Prevalence and demographics of 331 rare diseases and associated COVID-19-related mortality among 58 million individuals: a nationwide retrospective observational study

Johan H Thygesen, Huayu Zhang*, Hanane Issa, Jinge Wu, Tuankasfee Hama, Ana-Caterina Phiho-Gomes, Tudor Groza, Sara Khalid, Thomas R Lumbers, Mevhibe Hocaoglu, Kamlesh Khunti, Rouven Priedon, Amitava Banerjee, Nikolas Pontikos, Chris Tomlinson, Ana Torralbo, Paul Taylor, Cathie Sudlow, Spiros Denaxas, Harry Hemingway, Honghan Wu, on behalf of the CVD-COVID-UK/COVID-IMPACT Consortium*

Summary

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 a single-study design exists for hundreds of rare diseases. Consequently, for many rare conditions there is little population-level evidence, including prevalence and clinical vulnerability, resulting in an absence of evidence-based care that was prominent during the COVID-19 pandemic. We aimed to inform rare disease care by providing key descriptors from national data and explore the impact of rare diseases during the COVID-19 pandemic.

Methods In this nationwide retrospective observational cohort study, we used the electronic health records (EHRs) of more than 58 million people in England, linking nine National Health Service datasets spanning health-care settings for people who were alive on Jan 23, 2020. Starting with all rare diseases listed in Orphanet (an extensive online resource for rare diseases), we quality assured and filtered down to analyse 331 conditions mapped to ICD-10 or Systemized Nomenclature of Medicine–Clinical Terms that were clinically validated in our dataset. For all 331 rare diseases, we calculated population prevalences, analysed patients' clinical and demographic details, and investigated mortality with SARS-CoV-2. We assessed COVID-19-related mortality by comparing cohorts of patients for each rare disease and rare disease category with controls matched for age group, sex, ethnicity, and vaccination status, at a ratio of two controls per individual with a rare disease.

Findings Of 58 162 316 individuals, we identified 894 396 with at least one rare disease and assessed COVID-19-related mortality between Sept 1, 2020, and Nov 30, 2021. We calculated reproducible estimates, adjusted for age and sex, for all 331 rare diseases, including for 186 (56·2%) conditions without existing prevalence estimates in Orphanet. 49 rare diseases were significantly more frequent in female individuals than in male individuals, and 62 were significantly more frequent in male individuals than in female individuals; 47 were significantly more frequent in Asian or British Asian individuals than in White individuals; and 22 were significantly more frequent in Black or Black British individuals than in White individuals. 37 rare diseases were significantly more frequent in the White population compared with either the Black or Asian population. 7965 (0·9%) of 894 396 patients with a rare disease died from COVID-19, compared with 141 287 (0·2%) of 58 162 316 in the full study population. In fully vaccinated individuals, the risk of COVID-19-related mortality was significantly higher for eight rare diseases, with patients with bullous pemphigoid (hazard ratio 8·07, 95% CI 3·01–21·62) being at highest risk.

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.

Introduction

Rare diseases are a worldwide health-care challenge, with more than 10 000 known to exist,[1] affecting 3·5–5·9% of the global population.[2] Rare diseases are often disabling, life limiting, and expensive to manage and have a devastating impact on patients, their families and carers, and the health-care system.[3] Most rare diseases have a genetic component,[2] with serious multisystem facets that require a disproportionate number of health-care resources.[4] More crucially, a UK report estimated that about 30% of patients with a rare disease die before the age of 5 years.[5] Patients will often spend more than 5 years on an emotionally difficult diagnostic journey.[6] Such an experience not only has adverse consequences

Lancet Digit Health 2025; 7: e145–56

*Joint first authors

Institute of Health Informatics, University College London, London, UK (J H Thygesen PhD, H Issa MSc, J Wu MSc, T Hama MSc, A-C Phiho-Gomes PhD, T R Lumbers PhD, Prof A Banerjee PhD, C Tomlinson PhD, A Torralbo PhD, Prof P Taylor PhD, Prof S Denaxas PhD, Prof H Hemingway PhD, Prof H Wu PhD); **Advanced Care Research Centre, Usher Institute, University of Edinburgh, Edinburgh, UK** (H Zhang PhD, Prof H Wu); **European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK** (T Groza PhD); **Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK** (S Khalid PhD); **Cicely Saunders Institute of Palliative Care, Policy & Rehabilitation, King's College London, London, UK** (M Hocaoglu PhD); **College of Life Sciences, University of Leicester, Leicester, UK** (Prof K Khunti PhD); **Health Data Research UK, London, UK** (Prof C Sudlow PhD, Prof S Denaxas, Prof H Hemingway); **UCL Institute of Ophthalmology, University College London, London, UK** (N Pontikos PhD); **Moorfields Eye Hospital NHS Foundation Trust, London, UK** (N Pontikos); **British Heart Foundation Data Science Centre, Health Data Research UK, London, UK** (R Priedon BA, Prof C Sudlow, Prof S Denaxas); **Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK**

(Prof C Sudlow); **National Institute of Health Research, University College London Hospitals Biomedical Research Centre, London, UK** (Prof S Denaxas, Prof H Hemingway); **School of Health and Wellbeing, University of Glasgow, Glasgow, UK** (Prof H Wu)

Correspondence to: Prof Honghan Wu, School of Health and Wellbeing, University of Glasgow, Glasgow G11 6EW, UK
honghan.wu@ucl.ac.uk

or

Johan Thygesen, Institute of Health Informatics, University College London, London NW1 2DA, UK
j.thygesen@ucl.ac.uk

Research in context

Evidence before this study

Our previous study of COVID-19 outcomes across rare diseases was underpowered, with a sample size of 158 patients with rare diseases diagnosed with COVID-19 and 125 unaffected relatives, from Genomics England. We searched PubMed for publications from database inception to April 21, 2023, using the search terms "COVID-19" or "SARS-CoV-2" and "rare disease" or "ORPHANET", without language restrictions. Although many studies have examined the severity of COVID-19 in patients with rare diseases, to date, most have focused on a single disease or a few rare diseases and not taken a comprehensive rare disease-wide approach. Moreover, no studies have examined the effect of vaccination on mortality in patients with rare diseases, and the sample size used to examine rare diseases in most studies was small. The largest study we identified, conducted in England, included 168 680 individuals and focused on autoimmune rheumatic disease.

Added value of this study

In this study, we use national-scale electronic health record data from England to report point prevalences, adjusted for age and sex, for 331 rare diseases with clinically validated ICD-10 or Systemized Nomenclature of Medicine – Clinical Terms code lists, or both. Among these, 186 (56%) diseases did not have existing point prevalence data available in Orphanet (an online resource for rare diseases). 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 carriers of even very rare conditions (<1 case per million people). Our analysis of COVID-19-related mortality has shown the clinical relevance of national data for rare diseases. Specifically, we identified eight rare conditions associated with a significantly increased risk of mortality from COVID-19, even among fully vaccinated individuals (ie, patients for whom at least 14 days have elapsed since their second dose).

Implications of all the available evidence

These findings provide robust and reproducible estimates of prevalence, sex, and ethnicity for diseases that might often have been underprioritised and for which such information was mostly not previously available. Our COVID-19 mortality findings highlight the need for targeted governmental policy and support to address the high level of clinical vulnerability of these patients to COVID-19.

for patients but also affects inclusion and integration in mainstream education systems and causes discrimination, social exclusion, and a greater risk of impoverishment.[7–9]

For more on **Orphanet** see http://www.orpha.net

Globally, there has been a substantial push to collect data on patients with rare diseases. The Orphanet database is an international initiative providing essential resources for various rare diseases, such as clinical mappings and prevalence. Orphanet documents more than 800 established registries, databases, and cohorts.[10] These initiatives often target specific rare diseases or groups of rare diseases, offering varied international coverage. A notable example is the French National Rare Disease Registry, which captures a large number of cases for over 5000 rare diseases from patients treated in rare disease expert centres, but this registry does not currently provide prevalence information.[11]

Research into rare diseases has long been hindered by numerous challenges, including scarce data, reliance on case and family reports for single diseases, geographical variations, absence of consensus around definitions, use of heterogeneous epidemiological approaches for case ascertainment, and incomplete codification in health-care systems.[2] Consequently, many rare diseases have unknown or imprecise prevalence estimates, preventing data-driven health-care prioritisation, which can be detrimental in the context of events such as the COVID-19 pandemic, and hampering systematic assessment and comparison across health-care systems.

