapping to ICD10 was classified as “narrow to broad”, since our initial manual validation revealed that most if not all of these codes included a combination of rare and common diseases and hence would result in a high degree of misclassification (note that some of the diseases that were removed also had SNOMED-CT code mappings which were still included).

Fourth, we computed the point prevalence for all the eligible diseases in the TRE (n = 3,817 diseases with at least one affected individual), searching across the linked tables for SNOMED-CT codes in GDPPR, and ICD-10 codes in HES-APC and HES-OP. Fifth, manual curation was performed by a clinician to validate the accuracy of the matching of diseases to SNOMED-CT and ICD-10 codes by considering both the code definitions and the code frequencies from the EHR data. All SNOMED-CT codes were specific to the diseases but some ICD-10 codes were still too broad and introduced misclassification as they matched to both rare and common diseases, these were excluded. This resulted in the inclusion of 331 RDs with highly specific mapping, of which 164 had ICD-10 mappings, 140 had SNOMED-CT mappings and 27 had both (figure S1). Participants were defined as having a rare disease if they had one or more codes of each disease’s ICD-10 or SNOMED-CT codes.
Estimation of point prevalence of rare diseases in the English population

Point prevalence, estimates per 1,000,000 individuals, was calculated including anyone who were alive and had a RD diagnosis between January 1990 and the study start. Age and gender adjusted prevalence estimates were calculated using the 2021 census data from England as the reference population20.

We compared our gender- and age-adjusted estimated prevalence with the point prevalence of RD provided by Orphanet. Giving preference to the Orphanet estimated from the United Kingdom, followed by European estimates (regional or national specific), and worldwide estimates last.

For RDs in our sample identified in 5 or fewer individuals, adjusted prevalence estimates could not be given due to the risk of re-identification; these were compared to Orphanet as having a prevalence of 0.1 per million. To assess the quality and strength of evidence of the Orphanet prevalence data sources, we manually reviewed their study design and methodology.

COVID-19 phenotyping and vaccine data

To determine the impact of COVID-19 on RD patients we used five previously defined COVID-19 phenotypes21. In brief we identified 1) positive SARS-CoV-2 tests, 2) COVID-19 diagnosis recorded in primary care, 3) hospital admissions with a COVID-19 diagnosis, 4) ventilatory support related to COVID-19, and 5) COVID-19 mortality (Supplementary methods for details). Onset of COVID-19 was defined as the date of the earliest COVID-19-event (any of the five phenotypes defined above) for each individual. Date of the outcome was defined as the date of the earliest record of COVID-19 mortality. Time to the event was calculated as the day difference between the date of onset and date of the outcome.

Vaccination status was determined from the COVID-19 vaccination dataset, including all vaccinations administered after Dec 12, 2020 (when the first official dose was administered in England). Patients were classified as fully vaccinated once 14 days had elapsed after their second dose.

Demographic data

Date of birth, sex and ethnicity were extracted from GDPPR and HES datasets. Age at first RD diagnosis and COVID-19 was calculated as the difference between date of birth and the relevant first RD diagnostic code and/or the first COVID event divided by 365.25. Ethnicity was categorised as per the ONS22: Asian or Asian British, Black or Black British, mixed, white, other ethnicities and unknown, mapping ethnicity across primary and secondary care prioritising information from primary care23.

Fisher's exact test was performed to test for difference in sex and ethnicity ratios for the 219 RD with 100 or more affected. Comparing the ratio of patients with a specific RD with the ratio in all unaffected (with regards to the specific condition) in the study population. Ethnicity comparison was done relative to the majority White ethnicity group. Bonferroni correction for multiple testing was applied to give a statistical significance threshold of 0.000228 (0.05 / 219).

COVID-19-related mortality risk analysis and shielding analysis

We performed a retrospective cohort analysis comparing COVID-19 related mortality in people with a certain RD or a RD category, with matched controls from the general population. Differences in COVID-19 related mortality were addressed with a time-to-event analysis. See supplement methods, for full detail of the following analysis. In brief, The study period for assessing COVID-19 related mortality was set from 2020-09-01 to 2021-11-30. Cohorts were formed for each RD and RD category, using exact matching on age group, sex, ethnicity and
vaccination at the ratio of two controls (not affected by the condition) per individual with a RD. Survival functions were estimated using Kaplan–Meier estimator.

To examine the impact of governmental shielding recommendation on rare-diseases, we compared rare diseases with high risk of COVID-19-related deaths and diseases in the NHS digital shielding list\(^2^4\).

