BHF Data Science Centre: CVD-COVID-UK / COVID-IMPACT Project Proposal Form

Project reference:	CCU019
Project title:	Identification and personalised risk prediction for severe COVID-19 in patients with rare disorders impacting cardiovascular health
Proposal version:	4.0
Start date (best estimate):	01/06/21
End date (best estimate)[1]:	31/08/23
Named project lead and institution/organisation:	Honghan Wu, University College London

Plain English summary

Approx. 200 words[2] overall to succinctly summarise the project in language suitable for a non-specialist lay public.[3] This summary should be written using the headings below for clarity, with approximately one paragraph for each heading.

Describe the **challenge** or problem your project will address:

We know individuals with underlying health conditions have greater risk of developing severe COVID-19 and ending up with poorer outcomes. That is why governments and public health services have been providing dedicated and prioritised protections for these more clinically vulnerable people – for example, via recommending shielding or being prioritised to have COVID-19 vaccinations.

However, the majority of those living with rare diseases – around 5.8% of the UK population, or 3.7 million people - are often overlooked. Rare diseases are often poorly recorded in clinical data leading to a challenge in identifying patients whose rare condition makes them clinically vulnerable. We don't know the most effective way to personalise and manage treatments for patients with rare diseases who contracted COVID-19. Furthermore, there are many people who are not diagnosed but share similar clinical presentations (so-called phenotypes) to those diagnosed rare-disease patients. We know some of them are likely to share similar vulnerabilities. However, we don't know how to identify them currently.

How will your project be the **solution** to address/understand the challenge or problem?

In this project, we aim to tackle these challenges by bringing together a comprehensive set of knowledge about rare diseases, and applying the most up to date data science technologies to use such knowledge and resources on CVD-COVID-UK datasets.

What is the potential **impact** from this work, e.g. how will it benefit patients/NHS, inform policy etc?

In this way, we hope to develop a more accurate identification system for people living with rare diseases who are clinically vulnerable. We will also provide the much needed information on the risk of severe COVID-19 in people with rare diseases, hopefully leading to an improvement in their care

[1] Current approvals extend for about three years from June 2020, although we envisage that we should be able to extend these for some projects looking at longer term outcomes, if needed.
[2] Word counts are only a guide and can be exceeded, if necessary.
[3] Your plain English summary will appear on the BHF Data Science Centre CVD-COVID-UK / COVID-IMPACT webpages. Please use the sort of language you might use to describe your project to a non-specialist friend, relative or journalist. Please avoid using technical jargon and aim to keep sentences short for ease and clarity of reading.

by providing evidence on treatments that may work better for them. Furthermore, we will analyse the compound risks of severe COVID-19 in people bearing clinical risks and disadvantaged socioeconomic backgrounds, aiming to inform policy responses for providing better management and treatment for these most vulnerable groups who might previously have been overlooked.

We are matched with a data analyst from the department of health and social care. This will enable us a speedy dissemination of our work to the policy makers realising swift and actionable suggestions. We will also disseminate our findings to charities and societies of rare diseases in the UK and beyond to maximise the impact of our work.

Background

Approx. 300 words[2] summarising why the question(s) you are addressing matter, and how your project fits within the broad scope of CVD-COVID-UK / COVID-IMPACT

Several common comorbidities have been identified as risk factors for severe COVID-19. However, risk borne by patients with rare diseases is largely unknown, in part because rare diseases are often poorly recorded in routine data. Although individually rare, ‘rare diseases’ are cumulatively common, affecting approximately 1 in 17 people in UK. Of the 6002 rare diseases recorded in GARD (Genetic and Rare Diseases Information Center), 168 diseases directly affect the cardiovascular system and many more indirectly affect the cardiovascular system due to complications or medication use. Many people with rare diseases missed out on shielding support in the first wave of the pandemic. Effort is now urgently needed to estimate risk for COVID-19-related poor outcome for these patients, so that all vulnerable patients can be prioritised for vaccination.