Previous research has measured the indirect effect of COVID-19 on the life and care of patients with rare diseases, highlighting the exacerbated difficulties they face, such as health-care service disruption, difficulty of access due to physical distancing and safety concerns, declining mental health due to prolonged diagnosis, absence of effective treatment, and subsequently deteriorating symptoms.[12–14] Direct quantification of the risk of severe COVID-19 outcomes in rare disease populations remains scant. In 2022, Rutter and colleagues[15] showed an increase in COVID-19-related deaths among patients with rare autoimmune rheumatic diseases in England, using electronic health records (EHRs).[15] In a population-wide study conducted in Hong Kong in 2021, Chung and colleagues[16] described an increased risk of COVID-19-related death among hospitalised patients with rare diseases. We found similar results regarding the first COVID-19 wave in the UK in our previous study published in 2022, which used data from Genomics England.[17]

In many countries, including the UK, groups associated with increased risk of adverse outcomes from COVID-19 were prioritised for vaccination. The UK shielded patient list (those deemed clinically extremely vulnerable to the development of severe SARS-CoV-2 infections) was created in a data-driven way and refined with a risk prediction tool to consider additional risk groups.[18] Although a few rare respiratory (eg, bronchiectasis) and neurological (eg, myasthenia gravis) conditions were included in the model, the majority of rare diseases were

unaccounted for. This failure to include patients with most rare diseases underscores the need to examine adverse COVID-19 outcomes in patients with rare diseases within a large cohort and to improve codification for easier identification of rare diseases in EHRs. In this context, population-scale data are crucial to define and refine epidemiological estimates of rare diseases and the risk of severe COVID-19 outcomes. This approach is useful not only for risk management and vaccine prioritisation 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, we aimed to evaluate the prevalence of rare diseases and the demographic characteristics of the affected individuals and explore the impact of rare diseases during the COVID-19 pandemic, building on our previous findings that patients with rare diseases were disproportionately affected and underprioritised. We report reproducible estimates of population prevalence, sex and ethnicity ratios, and age of onset of those affected by rare diseases and assess differences in mortality from COVID-19 during a time period spanning multiple virus variants and vaccine roll-out. Such findings are important to explore the diversity of patients with rare diseases and to identify the demographic subpopulations that were specifically challenged by the pandemic.

Methods

Study design and population

We conducted a nationwide retrospective observational cohort study using nine linked datasets from National Health Service (NHS) England available within NHS England's Secure Data Environment (SDE) and accessed through the British Heart Foundation (BHF) Data Science Centre's CVD-COVID-UK/COVID-IMPACT Consortium. We used the following datasets: primary care data from the General Practice Extraction Service Extract for Pandemic Planning and Research (GDPPR);[19] COVID-19 testing data from the Public Health England Second Generation Surveillance System; COVID-19 hospital admission data from the Secondary Uses Service, from the Hospital Episode Statistics for admitted patient care (HES-APC), adult critical care, and outpatients (HES-OP), and from the COVID-19 Hospitalisations in England Surveillance System; COVID-19 vaccination status; and mortality information from the Office for National Statistics' Civil Registration of Deaths. The linked datasets include only structured data—ie, no clinical reports or imaging data were available. Wood and colleagues[20] previously provided a thorough description of these linked datasets, which were found to accurately replicate the age, biological sex at birth, and ethnicity distribution of the English population when compared with official UK Government statistics (further details are in appendix 1 [pp 2–3]).

The study start date was Jan 23, 2020, the date of the first recorded COVID-19 case in the UK,[21] and the end date was Nov 30, 2021, the latest date with overlapping data availability across the datasets when analysis commenced. We included individuals who were alive at the start of the study; registered with a general practitioner (GP) in England (minimum one patient record in GDPPR) between Jan 1, 1990 (the first year with good data coverage), and Jan 23, 2020; and associated with a valid person pseudo-identifier, enabling data linkage.

North East–Newcastle and North Tyneside 2 provided ethics approval for the CVD-COVID-UK/COVID-IMPACT research programme (Ethics Committee reference 20/NE/0161) to access—within secure trusted research environments—unconsented, whole-population, de-identified data from EHRs collected as part of patients' routine health care, as described previously.[20] The data used in this study, from NHS England's SDE, are not publicly available and were made available to accredited researchers only. The CVD-COVID-UK/COVID-IMPACT programme, led by the BHF Data Science Centre), requested approval to access data in NHS England's SDE service for England from the Independent Group Advising on the Release of Data via an application made in the Data Access Request Service online system (DARS-NIC-381078-Y9C5K). The CVD-COVID-UK/COVID-IMPACT Approvals & Oversight Board subsequently granted approval to researchers (JHT, HZ, HW, TH, and JW) for this project to access the data.

For the **NHS's Data Access Request Service** see https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services2

Identification of participants with rare diseases

We used Orphanet to identify and define rare diseases in individuals included in the nine linked datasets; data were downloaded on May 6, 2022. We adopted a stepwise approach to identify rare diseases that could be accurately mapped to the disease codes available in our data sources (full details are in appendix 1 [p 3–4]).

For more on the **NHS's Secure Data Environment** see https://digital.nhs.uk/services/secure-data-environment-service2

For the **British Heart Foundation Data Science Centre** see https://bhfdatasciencecentre.org/

First, we extracted all Orphanet diseases with mappings to ICD-10 or Systemized Nomenclature of Medicine–Clinical Terms (SNOMED CT), the clinical terminologies used in our data sources. Second, we kept diseases belonging to the following Orphanet disorder types: disease, morphological anomaly, malformation syndrome, or clinical syndrome (n=5864). Third, we excluded all diseases for which mapping to ICD10 was classified as narrow to broad, as these would encompass more general diseases that are typically not rare conditions. Fourth, we computed the point prevalence for all eligible diseases in the SDE (n=3817 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 done by two independent clinicians (A-CP-G and TH) to validate the accuracy of the matching of diseases to SNOMED CT and ICD-10 codes by considering both the code definitions and frequencies from the EHR data. Disparities were discussed by

See Online for appendix 1

A-CP-G and TH and resolved through mutual agreement. This process resulted in the inclusion of 331 rare diseases with highly specific mapping, of which 164 had ICD-10 mappings, 140 had SNOMED CT mappings, and 27 had both (appendix 1 p 6). Participants were defined as having a rare disease if they had one or more codes of these diseases.

See Online for appendix 2

Estimation of point prevalence of rare diseases in the English population

Point prevalences (estimates per 1 000 000 individuals) were calculated including anyone who was alive and had a rare disease diagnosis between Jan 1, 1990, and the study start. Prevalence estimates adjusted for age and sex were calculated with the 2021 census data from England as reference.[22]

We compared our sex-adjusted and age-adjusted estimated prevalences with the point prevalences of rare diseases provided by Orphanet, giving preference to the Orphanet estimate from the UK, followed by European estimates (regional or national specific) and then worldwide estimates. Ethnicity was categorised as Asian or Asian British, Black or Black British, mixed, White, other ethnicities, and unknown (further details are in appendix 1 [p 3]).

For rare diseases in our sample identified in five or fewer individuals, adjusted prevalence estimates could not be given due to the risk of re-identification; in these cases, we used a prevalence of 0·1 per million for comparison with Orphanet. 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 vaccination data

To evaluate the effect of COVID-19 on patients with rare diseases, we used five previously defined COVID-19 phenotypes:[23] a positive SARS-CoV-2 test, COVID-19 diagnosis recorded in primary care, hospital admission with a COVID-19 diagnosis, ventilatory support, or death (appendix 1 p 4). Onset of COVID-19 was defined as the date of the earliest COVID-19 event (ie, any of the five phenotypes). Date of the outcome was defined as the date of the earliest record of COVID-19 mortality. Time-to-event was calculated as the difference in days between the dates of onset and outcome. Vaccination status was identified from the COVID-19 vaccination dataset. Patients were classified as fully vaccinated once 14 days had elapsed after their second dose.

Statistical analysis

Fisher's exact test was done to test for differences in sex and ethnicity ratios for the 219 rare diseases affecting 100 or more individuals, comparing the ratio of patients with a specific rare disease with the ratio in all unaffected individuals (ie, without 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).