Results

Identifying rare disease patients

From the 10,563 clinical entities listed on Orphanet (6th of May 2022), we identified and examined 331 disorders where clinical mapping to either SNOMED-CT or ICD-10 was of high specificity (see method section for details) see figure S1.

Our study population comprised 58,162,316 individuals registered with a GP in England and were alive at the onset of the pandemic on January 23, 2020\(^1^8\). In this cohort we identified 894,396 individuals with at least one of the 331 rare disorders diagnosed prior to the onset of the pandemic, giving a prevalence of 15,377 per million. In total 7,965 of 894,396 (0.9%) of rare-disease patients died from COVID-19, as compared to 141,287 of 58,162,316 (0.2%) in the full study population cohort from the start of the pandemic until 30th November 2021.

We observed 3,359,077 clinical codes (384 unique codes) associated with the 331 conditions. The majority of these 2,816,568 (83.8%, 216 unique codes) were ICD-10 codes recorded during admitted patient care, 372,620 (11.1%, 167 unique codes) were SNOMED-CT codes from general practitioners and 169,889 (5.1%, 206 unique codes) were ICD-10 codes from hospital outpatient care. Where diagnostic codes were available from both primary and secondary care (ICD-10 and SNOMED), we obtained comparable agreement in prevalence between sources especially for the more frequent conditions such as Myasthenia gravis and Aorta (figure S2).

Numbers of identified patients for each condition and point prevalence adjusted for age and gender for all examined RDs can be found in table 1, and data S1. As expected our age and gender adjusted point prevalence was highly correlated with our raw point prevalence estimates, see figure S2. The most frequent RD in our cohort was Polymyalgia rheumatica, a condition that causes pain, stiffness and inflammation in muscles. This condition was predominantly identified in older individuals with 84.5% being above 70. With an adjusted point prevalence of 2831.8 per million (n = 158,648) this condition is more prevalent than what’s considered a RD, according to the UK prevalence definition of 500 or less cases per million\(^2^5\).
Table 1: Demographics of 21 selected rare diseases highlighting different prevalence spectra. Table is sorted first by reported Orphanet prevalence per million individuals, category and then study prevalence from low to high. *Adjusted population prevalence is age and gender adjusted prevalence per million individuals. Age is presented as the median age of first diagnosis (25% quantile and 75% quantile) and as numbers and percentages of cases within age groups calculated at study start. Age and sex was available for all diseases. Data for all rare-diseases studied can be seen in data S1 (link to google sheet).

<table>
<thead>
<tr>
<th>Rare disease</th>
<th>Prevalence</th>
<th>Gender</th>
<th>Age</th>
<th>Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>female</td>
<td>male</td>
<td></td>
</tr>
<tr>
<td>Leprechaunism</td>
<td>&lt;10</td>
<td>&lt;10</td>
<td>&lt;10</td>
<td></td>
</tr>
<tr>
<td>Seckel syndrome</td>
<td>59</td>
<td>34(57.6)</td>
<td>25(42.4)</td>
<td></td>
</tr>
<tr>
<td>Wells syndrome</td>
<td>577</td>
<td>247(42.8)</td>
<td>300(57.2)</td>
<td></td>
</tr>
<tr>
<td>Pломiomyelitis</td>
<td>1290</td>
<td>614(47.6)</td>
<td>676(52.4)</td>
<td></td>
</tr>
<tr>
<td>Lafora disease</td>
<td>&lt;10</td>
<td>&lt;10</td>
<td>&lt;10</td>
<td></td>
</tr>
<tr>
<td>Wolfraam syndrome</td>
<td>125</td>
<td>17(26.6)</td>
<td>19(33.3)</td>
<td></td>
</tr>
<tr>
<td>Tibial hemimelia</td>
<td>329</td>
<td>188(57.1)</td>
<td>141(42.9)</td>
<td></td>
</tr>
<tr>
<td>Moyaamoya disease</td>
<td>734</td>
<td>274(37.3)</td>
<td>460(62.7)</td>
<td></td>
</tr>
<tr>
<td>Fibular hemimelia</td>
<td>283</td>
<td>168(59.4)</td>
<td>115(40.6)</td>
<td></td>
</tr>
<tr>
<td>Angelman syndrome</td>
<td>1250</td>
<td>601(48.1)</td>
<td>649(51.9)</td>
<td></td>
</tr>
<tr>
<td>Hemophilia B</td>
<td>2004</td>
<td>1257(62.7)</td>
<td>747(37.3)</td>
<td></td>
</tr>
<tr>
<td>Intestinal communication</td>
<td>107513</td>
<td>53179(49.5)</td>
<td>54328(50.5)</td>
<td></td>
</tr>
<tr>
<td>Alopecia totalis</td>
<td>809</td>
<td>278(34.4)</td>
<td>531(65.6)</td>
<td></td>
</tr>
<tr>
<td>Huntington disease</td>
<td>6088</td>
<td>2836(46.6)</td>
<td>3252(53.4)</td>
<td></td>
</tr>
<tr>
<td>Hemophilia A</td>
<td>11635</td>
<td>7047(60.6)</td>
<td>4587(39.4)</td>
<td></td>
</tr>
<tr>
<td>Intestinal cystitis</td>
<td>22701</td>
<td>2648(11.7)</td>
<td>20053(88.3)</td>
<td></td>
</tr>
<tr>
<td>Good syndrome</td>
<td>10</td>
<td>6732(29.1)</td>
<td>655(58.7)</td>
<td></td>
</tr>
<tr>
<td>Trigeminal neuralgia</td>
<td>43396</td>
<td>13822(31.9)</td>
<td>29573(68.1)</td>
<td></td>
</tr>
<tr>
<td>Infantile apnea</td>
<td>49966</td>
<td>26722(53.5)</td>
<td>23242(46.5)</td>
<td></td>
</tr>
<tr>
<td>Polymyalgia rheumatic</td>
<td>158648</td>
<td>48551(30.6)</td>
<td>110097(69.4)</td>
<td></td>
</tr>
</tbody>
</table>