Currently there are no studies estimating the risk of severe COVID-19 in rare disease patients. Our preliminary analysis using data from Genomics England (GEL) revealed that patients with rare diseases have about three times higher COVID-19-related mortality than their healthy relatives (univariable OR=3.26). However, the low number of cases (in the GEL data) limits our ability to draw solid conclusions on any specific rare disease or to perform multivariable analysis. Therefore, we are seeking to use the datasets available through the CVD-COVID-UK consortium in England and Scotland (with extension to Wales if sufficient analyst resource becomes available) to provide the power to further develop personalised risk prediction tools for patients with rare cardiovascular and associated (co-morbid) rare diseases.

Research question(s)

Aims:

Focusing on patients with cardiovascular disease, we aim to meet two main challenges: 1) Identification of co-morbid rare disease from routinely collected data, given the structural coding of rare diseases is incomplete; 2) Estimation of the added risk for developing severe COVID-19 in this patient subgroup. Given the low prevalence, this study will benefit hugely from the large sample size available via the CVD-COVID-UK consortium.

Objectives:

Objective 1 - Improving the identification of people with rare diseases in routine datasets.

We will initially use the recruitment criteria of Genomics England to define computable disease models for rare diseases, containing rules regarding patient phenotypes. These disease models will then be applied to the population-wide data available via CVD-COVID-UK to identify potential rare disease patients, using data from both primary and secondary care. We will aim to identify a representative array of different types of rare diseases, to the degree that is possible with the available data. We will expand the 'definition of rareness' based on rare disease knowledge bases and the literature to capture more rare conditions having potentially high adverse impacts on cardiovascular health. This work will also contribute to the consortium's work on developing shareable phenotyping algorithms by generating computable phenotypes on rare diseases.

Objective 2 - Risk estimation and prediction for severe COVID-19 in patients with rare disease comorbidity.

We will perform a retrospective cohort study to compare the rate of target events (e.g. critical care dependency or mortality) in rare disease patients identified in *Objective 1* to the rate of the general population.

We will examine the risk of COVID-19-related poor outcome in people with a range of rare diseases (based on the preliminary findings in the Genomics England data analysis). We will shortlist rare diseases that plausibly make patients more vulnerable to COVID-19 related poor outcomes. Finally, we will derive and validate a machine learning model, which could take rare disease information and other conventional risk factors as input, to produce a prediction model for severe COVID-19.

Objective 3 – using rare disease phenotype models as a proxy for identifying COVID-19 related vulnerabilities

We address that rather than aiming for identifying uncoded rare disease patients this objective focuses on identifying people without a rare disease code **but sharing similar vulnerabilities**. To do that, we will first implement and evaluate Human Phenotype Ontology (HPO) phenotype models for rare diseases in CVD-COVID-UK/COVID-IMPACT instance of NHS Digital's TRE for England. Then, using phenotype models as feature sets, machine learning methods will be used to identify uncoded people with high risk of COVID-19 related poor prognosis. In addition, we will analyse the consequences of compound risk factors of rare disease phenotypes and socioeconomics background.

Patient/public contributor involvement

The BHF DSC works with patients and the public to ensure transparency, and to build trust in the use of health data for research. Please complete the relevant section below, to indicate your plans for involving patient/public contributors throughout your project.

Please contact bhfdsc@hdruk.ac.uk if you would welcome an initial conversation with the BHF DSC team and patient/public contributors, or any other support regarding patient/public contributor involvement in your project.

The research team **has** consulted with public/patients on plans for this project.
Please provide brief details (e.g., any specific groups you are engaging with, how this has influenced the project/research question, and how you will continue involving public/patients throughout the project)

We will involve patients/lay members from study design through to implementation.

The proposal has been discussed with a lay panel facilitated by the BHF Data Science Centre. We communicated in detail with the patient/lay members on the study design and the patient and public benefit of the project. Further similar meetings will be held in the follow-up stages of the project. Via the PPIE, we want to ensure that our research is kept in tune with day-to-day experience of patients and that we provide actionable and practical advice to patients as well as healthcare providers.