We compared COVID-19-related mortality in people with a particular rare disease or a condition in an Orphanet-defined rare disease category (whereby each rare disease might belong to multiple categories; appendix 2 p 1) with matched controls from the general population. Differences in COVID-19-related mortality were addressed with a time-to-event analysis (appendix 1 pp 5–6). In brief, the study period for assessing COVID-19-related mortality was set as Sept 1, 2020, to Nov 30, 2021. This period spanned the time after the roll-out of vaccination efforts during which COVID-19 testing capacity remained high. Cohorts were formed for each rare disease and rare disease category, with exact matching on age group, sex, ethnicity, and vaccination status at the ratio of two controls per individual with a rare disease. Survival functions were estimated with a Kaplan–Meier estimator. We estimated hazard ratio (HR) and 95% CI using a univariable Cox proportional hazards model, with time of event as the dependent variable and rare disease status as the independent variable. To examine the effect of governmental shielding recommendation on rare diseases, we compared rare diseases with high risk of COVID-19-related death and diseases in the NHS England shielding list.[24]

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Our study population comprised 58 162 316 individuals who were registered with a GP in England and alive at the onset of the pandemic on Jan 23, 2020.[21] In this cohort, 894 396 (1·5%) individuals had been diagnosed with at least one of 331 rare disorders before the onset of the pandemic, giving a prevalence of 15 377 per million. In total, 7965 (0·9%) of 894 396 patients with a rare disease died from COVID-19, compared with 141 287 (0·2%) of 58 162 316 individuals in the full study population cohort, from Jan 23, 2020, to Nov 30, 2021.

3 359 077 clinical codes (384 unique codes) were associated with the 331 conditions identified in the study cohort. The majority, 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 GPs; and 169 889 (5·1%; 206 unique codes) were ICD-10 codes from hospital outpatient care. When 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 coarctation (appendix 2 p 2).

	Prevalence			Sex		Age						Ethnicity				
	Individuals affected (n)	Adjusted population prevalence*	Prevalence in Orphanet	Female sex	Male sex	Median age at first diagnosis, years (IQR)	<18 years	18–29 years	30–49 years	50–69 years	≥70 years	White ethnicity	Asian or Asian British ethnicity	Black or Black British ethnicity	Mixed ethnicity	Other ethnicity
Leprechaunism	<10	0·13	<1	<10	<10	49 (21·–53)	<10	<10	<10	<10	<10	<10	<10	<10	<10	<10
Seckel syndrome	59	0·99	<1	25 (42·4%)	34 (57·6%)	7 (1–20·5)	10 (16·9%)	<10	36 (61·0%)	<10	0	36 (61)	18 (30·5%)	<10	<10	<10
Wells syndrome	577	9·99	<1	330 (57·2%)	247 (42·8%)	49 (34–62)	31 (5·4%)	29 (5·0%)	118 (20·5%)	230 (39·9%)	169 (29·3%)	488 (84·6%)	45 (7·8%)	23 (4·0%)	<10	11 (1·9%)
Poliomyelitis	1290	22·76	<1	676 (52·4%)	614 (47·6%)	65 (55–73)	29 (2·2%)	26 (2·0%)	86 (6·7%)	408 (31·6%)	741 (57·4%)	1018 (78·9%)	163 (12·6%)	62 (4·8%)	17 (1·3%)	24 (1·9%)
Lafora disease	<10	<0·1	1–9	<10	<10	27 (19–29)	<10	<10	<10	<10	<10	<10	<10	<10	<10	<10
Wolfram syndrome	125	2·12	1–9	71 (56·8%)	54 (43·2%)	19 (11–34)	23 (18·4%)	51 (40·8%)	25 (20%)	22 (17·6%)	<10	73 (58·4%)	46 (36·8%)	<10	<10	<10
Tibial hemimelia	329	5·69	1–9	141 (42·9%)	188 (57·1%)	11 (5–15)	143 (43·5%)	108 (32·8%)	69 (21%)	<10	<10	261 (79·3%)	31 (9·4%)	17 (5·2%)	12 (3·6%)	<10
Moyamoya disease	734	12·55	1–9	460 (62·7%)	274 (37·3%)	25 (12–41)	169 (23%)	167 (22·8%)	235 (32%)	127 (17·3%)	36 (4·9%)	440 (59·9%)	133 (18·1%)	116 (15·8%)	30 (4·1%)	13 (1·8%)
Fibular hemimelia	283	4·96	10–99	115 (40·6%)	168 (59·4%)	9 (2–14)	143 (50·5%)	94 (33·2%)	34 (12%)	11 (3·9%)	<10	237 (83·7%)	23 (8·1%)	<10	13 (4·6%)	<10
Angelman syndrome	1250	21·54	10–99	649 (51·9%)	601 (48·1%)	6 (2–17)	464 (37·1%)	386 (30·9%)	323 (25·8%)	59 (4·7%)	18 (1·4%)	1063 (85%)	85 (6·8%)	36 (2·9%)	35 (2·8%)	22 (1·8%)
Haemophilia B	2004	34·12	10–99	747 (37·3%)	1257 (62·7%)	31 (14–49)	347 (17·3%)	295 (14·7%)	653 (32·6%)	457 (22·8%)	252 (12·6%)	1683 (84·0%)	176 (8·8%)	58 (2·9%)	42 (2·1%)	35 (1·7%)
Interatrial communication	107 513	1934·18	10–99	54 328 (50·5%)	53 179 (49·5%)	4 (0–42)	56 615 (52·7%)	11 019 (10·2%)	12 753 (11·9%)	17 014 (15·8%)	10 112 (9·4%)	87 311 (81·2%)	10 459 (9·7%)	4303 (4·0%)	3189 (3·0%)	1891 (1·8%)
Alopecia totalis	809	13·95	100– 500	531 (65·6%)	278 (34·4%)	39 (19–58)	108 (13·3%)	122 (15·1%)	196 (24·2%)	242 (29·9%)	141 (17·4%)	628 (77·6%)	92 (11·4%)	52 (6·4%)	24 (3·0%)	<10
Huntington's disease	6088	104·41	100– 500	3252 (53·4%)	2836 (46·6%)	51 (40–62)	23 (0·4%)	210 (3·4%)	1656 (27·2%)	2892 (47·5%)	1307 (21·5%)	5721 (94·0%)	179 (2·9%)	54 (0·9%)	37 (0·6%)	37 (0·6%)
Haemophilia A	11 635	197·60	100–500	4587 (39·4%)	7047 (60·6%)	33 (18–51)	1619 (13·9%)	1797 (15·4%)	3746 (32·2%)	2742 (23·6%)	1731 (14·9%)	9890 (85·0%)	992 (8·5%)	337 (2·9%)	182 (1·6%)	191 (1·6%)
Interstitial cystitis	22 701	392·76	100–500	20 053 (88·3%)	2648 (11·7%)	48 (34–63)	43 (0·2%)	1816 (8·0%)	6389 (28·1%)	7722 (34·0%)	6731 (29·7%)	20 589 (90·7%)	1301 (5·7%)	349 (1·5%)	216 (1·0%)	208 (0·9%)
Good syndrome	0	0·18	NA	<10	<10	66 (55–75)	0	0	<10	<10	<10	<10	0	0	<10	0
Schilder disease	115	1·97	NA	69 (60·0%)	46 (40·0%)	51 (32–61)	<10	11 (9·6%)	23 (20·0%)	51 (44·3%)	24 (20·9%)	91 (79·1%)	14 (12·2%)	<10	<10	<10
Trigeminal neuralgia	43 396	760·57	NA	29 573 (68·1%)	13 822 (31·9)	61 (49–73)	70 (0·2%)	794 (1·8%)	6394 (14·7%)	16 183 (37·3%)	19 955 (46·0%)	40 170 (92·6%)	1842 (4·2%)	691 (1·6%)	334 (0·8%)	272 (0·6%)
Infantile apnoea	49 966	932·04	NA	23 242 (46·5%)	26 722 (53·5%)	0	44 956 (90·0%)	5006 (10·0%)	<10	<10	0	40 307 (80·7%)	4419 (8·8%)	1969 (3·9%)	2092 (4·2%)	905 (1·8%)
Polymyalgia rheumatica	158 648	2831·75	NA	110 097 (69·4%)	48 551 (30·6%)	76 (69–82)	24 (0)	232 (0·1%)	2211 (1·4%)	22 201 (14·0%)	133 980 (84·5%)	153 786 (96·9%)	2583 (1·6%)	839 (0·5%)	463 (0·3%)	615 (0·4%)

Data are n (%) unless otherwise specified. Data are ordered by reported Orphanet prevalence per million individuals (low to high), rare disease category (provided by Orphanet), and then estimated prevalence in study (low to high). Age group data were calculated at the study start; age and sex data were available for all diseases. Data (including 95% CIs) for all rare diseases studied are in appendix 2 (p 1). NA=not available. *Prevalence per million individuals adjusted for age and sex.