*Adjusted population prevalence is age and gender adjusted prevalence per million individuals. Age is presented as the median age of first diagnosis (25% quantile and 75% quantile) and as numbers and percentages of cases within age groups calculated at study start. Age and sex was available for all diseases.
Among all our studied RDs, patients were more often female (55% vs 50.2%), under 18 (25.2% vs 19.9%) or over 70 (28.2% vs 13.1%) as compared to the full study population. There were fewer individuals from the Asian or Asian British (6.7% vs 9.6%) and Black or Black British (2.9% vs 3.9%), communities as compared to the full study population. There were also fewer individuals with unknown ethnicity, which was to be expected from the increased healthcare use and thus registration in the RD group. As expected, RD patients are more likely to be on the shielding list 23.5%, had a SNOMED-CT code indicating shielding, as compared to 7% of the full study population (table 2).

We compared 219 RD where more than 100 affected individuals were identified, for differences in gender and ethnicity ratios as compared to unaffected in our study population. After Bonferroni correction for multiple testing, we found 111 RD which had significant gender differences. 49 being more prominent in females (replicating known findings for: Rett syndrome and Interstitial cystitis) and 62 being more prominent in males (examples: Kennedy's disease and Reactive arthritis). 100 RD had significant differences between Asian and white ethnicities, with 47 RD being overrepresented in individuals of Asian ethnicity, (Lamellar ichthyosis and Maple urine disease topped this list). Furthermore, 75 RD had significant differences between Black and white ethnicities, with 22 RD found in higher ratios in the Black community; Moyamoya disease and Quinquaud folliculitis decalvans had the largest ratios in the Black ethnicities. 37 RD were overrepresented in the white population as compared to both Black and Asian ethnicities (see supplementary data2).