Methods

*Provide a **brief overview** of methods to be used - a detailed plan is not required.*
Please also complete the table on the next page for information on TRE(s), datasets and years of data required and the analyst(s) who you propose will work with the data in the TRE(s)

Definition of rare diseases using rare disease ontology and knowledge base
The list of rare diseases will be obtained using The Orphanet Rare Disease Ontology and its linkage with other ontologies combined with rare disease information databases like Orphadata and GARD (Genetic and Rare Diseases Information Center). In the early stage, we will focus on rare diseases with increased risk of COVID-19 associated mortality from our preliminary study in Genomics England cohort. In recognition of the wide spectrum rare diseases we will aim to identify representative types of rare diseases with regards to their manifestation and origin (i.e., monogenic vs rare diseases with multigenic origin).

Identification of rare disease patients using computable disease models
The Genomics England project has comprehensive documentation of criteria used for patient recruitment. Using ontologies and rules, we will convert these criteria to ontology concepts and rules so that we can search for people who meet the criteria in a scalable way. Manual evaluation of shortlisted patients will be done to confirm the accuracy of the method.

Risk estimation and prediction
Using the linked datasets, we will perform a retrospective cohort study estimating the risk of severe COVID19 in rare disease patients. In addition, we will develop a personalized prediction model which takes account of pre-existing rare diseases.

Implement rare disease phenotype model

We will map Human Phenotype Ontology (HPO) terms to ICD-10 and SNOMED-CT terms, which will enable the use of linked health datasets to populate phenotype representations for each disease. Such disease models will be validated and updated using real-world data from coded patients' records using multivariate analysis on HPO terms. We will particularly prioritise those diseases associated with high risk of poor prognosis, which will be called prioritised diseases in the rest of the proposal.

Identify uncoded people with high risk of COVID-19 related poor prognosis

The validated phenotype data models from the above will be used as the core feature sets in this task, to use a range of machine learning approaches to identify high risk individuals. For each condition in the prioritised disease list, we will devise a binary classification task using coded data as ground truth. A machine learning model trained for such a task would identify false positive results, i.e., people deemed as patients by the model but not coded in the system.

Analyse the compound (disease phenotype, sex, ethnicity and social-economics) risk factors of poor COVID-19 prognosis

To understand the compound risk factors of poor prognosis, we will apply multivariate analysis on variables including sex, ethnicity, social-economics categories, phenotypes, and other demographics using a matched case-control study on the infected sub-cohorts.

Trusted Research Environments (TRE)

England: NHS Digital TRE for England
Scotland: Scottish National Data Safe Haven *(for more information, view the COVID-19 Research Database Dataset and Variable Specification)*
Wales: Secure Anonymised Information Linkage Databank (SAIL)
Northern Ireland: Northern Ireland Honest Broker Service

***** FOR COVID-IMPACT[4] PROJECTS, PLEASE COMPLETE THE ANALYST AND DATA SOURCE DETAILS FOR ENGLAND ONLY *****

DATA ANALYSTS

	PLEASE COMPLETE THIS COLUMN
TRE	**Analyst(s) requiring TRE access – please provide name, institution, and email if not already a consortium member**
England	Honghan Wu, UCL; Johan, Thygesen, UCL, j.thygesen@ucl.ac.uk; Huayu Zhang, UoE, huayu.zhang@ed.ac.uk
Scotland	Honghan Wu, UCL; Johan, Thygesen, UCL, j.thygesen@ucl.ac.uk; Huayu Zhang, UoE, huayu.zhang@ed.ac.uk
Wales	Honghan Wu, UCL; Johan, Thygesen, UCL, j.thygesen@ucl.ac.uk; Huayu Zhang, UoE, huayu.zhang@ed.ac.uk
Northern Ireland	

DATA SOURCES

						PLEASE COMPLETE THESE COLUMNS		
TRE	**Category**	**Dataset Name**	**Year data available from**	**Time lag**	**Available in TRE**	**Required (X)**	**Years of data required (ALL or range)**	**Brief justification of why you need each dataset / date range**
England	Primary care	GDPPR: GPES Data for Pandemic Planning and Research	From the start of each individual's records[5]		Yes	X	ALL	We will use this data form consistency checks and phenotype extraction of COVID patients.
England	Secondary care	HES: Hospital Episode Statistics **- Admitted Patient Care**	1997		Yes	X	ALL	HES datasets contain symptom and disease codes as input of disease models.
England	Secondary care	**- Adult Critical Care**	2013		Yes	X	ALL	

[4] COVID-related research projects not directly linked to cardiovascular disease

[5] Includes patients with active, current registrations at participating practices and deceased patients with a date of death on or after 1 November 2019. Note: prescriptions and numeric values (e.g. BP, laboratory test results) only go back two years.