Table 1: **Demographics of 21 selected rare diseases highlighting different prevalence spectra**

	Study population (N=58 162 316)	All patients with a rare disease (N=894 396)	Patients with a rare disease and COVID-19 (N=128 608)	Patients with a rare disease and hospitalised with COVID-19 (N=17 750)	Patients with a rare disease and COVID-19 on ventilatory support (N=2924)	Deaths of patients with a rare disease and COVID-19 (N=7965)
Deaths from COVID-19	141 287 (0·2%)	7965 (0·9%)	7965 (6·2%)	4641 (26·1%)	1322 (45·2%)	7965 (100·0%)
All deaths	880 041 (1·5%)	46 273 (5·2%)	11 746 (9·1%)	6517 (36·7%)	1336 (45·7%)	7557 (94·9%)
Sex						
Female (%)	29 175 170 (50·2%)	492 200 (55·0%)	72 476 (56·4%)	10 377 (58·5%)	1464 (50·1%)	4425 (55·6%)
Male (%)	28 985 457 (49·8%)	402 171 (45·0%)	56 129 (43·6%)	7372 (41·5%)	1460 (49·9%)	3540 (44·4%)
Age, years						
<18	11 549 034 (19·9%)	225 397 (25·2%)	35 720 (27·8%)	762 (4·3%)	185 (6·3%)	17 (0·2%)
18–29	8 917 574 (15·3%)	104 538 (11·7%)	18 525 (14·4%)	590 (3·3%)	110 (3·8%)	31 (0·4%)
30–49	16 158 828 (27·8%)	138 246 (15·5%)	24 234 (18·8%)	1424 (8·0%)	397 (13·6%)	161 (2·0%)
50–69	13 894 037 (23·9%)	173 619 (19·4%)	22 172 (17·2%)	3324 (18·7%)	1072 (36·7%)	1057 (13·3%)
≥70	7 642 843 (13·1%)	252 596 (28·2%)	27 957 (21·7%)	11 650 (65·6%)	1160 (39·7%)	6699 (84·1%)
Ethnic group						
White	46 008 525 (79·1%)	777 730 (87·0%)	111 746 (86·9%)	15 786 (88·9%)	2362 (80·8%)	7310 (91·8%)
Asian or Asian British	5 570 658 (9·6)	59 717 (6·7%)	9584 (7·5%)	1081 (6·1%)	316 (10·8%)	359 (4·5%)
Black or Black British	2 242 918 (3·9)	25 935 (2·9%)	3172 (2·5%)	511 (2·9%)	146 (5·0%)	159 (2·0%)
Mixed	1 261 443 (2·2)	16 652 (1·9%)	2337 (1·8%)	166 (0·9%)	54 (1·8%)	47 (0·6%)
Other	1 202 022 (2·1)	10 541 (1·2%)	1387 (1·1%)	180 (1·0%)	42 (1·4%)	67 (0·8%)
Unknown	1 876 750 (3·2)	3821 (0·4%)	373 (0·3%)	26 (0·1%)	<10	24 (0·3%)
Social deprivation, fifths*						
1 (most deprived)	11 993 725 (20·6%)	188 047 (21·0%)	29 696 (23·1%)	4271 (24·1%)	731 (25·0%)	1574 (19·8%)
5 (least deprived)	11 103 838 (19·1%)	171 851 (19·2%)	23 327 (18·1%)	2967 (16·7%)	448 (15·3%)	1506 (18·9%)
Unknown	46 176 (0·1%)	662 (0·1%)	60 (<0·1%)	<10	<10	<10
Patients with comorbidities (on shielded patient list)	4 096 093 (7·0%)	210 571 (23·5%)	32 394 (25·2%)	9578 (54·0%)	1614 (55·2%)	3397 (42·6%)

*Information on social deprivation was provided by the datasets used in the study.

Table 2: **Study population demographics**

Numbers of identified patients for each condition and point prevalences adjusted for age and sex for all examined rare diseases are in table 1 and appendix 2 (p 1). As expected, the adjusted point prevalences were highly correlated with the raw point prevalence estimates (appendix 1 p 7). The most frequent rare disease 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 133 980 (84·5%) of 158 648 people with polymyalgia rheumatica being older than 70 years. With an adjusted point prevalence of 2831·8 per million (n=158 648), this condition exceeds the UK prevalence definition of 500 or fewer cases per million for rare diseases.[25]

Compared with the full study population, more individuals with rare diseases were female (55% *vs* 50·2%), younger than 18 years (25·2% *vs* 19·9%), or older than 70 years (28·2% *vs* 13·1%; table 2), and fewer were Asian or Asian British (6·7% *vs* 9·6%) or Black or Black British (2·9% *vs* 3·9%). As expected, compared with the full study population, individuals with rare diseases were more likely to be on the shielding list (23·5% *vs* 7·0%, as indicated by the SNOMED CT code; table 2).

After Bonferroni correction for multiple testing, we found significant differences in male to female sex ratios for 111 of 219 rare diseases affecting at least 100 individuals (appendix 2 p 2), compared with the unaffected study population. 49 rare diseases were more prominent in female patients than in male patients (replicating known findings for Rett syndrome[26] and interstitial cystitis[27]), and 62 were more prominent in male patients than in female patients (eg, Kennedy's disease and reactive arthritis). The ratios of individuals of Asian or Asian British ethnicity to those of White ethnicity were significantly different for 100 of the 219 rare diseases, compared with the unaffected study population, with 47 rare diseases significantly more frequent in individuals of Asian or Asian British ethnicity (eg, lamellar ichthyosis and maple urine disease). The ratios of individuals of Black or Black British ethnicity to those of White ethnicity were significantly different for 75 rare diseases, compared with the unaffected study population, with 22 rare diseases significantly more frequent in individuals of Black or Black British ethnicity (ratios were highest for moyamoya disease and Quinquaud folliculitis decalvans). 37 rare diseases were more

frequent in White individuals compared with individuals of either Black or Asian ethnicity (appendix 2 p 2).

Point prevalence data were not available from Orphanet for 186 (56%) of the 331 rare diseases in our study cohort. For 86 (59%) of the 145 diseases with available data, the adjusted point prevalences observed in our sample were within the prevalence ranges reported by Orphanet (figure 1). 25 (17%) diseases had higher estimates than Orphanet, with the largest discrepancy observed for interatrial communication—1848 per million individuals in our study versus 10–99 per million in Orphanet. 34 (23%) of our estimates were lower than those of Orphanet—eg, prevalence of chronic actinic dermatitis was 2 per million, with its Orphanet estimate being 100–500 per million.

To understand these differences, we explored the data sources underlying the Orphanet point prevalence estimates for the 145 rare diseases. The data sources were variable, and most did not present reproducible definitions; 63 (43%) listed expert or Orphanet as the only source and could not be verified. Of the 83 with a PubMed unique identifier, 23 (28%) were case reports, 20 (24%) were systematic reviews that referred to either case reports or epidemiological studies, 12 (8%) were reports by the European Medicines Agency, and 28 (34%) referred back to articles describing the conduct of an epidemiological study; for 13 (46%) of these 28, no sample size could be found. Three data sources had no reference data (table 3).

We compared the risk of COVID-19-related death in people with rare diseases with that in matched controls from the general population. Risk of COVID-19 related death was significantly increased for eight of 331 rare diseases in vaccinated individuals and for 28 of 331 rare diseases in individuals who were not fully vaccinated (figure 2; appendix 2 p 3). Of the 25 rare disease categories (appendix 1 p 12), we observed significantly increased COVID-19-related mortality among vaccinated individuals in 11 categories (encompassing 323 diseases, of which 179 were included in the analysis) and among individuals who were not fully vaccinated in 14 categories (encompassing 325 diseases, of which 181 were included in the analysis; appendix 2 p 4).