### Table 2: Demographics table (Link to google sheets version)

<table>
<thead>
<tr>
<th>Overall</th>
<th>Study population</th>
<th>All Rare-disease patients</th>
<th>Rare-disease with COVID-19</th>
<th>Rare-disease with COVID-19 Hospitalisation</th>
<th>Rare-disease with COVID-19 Ventilatory Support</th>
<th>Rare-disease with COVID-19 Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>58162316</td>
<td>894396</td>
<td>128608</td>
<td>17750</td>
<td>2924</td>
<td>7965</td>
</tr>
<tr>
<td>COVID Deaths</td>
<td>141287 (0.2)</td>
<td>7965 (0.9)</td>
<td>7965 (6.2)</td>
<td>4641 (26.1)</td>
<td>1322 (45.2)</td>
<td>7965 (100.0)</td>
</tr>
<tr>
<td>All Deaths (%)</td>
<td>880041 (1.5)</td>
<td>46273 (5.2)</td>
<td>11746 (9.1)</td>
<td>6517 (36.7)</td>
<td>1336 (45.7)</td>
<td>7557 (94.9)</td>
</tr>
<tr>
<td>Sex (%)</td>
<td>Female (50.2)</td>
<td>492200 (55.0)</td>
<td>72476 (56.4)</td>
<td>10377 (58.5)</td>
<td>1464 (50.1)</td>
<td>4425 (55.6)</td>
</tr>
<tr>
<td>Male (%)</td>
<td>28985457 (49.8)</td>
<td>402171 (45.0)</td>
<td>56129 (43.6)</td>
<td>7372 (41.5)</td>
<td>1460 (49.9)</td>
<td>3540 (44.4)</td>
</tr>
<tr>
<td>Age (%)</td>
<td>Under 18</td>
<td>225397 (25.2)</td>
<td>35720 (27.8)</td>
<td>762 (4.3)</td>
<td>185 (6.3)</td>
<td>17 (0.2)</td>
</tr>
<tr>
<td></td>
<td>18 - 29</td>
<td>104538 (11.7)</td>
<td>18525 (14.4)</td>
<td>590 (3.3)</td>
<td>110 (3.8)</td>
<td>31 (0.4)</td>
</tr>
<tr>
<td></td>
<td>30 - 49</td>
<td>138246 (15.5)</td>
<td>24234 (18.8)</td>
<td>1424 (8.0)</td>
<td>397 (13.6)</td>
<td>161 (2.0)</td>
</tr>
<tr>
<td></td>
<td>50 - 69</td>
<td>22172 (17.2)</td>
<td>3324 (18.7)</td>
<td>1072 (36.7)</td>
<td>1057 (13.3)</td>
<td>606 (15.1)</td>
</tr>
<tr>
<td></td>
<td>70 or over</td>
<td>72957 (21.7)</td>
<td>11650 (65.6)</td>
<td>1160 (39.7)</td>
<td>6699 (84.1)</td>
<td>130 (19.3)</td>
</tr>
<tr>
<td>Ethnic group (%)</td>
<td>White</td>
<td>777730 (87.0)</td>
<td>111746 (86.9)</td>
<td>15786 (88.9)</td>
<td>2362 (80.8)</td>
<td>7310 (91.8)</td>
</tr>
<tr>
<td></td>
<td>Asian or Asian British</td>
<td>59717 (6.7)</td>
<td>9584 (7.5)</td>
<td>1081 (6.1)</td>
<td>316 (10.8)</td>
<td>359 (4.5)</td>
</tr>
<tr>
<td></td>
<td>Black or Black British</td>
<td>25935 (2.9)</td>
<td>3172 (2.5)</td>
<td>511 (2.9)</td>
<td>146 (5.0)</td>
<td>159 (2.0)</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>16652 (1.9)</td>
<td>2337 (1.8)</td>
<td>166 (0.9)</td>
<td>54 (1.8)</td>
<td>47 (0.6)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>1387 (1.1)</td>
<td>180 (1.0)</td>
<td>42 (1.4)</td>
<td>67 (1.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>373 (0.3)</td>
<td>26 (0.1)</td>
<td>&lt;10</td>
<td>24 (0.3)</td>
<td></td>
</tr>
<tr>
<td>Social deprivation fifths (%)</td>
<td>1 (most deprived)</td>
<td>11993725 (20.6)</td>
<td>188047 (21.0)</td>
<td>29696 (23.1)</td>
<td>4271 (24.1)</td>
<td>731 (25.0)</td>
</tr>
<tr>
<td></td>
<td>5 (least deprived)</td>
<td>11103838 (19.1)</td>
<td>171851 (19.2)</td>
<td>23327 (18.1)</td>
<td>2967 (16.7)</td>
<td>448 (15.3)</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>662 (0.1)</td>
<td>60 (0.0)</td>
<td>&lt;10</td>
<td>&lt;10</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Comorbidities (%)</td>
<td>High risk (Shielding List)</td>
<td>4096093 (7.0)</td>
<td>210571 (23.5)</td>
<td>32394 (25.2)</td>
<td>9578 (54.0)</td>
<td>1614 (55.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3397 (42.6)</td>
</tr>
</tbody>
</table>

### Comparison of point prevalences with Orphanet

Of the 331 RDs 186 (56.2%) did not have point prevalence data available from Orphanet. Figure 2, compares the gender- and age-adjusted point prevalence observed in our sample with prevalence range estimates from Orphanet, for the 145 (43.8%) diseases with data. This showed that 86 of 145 (59.3%) diseases fall within the same prevalence ranges reported by Orphanet. 25 (17.2%) diseases have higher estimates, with the largest discrepancy observed for interatrial communication: 1,848 per million in our study versus 10-99 per million in Orphanet. 34 (23.4%) of our estimates were
lower than Orphanet. For instance, prevalence for chronic actinic dermatitis was 2 per million, while the higher Orphanet estimate is 100-500 per million.