						PLEASE COMPLETE THESE COLUMNS		
TRE	**Category**	**Dataset Name**	**Year data available from**	**Time lag**	**Available in TRE**	**Required (X)**	**Years of data required (ALL or range)**	**Brief justification of why you need each dataset / date range**
England	Secondary care	- **Outpatients**	2019		Yes	X	ALL	
England	Secondary care	- **Accident & Emergency**	2007		Yes	X	ALL	
England	Secondary care	**SUS: Secondary Uses Service**	2019 / earlier		Yes			
England	Secondary care	**SUS/Uncurated Low Latency Hospital Data (Admitted Patient Care, Outpatients, Critical Care)**			Yes			
England	Secondary care	**Emergency Care Data Set (ECDS)**			Yes			
England	COVID testing	**COVID-19 SGSS: Second Generation Surveillance System**[6]	From start of records (2020)		Yes	X	ALL	Key data for COVID research
England	COVID testing	**Pillar 2 Antigen**	April 2020		Yes			
England	COVID testing	**Pillar 3 Antibody**	September 2020		Yes			
England	COVID testing	**Variant strain data (COG-UK)**			Expected TBC			
England	COVID vaccinations	**Vaccination Status**	December 2020		Yes			
England	COVID vaccinations	**Vaccination Adverse Reactions**	December 2020		Yes			
England	Deaths	**Civil Registration – Deaths** (ONS guidance / NHSD mortality data review)	1993		Yes	X	ALL	We use the ONS data for determining COVID-related death and filtering patents who are alive.
England	ITU	**ICNARC: Intensive Care National Audit and Research Centre**			Yes	X	ALL	Data useful for defining disease severity
England	ITU/HDU admissions	**COVID-19 SARI-Watch (formerly CHESS: COVID-19 Hospitalisation in England Surveillance System)**	From start of records (2020)		Yes	X	ALL	Data useful for defining disease severity

[6] Pillar 1 and 2 positive tests

						PLEASE COMPLETE THESE COLUMNS		
TRE	**Category**	**Dataset Name**	**Year data available from**	**Time lag**	**Available in TRE**	**Required (X)**	**Years of data required (ALL or range)**	**Brief justification of why you need each dataset / date range**
England	Prescribing/ dispensing	Medicines Dispensed in Primary Care (NHS BSA)	April 2015		Yes	X	ALL	Prescription data provides the indicators whether a rare disease patient is vulnerable because of the medication used.
England	Prescribing/ dispensing	**Secondary Care Prescribed Medicines (EPMA)**			Yes			
England	NICOR CVD audits	**NICOR – MINAP: Myocardial Ischaemia National Audit Project**			Yes			
England	NICOR CVD audits	**NICOR – PCI: Percutaneous Coronary Interventions**			Yes			
England	NICOR CVD audits	**NICOR – NHFA: National Heart Failure Audit**			Yes			
England	NICOR CVD audits	**NICOR – NACSA: National Adult Cardiac Surgery Audit**			Yes			
England	NICOR CVD audits	**NICOR – NACRM: National Audit of Cardiac Rhythm Management**			Yes			
England	NICOR CVD audits	**NICOR – NCHDA: National Congenital Heart Disease Audit**			Yes			
England	NICOR CVD audits	**NICOR – TAVI: Transcatheter Aortic Valve Implantation**			Yes			
England	Stroke audit	**SSNAP: Sentinel Stroke National Audit Programme**			Yes			
England	National Vascular Registry	**National Vascular Registry Audit**			Expected TBC			
England	Other	Diagnostic Imaging Dataset			Expected TBC			
England	Other	Improving Access to Psychological Therapies (IAPT) v2.0 & v2.1	Sep 2020		Yes			
England	Other	Maternity Services Dataset (MSDS)	April 2019		Yes			
England	Other	Mental Health Services Dataset (MHSDS)	April 2019		Yes			