For vaccinated individuals, the three rare diseases with the highest risk of COVID-19-related mortality were bullous pemphigoid (HR 8·07, 95% CI 3·01–21·62), an autoimmune bullous skin disease; osteogenesis imperfecta (10·14, 2·22–46·26), a rare, genetic, primary bone dysplasia; and autosomal dominant polycystic kidney disease (4·06, 2·03–8·13), a genetic, renal tubular disease (figures 2, 3; appendix 2 p 3). For individuals who were not fully vaccinated, the three rare diseases with the highest risk of COVID-19-related mortality 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 3; appendix 2 p 3). Of the 28 rare diseases with significantly increased risk of COVID-19-related death in individuals who were not fully vaccinated, only interatrial communication was included in the list of shield diseases.

For vaccinated individuals, the three rare disease categories with the highest risk of COVID-19-related mortality were infertility disorders (HR 7·54, 95% CI

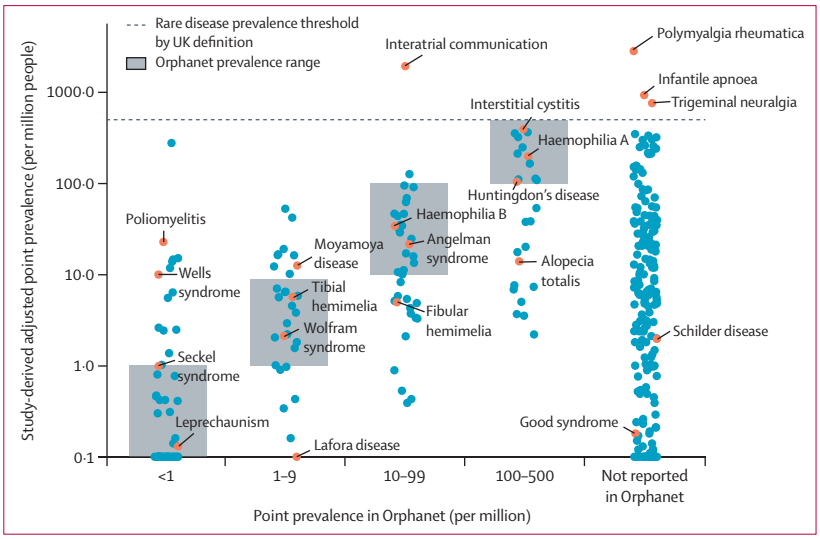

Figure 1: **Comparison of study-derived adjusted point prevalence estimates for rare diseases with Orphanet prevalence ranges**

The 331 rare diseases in this study are stratified by relative rarity (increasing in rarity along the x-axis), based on Orphanet estimates. The points indicate our estimated point prevalences, adjusted for age and sex, from population data in England. The labelled diseases are the 21 selected rare diseases highlighted in table 1 as representing a range of prevalences. The grey boxes indicate prevalence ranges reported for the diseases in Orphanet data. The dashed line represents a prevalence of 1 per 2000 individuals, the upper limit definition for a rare disease suggested by the UK Government.[25]

	<1 (N=52)	1–9 (N=27)	10–99 (N=40)	100–500 (N=26)	All diseases (n=145)
Location					
UK	3 (6%)	5 (19%)	10 (25%)	5 (19%)	23 (16%)
EU	6 (12%)	17 (63%)	28 (70%)	17 (65%)	68 (47%)
Worldwide	43 (83%)	5 (19%)	2 (5%)	4 (15%)	54 (37%)
Source type*					
Review	7 (13%)	3 (11%)	7 (18%)	3 (12%)	20 (13%)
Article (observational study or survey)	6 (11%)	9 (33%)	5 (13%)	5 (19%)	25 (17%)
Case series report	16 (30%)	2 (7%)	3 (8%)	2 (8%)	23 (15%)
Guideline	0	1 (4%)	0	1 (4%)	2 (1%)
European Medicines Agency report	0	2 (7%)	3 (8%)	7 (27%)	12 (8%)
Expert only	24 (45%)	10 (37%)	22 (55%)	7 (27%)	63 (42%)
Randomised controlled trial	0	0	0	1 (4%)	1 (<1%)
Unclear†	0	3 (11%)	0	0	3 (2%)

Data are n (%). Data are categorised by Orphanet prevalence per million. *Source types are not mutually exclusive—some rare diseases have multiple sources listed. †No reference data.

Table 3: **Overview of Orphanet prevalence data sources for the 145 of 331 rare diseases with available data**

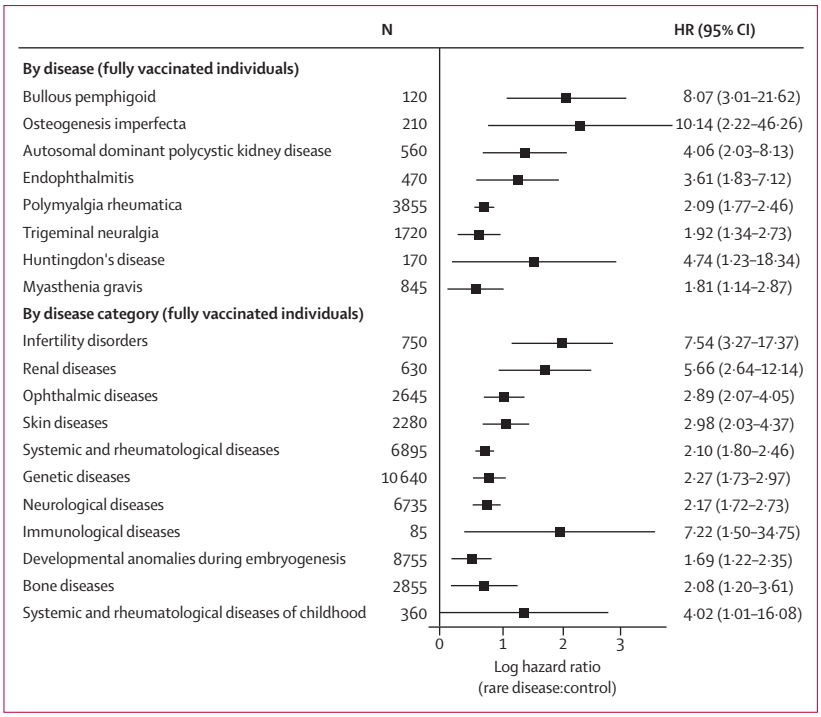

Figure 2: **Risk of COVID-19-related mortality in people with rare diseases**
Forest plot showing log hazard ratios of COVID-19-related death for rare diseases and rare disease categories with a significant increase of risk in fully vaccinated individuals, based on the lower bounds of the 95% CIs of the HRs for individuals with a rare disease and matched controls. HR=hazard ratio.

3·27–17·37), renal diseases (5·66, 2·64–12·14), and ophthalmic diseases (2·89, 2·07–4·05; figures 2, 3; appendix 2 p 4). For individuals who were not fully vaccinated, the three rare disease categories with the highest risk of COVID-19-related mortality 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 3, appendix 2 p 4).

Sample size influences the power to detect significant HRs for COVID-19 mortality in individuals with rare diseases. All significant HR ratios for specific rare diseases were found in disorders that had an adjusted prevalence of more than 17 per million people (for both vaccinated and not fully vaccinated individuals). Moreover, significant HR ratios for vaccinated individuals were seen only in diseases with a later age of onset (ie, median age of first diagnosis >34 years), as fewer COVID-19 deaths were observed in younger individuals. However, when considering significant HR estimates based on individuals affected by any rare disease belonging to a particular category, the identified categories with significant HRs for COVID-19 mortality in fully vaccinated individuals covered 323 of the 331 rare diseases we examined (appendix 1 p 10).

Lastly, projection of our estimated prevalence into smaller (national) population sizes of 25 million, 10 million, and 5 million individuals show the effectiveness of using large-scale linked EHR resources to identify sufficient patients with rare diseases, enabling these patients to be meaningfully described in terms of key demographic strata (appendix 1 p 10).