To better understand these differences we explored the Orphanet data source of point prevalence estimates for the 145 RDs with data. The data sources were variable and most did not present reproducible method definitions; 63 (43%) had “expert” or “orphanet” as the only source, and could not be verified. Of the 83 with a PubMed Unique Identifier, 23 (28% of 83) were case reports, and 20 (24% of 83) were systematic reviews which might be referring to a case report or an epidemiological study. 28 (34% of 83) referred back to articles where an epidemiological study was conducted, for 13 (46% of 28) of those no sample size could be found and 3 had no reference data (Table 3).

Figure 2. Comparing gender- and age-adjusted rare disease point prevalence estimates from the study data with Orphanet prevalence ranges. The 331 diseases are stratified in groups of relative rarity going from more common to more rare on the x-axis, based on Orphanet estimates. The points indicate our estimated age and gender adjusted point prevalences from population data in England. The labelled diseases are the same 21 selected rare diseases highlighted in table 1. The grey boxes indicate the range of prevalence reported for each disease in Orphanet data. The dashed line represents a prevalence of 1:2000 the upper limit definition for a rare disease suggested by the UK government. COVID19-related mortality in people with rare diseases

CC-BY-NC-ND 4.0 International license
It is made available under a CC-BY-NC-ND 4.0 International license.
Table 3. Overview of Orphanet prevalence data sources, used in figure 2 for the 145 (out of 331) rare diseases that had available data. Source types are not mutually exclusive, some rare diseases have multiple sources listed. Unclear Little information from these 3 as there is no reference data. Numbers in parentheses are column wise percentages.

<table>
<thead>
<tr>
<th>Orphanet prevalence (per million)</th>
<th>&lt;1</th>
<th>1-9</th>
<th>10-99</th>
<th>100-500</th>
<th>All diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>52</td>
<td>27</td>
<td>40</td>
<td>26</td>
<td>145</td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worldwide</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source type*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Review</td>
<td>7</td>
<td>3</td>
<td>7</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Article (observational study, survey)</td>
<td>6</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>Case series report</td>
<td>16</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>Guideline</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>European Medicine Agency Report</td>
<td>24</td>
<td>10</td>
<td>22</td>
<td>7</td>
<td>63</td>
</tr>
<tr>
<td>Expert only</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>RCT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Unclear**</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

COVID-19 mortality risk

We compared the risk of COVID-19 related death in people with RDs with that of the matched controls from the general population. Of the 331 RDs, compared to matched controls, 8 diseases in vaccinated individuals and 28 in unvaccinated individuals (Figure 3, Supplementary data 3) had significantly increased risk for COVID-19 related death. Of the 25 RD categories (table S1), we observed significantly increased COVID-19 related mortality among vaccinated patients in 11 categories (including 323 diseases of which 179 were included in the analysis) and unvaccinated individuals in 14 categories (including 325 diseases of which 181 were included in the analysis) (Supplementary data 4).

Figure 3: Forest plot showing Log Hazard Ratio of COVID-19-related death for rare diseases and categories with significant increase of risk in fully vaccinated individuals, based on lower-bound of the 95% confidence intervals (95% CI) of the hazard ratio (HR) between individuals with rare diseases (RD) and matched controls (CTRL). HRs and 95% CIs are displayed numerically on the right.
For vaccinated individuals, the three worst affected RDs were bullous pemphigoid (HR [95% CI] = 8.07[3.01-21.62], an autoimmune bullous skin disease), osteogenesis imperfecta (10.14[2.22-46.26], a rare, genetic, primary bone dysplasias), autosomal dominant polycystic kidney disease (4.06[2.03-8.13], a genetic, renal tubular disease). For unvaccinated individuals, the three worst affected RDs were progressive supranuclear palsy (4.11[2.78-6.09], a late-onset neurodegenerative disease), infantile spasms syndrome (16.14[2.02-129.04], a rare epilepsy syndrome) and severe intellectual disability-progressive spastic diplegia syndrome (3.41[1.94-6.02], a genetic, syndromic intellectual disability disorder) (Figure 4). Of the 28 RDs with significantly increased risk of COVID-19 related death in unvaccinated individuals, only Interatrial communication was included in the list of shield diseases.