						PLEASE COMPLETE THESE COLUMNS		
TRE	**Category**	**Dataset Name**	**Year data available from**	**Time lag**	**Available in TRE**	**Required (X)**	**Years of data required (ALL or range)**	**Brief justification of why you need each dataset / date range**
England	Other	Patient Reported Outcome Measures (PROMs)			Expected TBC			
Scotland	Primary care	Primary care[7]			Yes	X	ALL	Primary care data contains record of rare disease patient care
Scotland	Secondary care	Outpatient Appointments and Attendances - Scottish Morbidity Record (SMR00)	1997		Yes	X	ALL	SMR datasets contain symptom and disease codes as input of disease models.
Scotland	Secondary care	General Acute Inpatient and Day Case - Scottish Morbidity Record (SMR01)	1997		Yes	X	ALL	SMR datasets contain symptom and disease codes as input of disease models.
Scotland	Secondary care	Accident & Emergency	2007		Yes	X	ALL	SMR datasets contain symptom and disease codes as input of disease models.
Scotland	COVID testing	COVID-19 laboratory and lighthouse testing (ECOSS)[8]	From start of records (2020)		Yes	X	ALL	Key data for COVID19 research
Scotland	COVID testing	**Covid Tests**[9]			Yes	X	ALL	Key data for COVID19 research
Scotland	COVID testing	Variant strain data (COG-UK)			Yes			
Scotland	COVID vaccinations	**Vaccination data**			Yes			
Scotland	Deaths	Deaths			Yes	X	ALL	Key data for COVID19 research
Scotland	ITU	Intensive care data - Daily (SICSAG)[10]			Yes	X	ALL	Data useful for defining disease severity
Scotland	ITU	Intensive care data - Episodes (SICSAG)[11]			Yes	X	ALL	Data useful for defining disease severity

[7] Data provided comprises a single cut of the data as at June 2020 with no current updates. Based on data used in the EAVEII project.

[8] Contains the first positive test result per person or earliest test result if they have never tested positive (dataset not updated after August 2021 – replaced by Covid Tests)

[9] Contains all test results (positive and negative) and replaced the ECOSS dataset from August 2021.

[10] Additional approval process required for this dataset.

[11] Additional approval process required for this dataset.

						PLEASE COMPLETE THESE COLUMNS		
TRE	**Category**	**Dataset Name**	**Year data available from**	**Time lag**	**Available in TRE**	**Required (X)**	**Years of data required (ALL or range)**	**Brief justification of why you need each dataset / date range**
Scotland	Prescribing/ dispensing	**Dispensed/Prescribed/Paid (Prescribing Information System)**	2015		Yes	X	ALL	Prescription data provides the indicators whether a rare disease patient is vulnerable because of the medication used.
Scotland	Stroke audit	**Scottish Stroke Care Audit**	TBC		Yes			
Scotland	Other	**Diabetes covariates**			Yes			
Scotland	Other	**Scottish Renal Registry**[12]	2019		Yes			
Wales	Primary care	**GPCD: Welsh Longitudinal General Practice (Daily COVID codes only)**	2020		Yes	X	ALL	Primary care data contains record of rare disease patient care
Wales	Primary care	**WLGP: Welsh Longitudinal General Practice**	2000		Yes	X	ALL	Primary care data contains record of rare disease patient care
Wales	Secondary care	**CCDS: Critical Care Dataset**	2007		Yes	X	ALL	Data useful for defining disease severity
Wales	Secondary care	**EDDD: Emergency Department Dataset Daily**	2010		Yes	X	ALL	Data useful for defining disease severity
Wales	Secondary care	**EDDS: Emergency Department Dataset**	2009		Yes	X	ALL	Data useful for defining disease severity
Wales	Secondary care	**OPDW: Outpatient Dataset for Wales**	2004		Yes	X	ALL	OPDW dataset contains symptom and disease codes as input of disease models.
Wales	Secondary care	**OPRD: Outpatient Referral Dataset**	2009		Yes	X	ALL	OPRD dataset contains symptom and disease codes as input of disease models.
Wales	Secondary care	**PEDW: Patient Episode Dataset for Wales**	1995		Yes	X	ALL	PEDW dataset contains symptom and disease codes as input of disease models.
Wales	COVID testing	**PATD: COVID-19 Test Results (Laboratory Information Management System [Pillar 1&2 NHS/Lighthouse Labs Results & Pillar 3 Antibody Results])**	March 2020		Yes	X	ALL	Key data for COVID19 research
Wales	COVID testing	**CTTP: COVID-19 Test, Trace and Protect**			Yes	X	ALL	Key data for COVID19 research