Discussion

In this study, we used linked EHR data for more than 58 million people in England to provide detailed information for 331 rare diseases and explore the effect of rare diseases on COVID-19 mortality. We present fully reproducible point prevalence estimates adjusted for age and sex for all 331 diseases, with counts of the specific clinical codes observed in our population, including prevalence estimates for 184 rare diseases that did not previously have such data available from Orphanet. We show that for eight rare diseases, mortality was significantly increased in fully vaccinated individuals compared with matched individuals from the general population. Increased mortality was also seen for 11 of 25 categories of rare disease. This study shows the increased power in rare disease epidemiology when linked national data are made available for research.

The 331 rare diseases described in this study are just a fraction of the total number of rare diseases. Orphanet's rare disease database lists 10 563 clinical entities, but many of these (28%) are currently unmapped to clinical codes (ie, ICD-10 and SNOMED CT), complicating case ascertainment. A large proportion (52·4%) of Orphanet-listed rare diseases were excluded from our study, either because none of the related codes were observed in our cohort (20·1%; eg, the rare inflammatory eye disease ocular cicatricial pemphigoid was not found in our data) or because the frequency of the mapped clinical codes was excessively high due to a too-broad definition (32·3%; eg, a range of rare types of Parkinson's disease, such as X-linked parkinsonism-spasticity syndrome, map directly to the broader ICD-10 code for Parkinson's disease). We believe the difficulty of using routinely coded data for rare disease analysis is a finding in its own right, which further emphasises the need for improved coding and diagnostic awareness for a very large proportion of rare diseases, a sentiment that is already well described in the literature.[6,7] Future studies should explore what could be done to improve this shortcoming. This result highlights that even with population-wide data, better mappings to more specific codes would substantially improve case identification. The more fundamental issue is that classification systems such as ICD-10 do not possess the granularity needed for all rare diseases—or the granularity is not recorded in the clinic. For example, ICD-10 coding is often only recorded to three digits, whereas the fourth-digit level of ICD-10-Clinical Modification would provide more precision. Orphanet currently only includes mappings to three-digit ICD-10 codes.

Comprehensive clinical terminologies such as SNOMED CT have contributed to the release of mappings covering thousands of rare diseases,[28] and some

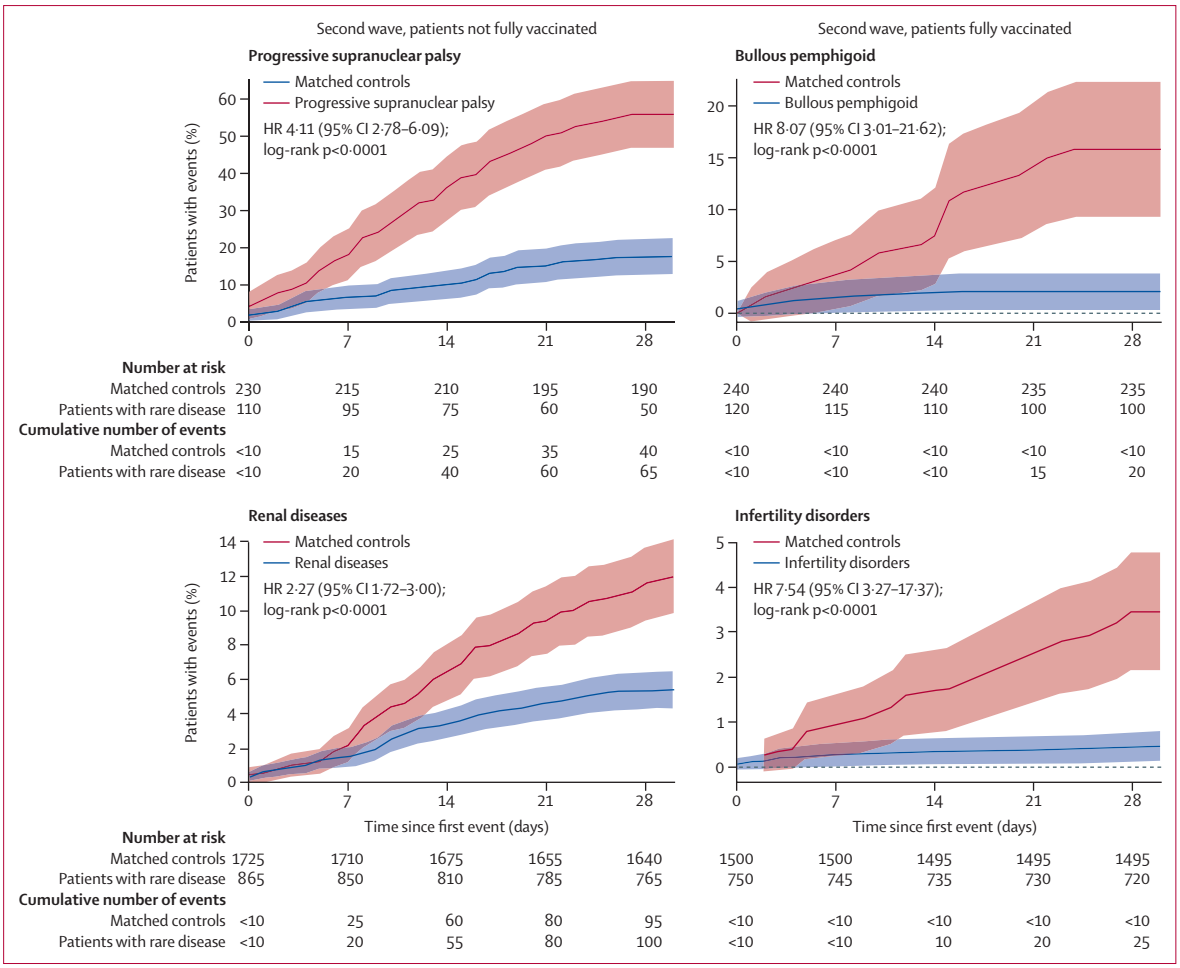

Figure 3: **Survival analysis of rare diseases and categories with the highest risk of COVID-19-related mortality**
Kaplan–Meier plots for the rare diseases or disease categories with the highest mortality risks, based on the lower bounds of the 95% CIs of the HRs for individuals with a rare disease and matched controls; stratified by patients' COVID-19 vaccination status. Events are defined as COVID-19-related death. Kaplan–Meier plots for all conditions are in appendix 1 (pp 19–45). HR=hazard ratio.

improvements in the most recent ICD-11 might also better meet the need for increased specificity. Nonetheless, the use of very specific codes in contemporary health systems remains inconsistent, and systematic efforts are needed to promote it. For better case identification, previous work has shown the great potential of free text in EHRs,[29] which, however, comes with its own set of challenges.[30] As another promising approach to identifying undercoded or undiagnosed patients with rare diseases, phenotypes associated with a disease might constitute a computational disease model. Both approaches, however, would require more comprehensive and granular EHR data, which are not yet available on a population scale. Using terminologies or ontologies that are larger than SNOMED CT or ICD-10, such as the Unified Medical Language System,[31] would probably yield better granularity and improve mappings to other ontologies. However, the Unified Medical Language System has the same challenge of ensuring high-quality mapping of concepts to ICD-10 or SNOMED CT as an integrated part of routine care.

Orphanet and Rare Diseases Europe have made great progress in assessing the prevalence of rare diseases.[10,14] However, most of the rare disease prevalences estimated in our study (56%) were not previously available in Orphadata, Orphanet's data platform, which is one of the most comprehensive repositories of such estimates. We recognise that other rare disease registers exist,[10] some with data on the examined rare diseases, but comparing our data with all these registers was beyond the scope of this study. Our review of the Orphadata sources for the available prevalence estimates showed that the sources varied in terms of methods, with 42% denoted as expert view, making a direct study comparison challenging. Among diseases with prevalence available in Orphadata, 60% were consistent with our data, whereas the prevalence of some differed greatly, with both over-estimates and underestimates.