For vaccinated individuals, the three worst affected RD categories were infertility disorders (7.54[3.27-17.37]), renal diseases (5.66[2.64-12.14]), and ophthalmic diseases (2.89[2.07-4.05]). For unvaccinated individuals, the three worst affected RD categories were renal diseases (2.27[1.72-3.00]), infertility diseases (2.21[1.62-3.00]) and ophthalmic diseases (1.82[1.62-2.04]).

Figure 4: Kaplan-Meier Plot for the most affected rare diseases or categories stratified by if individuals were fully vaccinated, based on lower-bound of the 95% confidence intervals (95% CI) of the hazard ratio (HR) between individuals with rare diseases (RD) and matched controls (CTRL). Events are defined as COVID-related death. Kaplan-Meier Plot for all conditions can be seen in supplementary data 5.
Figure 5: (A) Hazard ratio for COVID-19 mortality for 290 Rare diseases in relation to median age at first diagnosis and prevalence. Point sizes are relative to the estimated Hazard Ratio (HR) for COVID-19 mortality. Colours represents the method by which the HR was estimated, and are prioritised in the following order based on what significant HR estimate is available for each disease; (Red, n = 8) estimates from fully vaccinated affected individuals of a specific rare disease, (Orange, n = 20) estimates from none fully vaccinated affected individuals of a specific rare diseases, (Light red, n = 296) estimates based on the highest HR from any of the rare disease category the specific disease belongs to, (dark grey, n = 7) no significant association found for the specific disease or disease category.

Age and gender adjusted prevalence rate per million is shown on log10 scale. (B) Number of diseases with at least 10, 100 and 1000 identified rare disease patients, observed in our current study sample of ~58 million, and the estimated numbers of patients if we project our rare disease prevalence into sample sizes of 25, 10, 5 million populations. All current study prevalence estimates are available for all diseases as supplementary data 1.

Sample size influences power to detect significant HR for COVID-19 mortality in rare disease patients. All significant HR ratios for specific rare diseases were found in disorders which had an adjusted prevalence of more than 17 per million (both for vaccinated and unvaccinated). Significant HR ratios for vaccinated individuals was also only seen in diseases with a later age of onset (median age of first diagnosis > 34 years), as fewer COVID-19 mortalities were observed in younger individuals. However, when considering significant HR estimates based on individuals affected by any rare disease belonging to a certain category, the identified categories with significant HR for COVID-19 mortality in fully vaccinated covered 324 of the 331 rare diseases we examined (Figure 5, panel A).

Lastly, projection of our estimated prevalences into smaller population sizes of 25, 10 and 5 million individuals showcase the need for large scale linked EHR resources in order to identify sufficient rare diseases patients to meaningfully describe these in terms of key demographic strata (Figure 5, panel A).
Discussion

This study uses linked EHR data for more than 58 million people in England, to provide detailed information for 331 RDs and explore the impact of RDs on COVID-19 mortality. We provide prevalence estimates for 184 RDs which previously did not have such data available from Orphanet. We present, fully reproducible, age and sex adjusted point prevalence estimates for all 331 diseases, including counts of the specific clinical codes observed in our population.

We highlight eight specific RDs where significant increased mortality was observed in fully vaccinated individuals compared to matched individuals from the general population. This is relevant both for countries with high vaccination uptake and those where vaccination rollout is still falling behind. Increased mortality was also seen for 11 out of 25 categories of RD, including RDs where the sizes of individual-disease cohorts did not permit single disease analyses. This paper demonstrates the increased power in RD epidemiology, when linked national data are made available for research.

Rare-disease selection and case identification

The 331 RDs described in this study is just a fraction of the total number of RDs which have been described in the published literature. Orphanet lists 10,563 clinical entities in their RD database, but many of these (28%) are currently unmapped to clinical codes (ICD-10, SNOMED-CT), complicating case ascertainment from clinical codes. A large proportion (52.5%) of Orphanet listed RD were dropped from our study, either because none of the related codes were observed in our cohort (20.1%), or because the frequency of the mapped clinical codes were excessively common, due to a too broad definition (32.3%). This highlights that even with population wide data, better mappings, to more specific codes, would significantly improve case identification. The more fundamental issue is that classification systems like ICD-10 do not possess the granularity for all RD, or the granularity is not recorded in the clinic. An example of this being that ICD-10 coding is often only recorded to three digits, where the fourth digit level of ICD-10-CM would provide more precision. Orphanet currently only includes mappings to three digit ICD-10 codes in their mappings.