For more on **Orphadata** see http://www.orphadata.org

Differences between prevalence estimates provided by this study and by Orphadata could have many causes. Traditionally, prevalence of rare diseases was estimated indirectly from reported incidence[32] or from meta-analyses of multiple sources of prevalence. Such estimates might be low if there is a poor coverage of incidences. Estimating denominators can also be difficult, leading to both underestimates and overestimates. In this study, interatrial communication showed the greatest discrepancy in terms of prevalence. Our analysis identified 1848 cases per million individuals in the English population, significantly higher than the Orphanet prevalence range of 10–99 per million. Although the Orphanet prevalence is not based on an external reference, the main reason for this discrepancy probably stems from the diagnostic mapping method. Interatrial communication is linked to ICD-10 code Q21.1 (atrial septal defect) in Orphanet, which also includes more prevalent conditions. Our observed prevalence aligns with findings from a large German study on congenital heart defects in newborns[33] that focused on atrial septal defect. However, both prevalences probably overestimate the true prevalence of rare interatrial communication. 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 derived was restricted to individuals aged 75 years and older.[34] Our results highlight the importance of data-driven, reproducible estimation of rare disease point prevalence, adjusted for age and sex, to inform rare disease care and research and enable reasonable comparison between studies.

Detailed analysis on sex and ethnicity from our study highlights a range of novel insights with regard to differences in disease occurrence between groups and, in some cases, disease disparities indicating inequalities of disease burden or underdiagnosis in subpopulations. To address these issues, joint efforts from governments, research communities, and charities might be required to offer better screening mechanisms for rare disease, targeted public health policies, and effective and accessible diagnostic approaches.

A strength of our study was the size of the cohort, covering almost the entire population of England. Large population cohorts are important when estimating prevalence of rare diseases because they offer statistical power and a nationally representative sample. Previous studies in South Korea[35] and Japan[36] have used insurance claim data to enable such estimation at population scale. Our study adds to the evidence with observations from the English population, with the added benefit of using linked EHR data. The dataset used in this study provided granular information from both primary and secondary care, enabled further epidemiological analysis of rare diseases to explore the impact of COVID-19-related deaths, and facilitated analysis of the demographics and prevalence of hundreds of rare diseases.

An important limitation to highlight for this study and other studies of rare diseases relying on EHR diagnostic data is that most patients with a rare disease have to wait for several years for a diagnosis, and some remain undiagnosed.[6] This limitation means that overall prevalences relying on diagnostic information are likely to be underestimated. A second limitation is that during this study, we did not have access to clinical notes, documents, laboratory reports, or results from genetic analysis that could be used to clinically verify the rare disease diagnosis of the identified patients. Therefore, it was not possible to estimate the sensitivity of our approach for each of the diseases. Third, our COVID-19 mortality analysis only included demographic covariates for cohort matching; other potential confounding factors, such as BMI and comorbidities, were considered but not included. Our research question concerned risk of COVID-19-related death in people with rare diseases compared with matched cohorts of the general population—controlling for factors such as BMI and comorbidities could result in underestimation of the effect of rare diseases on COVID-19-related deaths because these factors can mediate the effect of rare diseases. Fourth, despite the comprehensive nature of the GDPPR, our cohort included more individuals than the 2021 Office for National Statistics census population for England (56 489 800),[37] reflecting the absence of registration data for residents in England; without these data, temporary residents might persist in the dataset, elevating the population denominator and leading to an underestimation of prevalence. Fifth, the GDPPR contains only a subset of SNOMED CT codes—ie, those which were thought to be related to COVID-19 research[19]—and might, therefore, also cause underestimation of rare disease prevalence. Sixth, our study period spans what is commonly referred to as the second wave and the first half of the third wave of the pandemic in the UK, mainly driven by the alpha and delta SARS-CoV-2 variants. As variant information was not available for individual positive tests, we were not able to estimate variant-specific hazard ratios of COVID-19-related deaths. Finally, when interpreting the results of the COVID-19 mortality analysis, it is important to consider that each rare disease might belong to multiple categories. This factor might, in part, explain the high mortality seen for categories that we would not directly relate to COVID-19, such as infertility and ophthalmic diseases, as these might be driven by rare diseases affecting multiple organs.

Federated systems that allow for analysis across multiple unlinked datasets, such as 4CE,[38] with detailed COVID-19 phenotypes, and the Miracum consortium[39] for rare diseases, are promising approaches in terms of future work, whereby the needed data for rare diseases

might become available with sufficient sample size. Future work should also focus on identifying undiagnosed patients with rare diseases and patients who might be diagnosed but are unrecorded in a structured format in the EHR. An improved understanding of the shared medical history and symptoms of these conditions, from both structured data and leveraging of clinical documents, could aid in the swifter identification of undiagnosed individuals who share similar medical histories. We hope that the detailed data in the appendices of this Article, alongside the shared analysis code, will help to foster such future work.

In conclusion, national-scale EHRs offer an unprecedented resource to study the epidemiology of rare diseases because of the increased statistical power and granularity of health events they provide. In this study, we used large-scale linked EHR data to systematically estimate the prevalence of rare diseases and risk of COVID-19-related death. Such basic epidemiological descriptors are crucial for improved planning and prioritisation of rare disease diagnosis and treatment on a national level.

Contributors

JHT, HZ, and HW were responsible for conceptualisation, data curation, methodology, investigation, formal analysis, and data visualisation. JHT, HZ, HW, HI, JW, and TH wrote the first draft. JW, TH, HI, and A-CP-G assisted with data curation, methodology, and analysis. TRL, CT, AT, PT, SD, and HH assisted with conceptualisation and methodology. All authors were responsible for writing, reviewing, and editing the manuscript. As Director of the British Heart Foundation Data Science Centre, CS coordinated approvals for and access to data within NHS England's Secure Data Environment for the CVD-COVID-UK/COVID-IMPACT Consortium. All listed authors meet authorship criteria, and no others meeting the criteria have been omitted. All authors have seen and approved the final manuscript. JHT, HZ, and HW are the guarantors and had full access to the raw data. As this work was done in a secure environment, not all authors had access to the raw data. All authors have access to the exported results. All authors had access to the derived study data. Drafts of project proposals and manuscripts were shared with all consortium members for feedback.

Declaration of interests

SD and CS are supported by a core grant award from the UK Medical Research Council (MRC) and partner funders to Health Data Research (HDR) UK and a British Heart Foundation (BHF) grant award to HDR UK for the BHF Data Science Centre. CT and AT report research funding paid to University College London from GSK for research outside the scope of this submitted work. KK was supported by grants and contracts from AstraZeneca, Boehringer Ingelheim, Lilly, Merck, Novo Nordisk, Roche, Sanofi, Servier, Oramed Pharmaceuticals, Roche, Daiichi-Sankyo, and Applied Therapeutics. NP was funded by a UK National Institute of Health and Care Research (NIHR) Artificial Intelligence in Health and Care Award (AI_AWARD02488) on inherited retinal diseases. RP receives support from the BHF Data Science Centre, HDR UK. TRL and SK receive support from HDR UK. All other authors declare no competing interests.

Data sharing

The analytical code and protocol are available under an open source licence at https://github.com/BHFDSC/CCU019_01. For inquiries about data access, please see www.healthdatagateway.org/dataset/7e5f0247-f033–4f98-aed3–3d7422b9dc6d or email bhfdsc@hdruk.ac.uk. Results will be disseminated through the BHF Data Science Centre's CVD-COVID-UK/COVID-IMPACT webpage, BHF communication channels, the BHF Data Science Centre's public contributor panel, and NHS England's communications channels.

Acknowledgments

This work was supported by the BHF Data Science Centre led by HDR UK (BHF grant SP/19/3/34678). The BHF Data Science Centre funded co-development (with NHS England) of the Secure Data Environment service for England, 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 Government 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 also supported by the UK Research and Innovation (UKRI)-funded Longitudinal Health and Wellbeing COVID-19 National Core Study (grants MC_PC_20030 and MC_PC_20059) and by the NIHR (grant NIHR202639), the NIHR/HDR UK Winter Pressure Award (WP0006), and awards from the MRC (MR/S004149/2 and MR/X030075/1). The views expressed are those of the authors and not necessarily those of the NIHR or the UK Department of Health and Social Care. This work benefits from the infrastructure and partnerships assembled by HDR UK, including through the Data and Connectivity National Core Study, funded by UKRI (grant MC_PC_20058) and the British Council (UCL-NMU-SEU International Collaboration On Artificial Intelligence In Medicine: tackling challenges of low generalisability and health inequality, and facilitating better urology care with effective and fair use of artificial intelligence—a partnership between UCL and Shanghai Jiao Tong University School of Medicine). HZ and HW were partly funded by the Legal & General Group (via a research grant to establish the independent Advanced Care Research Centre at University of Edinburgh). The views expressed are those of the authors and not necessarily those of Legal & General. Additional financial support was provided by the HDR UK discretionary fund Rare Disease Phenomics (TF2022·42), which receives its funding from HDR UK (HDRUK 2022·0137), and by the NIHR University College London Hospitals Biomedical Research Centre, which also receives funding from HDR UK (HDR-9006). HDR UK receives funding from the UK MRC, 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 of the Welsh Government, Northern Ireland's Public Health Agency, BHF, and Wellcome Trust. This study made use of de-identified data held in NHS England's Secure Data Environment service for England and made available via the BHF Data Science Centre's CVD-COVID-UK/COVID-IMPACT Consortium. This work used 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 health relevant data available for research. The study team would also like to thank the BHF Data Science Centre's public contributors for their input and NHS England's Data Access Environment output checkers Lisa Gray and Hanna McLean.

References

1 Haendel M, Vasilevsky N, Unni D, et al. How many rare diseases are there? *Nat Rev Drug Discov* 2020; **19:** 77–78.
2 Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* 2020; **28:** 165–73.
3 Ferreira CR. The burden of rare diseases. *Am J Med Genet A* 2019; **179:** 885–92.
4 Marshall DA, Benchimol EI, MacKenzie A, et al. Direct health-care costs for children diagnosed with genetic diseases are significantly higher than for children with other chronic diseases. *Genet Med* 2019; **21:** 1049–57.
5 Davies SC. Annual report of the Chief Medical Officer 2016: generation genome. UK Department of Health, 2017.
6 Bauskis A, Strange C, Molster C, Fisher C. The diagnostic odyssey: insights from parents of children living with an undiagnosed condition. *Orphanet J Rare Dis* 2022; **17:** 233.
7 Stoller JK. The challenge of rare diseases. *Chest* 2018; **153:** 1309–14.
8 Delaye J, Cacciatore P, Kole A. Valuing the "burden" and impact of rare diseases: a scoping review. *Front Pharmacol* 2022; **13:** 914338.

9 Tisdale A, Cutillo CM, Nathan R, et al. The IDeaS initiative: pilot study to assess the impact of rare diseases on patients and healthcare systems. *Orphanet J Rare Dis* 2021; **16:** 429.

10 Orphanet. Rare disease registers in Europe. Dec, 2021. https://www.orpha.net/pdfs/orphacom/cahiers/docs/GB/Registries.pdf (accessed May 6, 2022).

11 Pichon T, Messiaen C, Soussand L, et al. Overview of patients' cohorts in the French National Rare Disease Registry. *Orphanet J Rare Dis* 2023; **18:** 176.

12 Peach E, Rutter M, Lanyon P, et al. Risk of death among people with rare autoimmune diseases compared with the general population in England during the 2020 COVID-19 pandemic. *Rheumatology* 2021; **60:** 1902–09.

13 Lampe C, Dionisi-Vici C, Bellettato CM, et al. The impact of COVID-19 on rare metabolic patients and healthcare providers: results from two MetabERN surveys. *Orphanet J Rare Dis* 2020; **15:** 341.

14 Rare Diseases Europe. How has COVID-19 impacted people with rare diseases? October, 2020. https://www.eurordis.org/publications/how-has-covid-19-impacted-people-with-rare-diseases/ (accessed May 6, 2022).

15 Rutter M, Lanyon PC, Grainge MJ, et al. COVID-19 infection, admission and death among people with rare autoimmune rheumatic disease in England: results from the RECORDER project. *Rheumatology* 2022; **61:** 3161–71.

16 Chung CCY, Wong WHS, Chung BHY. Hospital mortality in patients with rare diseases during pandemics: lessons learnt from the COVID-19 and SARS pandemics. *Orphanet J Rare Dis* 2021; **16:** 361.

17 Zhang H, Thygesen JH, Shi T, et al. Increased COVID-19 mortality rate in rare disease patients: a retrospective cohort study in participants of the Genomics England 100,000 Genomes project. *Orphanet J Rare Dis* 2022; **17:** 166.

18 Reynolds M. How we created the Shielded Patient List. Nov 17, 2020. https://digital.nhs.uk/blog/tech-talk/2020/how-we-created-the-shielded-patient-list (accessed Oct 25, 2022).

19 NHS Digital. General Practice Extraction Service (GPES) data for pandemic planning and research. June 8, 2022. https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data (accessed Sept 14, 2022).

20 Wood A, Denholm R, Hollings S, et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021; **373:** n826.

21 Lillie PJ, Samson A, Li A, et al. Novel coronavirus disease (Covid-19): the first two patients in the UK with person to person transmission. *J Infect* 2020; **80:** 578–606.

22 Office for National Statistics. Population and household estimates, England and Wales: census 2021. June 28, 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationandhouseholdestimatesenglandandwalescensus2021 (accessed Feb 2, 2023).

23 Thygesen JH, Tomlinson C, Hollings S, et al. COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records. *Lancet Digit Health* 2022; **4:** e542–57.

24 The National Archives, NHS Digital. Shielded patient list methodology—additions and subtractions. NHS Digital (archive). https://webarchive.nationalarchives.gov.uk/ukgwa/20220610000338/https://digital.nhs.uk/coronavirus/shielded-patient-list/methodology/additions-and-subtractions?key= (accessed Feb 2, 2023).

25 Department of Health & Social Care. The UK rare diseases framework. Jan 9, 2021. https://www.gov.uk/government/publications/uk-rare-diseases-framework/the-uk-rare-diseases-framework (accessed Sept 21, 2022).

26 Petriti U, Dudman DC, Scosyrev E, Lopez-Leon S. Global prevalence of Rett syndrome: systematic review and meta-analysis. *Sys Rev* 2023; **12:** 5

27 Clemens JQ, Meenan RT, Rosetti MC, Gao SY, Calhoun EA. Prevalence and incidence of interstitial cystitis in a managed care population. *J Urol* 2005; **173:** 98–102.

28 SNOMED International. SNOMED-CD to Orphanet map package production release notes—July 2021. Jan 7, 2021. https://confluence.ihtsdotools.org/display/RMT/SNOMED+CT+to+Orphanet+Map+package+Production+Release+Notes+-+July+2021 (accessed Feb 20, 2024).

29 Dong H, Suárez-Paniagua V, Zhang H, et al. Ontology-driven and weakly supervised rare disease identification from clinical notes. *BMC Med Inform Decis Mak* 2023; **23:** 86.

30 Groza T, Köhler S, Moldenhauer D, et al. The human phenotype ontology: semantic unification of common and rare disease. *Am J Hum Genet* 2015; **97:** 111–24.

31 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids Res* 2004; **32** (suppl 1): D267–70.

32 Kruger E, McNiven P, Marsden D. Estimating the prevalence of rare diseases: long-chain fatty acid oxidation disorders as an illustrative example. *Adv Ther* 2022; **39:** 3361–77.

33 Lindinger A, Schwedler G, Hense H-W. Prevalence of congenital heart defects in newborns in Germany: results of the first registration year of the PAN Study (July 2006 to June 2007). *Klin Padiatr* 2010; **222:** 321–26.

34 Dawe RS. Chronic actinic dermatitis in the elderly: recognition and treatment. *Drugs Aging* 2005; **22:** 201–07.

35 Lim S-S, Lee W, Kim Y-K, et al. The cumulative incidence and trends of rare diseases in South Korea: a nationwide study of the administrative data from the National Health Insurance Service database from 2011–2015. *Orphanet J Rare Dis* 2019; **14:** 49.

36 Ninomiya K, Okura M. Nationwide comprehensive epidemiological study of rare diseases in Japan using a health insurance claims database. *Orphanet J Rare Dis* 2022; **17:** 140.

37 Office for National Statistics. Population estimates. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates (accessed Feb 14, 2023).

38 Sperotto F, Gutiérrez-Sacristán A, Makwana S, et al. Clinical phenotypes and outcomes in children with multisystem inflammatory syndrome across SARS-CoV-2 variant eras: a multinational study from the 4CE consortium. *EclinicalMedicine* 2023; **64:** 102212.

39 Zöller D, Haverkamp C, Makoudjou A, et al. Alpha-1-antitrypsin-deficiency is associated with lower cardiovascular risk: an approach based on federated learning. *Respir Res* 2024; **25:** 38.