Comprehensive clinical terminologies such as SNOMED-CT or the upcoming ICD-11 may better meet the need for increased specificity, but the use of very specific codes in contemporary health systems remains inconsistent, and systematic efforts will be needed to promote this. For better case identification, previous work has shown the great potential of free-text components of EHRs, which does however come with its own set of challenges. Phenotypes associated with a disease constitute a computational disease model, is another promising approach to identifying undercoded/undiagnosed RD patients or patients sharing similar vulnerabilities. Both approaches, however, would require more comprehensive and granular EHR data, which currently are not available in national registries, or at least not on a population scale.

Prevalence comparisons to Orphanet

The European Organization for RDs (Eurordis) and Orphanet have made great progress in assessing the prevalence of RD. However, it is suggested that more reliable data as well as high efficiency with regards to the methodological quality of epidemiological studies is required for the future. Compared to their study, our approach can be generalised for rapid response and continuous monitoring of RD epidemiologic trends, thus filling a gap in the current literature about RD.

The majority of the RD prevalences estimated in our study 184 (55.6%) was not previously known in OrphaData. A significant number of examined Orphanet estimates are not backed by data-driven evidence. Listed sources are not re-assessed, which limits the systematic comparison of estimates due to low inter-study consistency and poor documentation of study methodology. Among diseases with prevalence curated in Orphadata, most of the disease prevalence were consistent (59.6%), although the prevalence of some diseases differed dramatically compared to our study, with both over and under-estimates.

Differences in prevalence estimates may have many causes. Traditionally, prevalence of RD was estimated indirectly from reported incidence or from meta-analysis of multiple sources of prevalence. Such estimates may be low if there is a poor coverage of incidences. It can also be difficult to estimate the denominator, leading to both under and over-
estimates. Some diseases, like interatrial communication, may also be underreported and underestimated due to their asymptomatic nature in early and mild cases. Additionally, the prevalence studies referenced in Orphanet comprise a range of heterogeneous epidemiological approaches, limiting the degree of inter study comparability. For instance, the population from which the published estimate for Chronic Actinic Dermatitis was obtained was restricted to individuals aged 75 and older. Our results highlight the importance of data-driven, reproducible, and age and gender adjusted estimation of RD point prevalence to inform healthcare service planning, clinical trials, pharmacological research, and to allow for precise comparison between studies.

The difference in gender, and ethnicity ratios found for a range of diseases, both replicate known findings and highlight a range of potentially novel insights with regards to differences in disease occurrence between groups, and/or in what may in some cases be disease disparities, indicating inequalities of disease burdens or underdiagnosis in subpopulation groups.

**Strengths and limitations**

Statistical power is a major limitation for epidemiologic studies investigating RD. For studies of COVID-19, this issue becomes even more pronounced when the numbers of COVID-19 related mortalities are low, especially in the analyses where individuals were fully vaccinated (visible by larger error bars of the corresponding hazard ratios). Our study utilised large-scale linked EHR data, which greatly improved the statistical power in our analyses.

Our analysis only included demographic covariates for cohort matching. One consideration for this approach is that it is often hard to distinguish whether a non-demographic covariate mediates the effect of RD on risks of COVID-19 related death. Conditioning on such covariates may lead to misleading messages that the people with the RD did not have increased COVID-related mortality. However, for individual RDs, there may be disease-specific covariates that can effectively further stratify people into different risk groups, curation of such information for different diseases is beyond the scope of the current study. We will welcome collaboration of domain-experts to further investigate their diseases of interest.

Despite the comprehensive nature of GDPPR, our cohort included more individuals than the 2021 ONS census population for England (56 489 800), reflecting the lack of registration data which may allow temporary residents to persist in the dataset, elevating the population denominator and leading to an underestimate of prevalence. Additionally GDPPR contains only a limited subset of SNOMED-CT codes, those which were thought to be related to COVID-19 research, and therefore may underestimate RD prevalence.

It should also be noted that the current analyses focused on the difference of COVID-19 related mortality between people with RDs and the general population. Therefore, the results cannot be seen as an analysis of vaccine effectiveness. Estimation of vaccine effectiveness will be done in a separate study. Further analyses looking at additional outcomes, such as hospitalisation and ventilatory treatment, are also being explored but were out of scope for this study.

**Conclusion**

National scale EHRs offer an unprecedented resource to study the epidemiology of RD that may not be possible using other sources. Such basic epidemiological descriptors are crucial for improved planning and prioritisation of RD diagnosis and treatment at national level. This study highlights diseases associated with increasing risk for COVID-19 mortality, which were under-protected by the shielding lists in England at the outbreak of the pandemic. In the current study, we directly and systematically estimated the prevalence of RDs using data-driven methods from large scale linked EHR data. The use of linked primary and secondary data ensured the coverage of RD cases across time and locations. Moreover, our approach is high throughputs, allowing the estimation of prevalences of many diseases at the same time, with consistent methodology.
Acknowledgements

This work is carried out with the support of the BHF Data Science Centre led by HDR UK (BHF Grant no. SP/19/3/34678). This work uses data provided by patients and collected by the NHS as part of their care and support. We would like to acknowledge all data providers who make anonymised data available for research.

The views expressed are those of the authors and not necessarily those of the organisations listed. The funders of this work played no role in the collection, analysis, or interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: support from the funders listed above; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work. SH works as a data scientist and data curator for NHS Digital, which holds and processes the data.

Code: The analytical code and protocol are available under an open source licence at the following repository: https://github.com/BHFDSC/CCU019_01

Data sharing: The authors and colleagues across the CVD-COVID-UK consortium have invested considerable time and energy in developing the data resource described here and are keen to ensure that it is used widely to maximise its value. For inquiries about data access, please see www.healthdatagateway.org/dataset/7e5f0247-f033-4f98-aed3-3d7422b9dc6d or email bhfdsc@hdruk.ac.uk.

Dissemination to participants and related patient and public communities: Results will be disseminated through the British Heart Foundation (BHF) Data Science Centre and CVD-COVID-UK webpages on the Health Data Research UK website, BHF communication channels, the BHF Data Science Centre’s lay members panel, and NHS Digital communications channels.

The study team would like to thank the BHF Data Science Centre’s lay members panel for their input and NHS DAE output checkers Lisa Grat, James Walker, Hanna McLean and others.

Ethical and Regulatory Approvals

The North East-Newcastle and North Tyneside 2 research ethics committee provided ethical approval for the CVD-COVID-UK/COVID-IMPACT research programme (REC No 20/NE/0161). This has been described in detail previously. Data access approval was granted to the CVD-COVID-UK consortium (under project proposal CCU013 High-throughput electronic health record phenotyping approaches) through the NHS Digital online Data Access Request Service (ref. DARS-NIC-381078-Y9C5K). NHS Digital data have been made available for research under the Control of Patient Information (COPI) notice which mandated the sharing of national electronic health records for COVID-19 research (more info: https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/control-of-patient-information-copi-notice). For further detail see supplementary methods.

Funding

The British Heart Foundation Data Science Centre (grant No SP/19/3/34678, awarded to Health Data Research (HDR) UK) funded co-development (with NHS Digital) of the trusted research environment, provision of linked datasets, data access, user software licences, computational usage, and data management and wrangling support, with additional contributions from the HDR UK data and connectivity component of the UK governments’ chief scientific adviser’s national core studies programme to coordinate national covid-19 priority research. Consortium partner organisations funded the time of contributing data analysts, biostatisticians, epidemiologists, and clinicians.
This work was funded by the Longitudinal Health and Wellbeing COVID-19 National Core Study, which was established by the UK Chief Scientific Officer in October 2020 and funded by UK Research and Innovation (grant references MC_PC_20030 and MC_PC_20059).

This work was supported by National Institute for Health and Care Research (NIHR202639), NIHR/HDR UK Winter Pressure Award (WP0006) and Medical Research Council (MR/S004149/2). 

The work is also founded by the HDR UK Discretionary fund - Rare Disease Phenomics (TF2022.42), which receives its funding from HDR UK Ltd (HDRUK 2022.0137) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Wellcome Trust.

This work was supported by National Institute of Health Research University College London CL Hospitals Biomedical Research Centre (UCLH BRC), London, UK, Health Data Research UK, which receives its funding from HDR UK Ltd (HDR-9006) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Wellcome Trust.
References


6. Rare Barometer (EURORDIS survey initiative). How has COVID-19 impacted people with rare diseases?.