[12] Contains data to identify patients receiving hospital based renal replacement therapy – haemodialysis – only (from January 2019).

						PLEASE COMPLETE THESE COLUMNS		
TRE	**Category**	**Dataset Name**	**Year data available from**	**Time lag**	**Available in TRE**	**Required (X)**	**Years of data required (ALL or range)**	**Brief justification of why you need each dataset / date range**
Wales	COVID testing	**CVSP: COVID-19 Shielded People List**	May 2020		Yes	X	ALL	Key data for COVID19 research
Wales	COVID testing	**CVSD: COVID-19 Sequence Data**[13]			Yes			
Wales	COVID testing	**ONS COVID-19 Infection Survey**[14]			Expected TBC			
Wales	COVID vaccinations	**CVVD: Covid Vaccination Dataset**			Yes			
Wales	Deaths	**ADDD: Annual District Death Daily (ONS Deaths)**	2016		Yes	X	ALL	We use the ONS data for determining COVID-related death and filtering patents who are alive.
Wales	Deaths	**ADDE: Annual District Death Extract (ONS Deaths)**	1996		Yes	X	ALL	We use the ONS data for determining COVID-related death and filtering patents who are alive.
Wales	Deaths	**CDDS: COVID-19 Consolidated Deaths**	2019		Yes	X	ALL	We use the ONS data for determining COVID-related death and filtering patents who are alive.
Wales	ITU	**ICCD: ICNARC – Intensive Care National Audit & Research Centre *(COVID-19 only admissions)***	March 2020		Yes	X	ALL	Data useful for defining disease severity
Wales	ITU	**ICNC: ICNARC – Intensive Care National Audit & Research Centre *(All admissions)***			Yes	X	ALL	Data useful for defining disease severity
Wales	Prescribing/ Dispensing	**WDDS: Wales Dispensing Dataset**	2015		Yes	X	ALL	Prescription data provides the indicators whether a rare disease patient is vulnerable because of the medication used.
Wales	NICOR CVD audits	**NICO: NICOR Audits and Registers**			Expected TBC			

[13] Additional approval process required for this dataset

[14] Additional approval process required for this dataset (4-6 week lead time). Analysts requiring access must be ONS Safe Researcher Training certified and have a valid Accredited Researcher (AR) number.

						PLEASE COMPLETE THESE COLUMNS		
TRE	**Category**	**Dataset Name**	**Year data available from**	**Time lag**	**Available in TRE**	**Required (X)**	**Years of data required (ALL or range)**	**Brief justification of why you need each dataset / date range**
Wales	Stroke audit	**HQIP: HQIP Stroke Audit**			Expected TBC			
Wales	National Vascular Registry	**NVR: National Vascular Registry**			Expected TBC			
Wales	Other	**ADBE: Annual District Birth Extract**	1996		Yes			
Wales	Other	**MIDS: Maternity Indicators Dataset**	2014		Yes			
Wales	Other	**NCCH: National Community Child Health**			Yes			
Wales	Other	**CARE: Care Homes Index**	2018		Yes			
Wales	Other	**CARS: (CARIS – Congenital Anomaly Register and Information Service)**	1998	1-2 months	Yes			
Wales	Other	**CENW: Office of National Statistics Census (2011)**[15]	March 2011 only		Yes			
Wales	Other	**RTTD: Referral to Treatment Times**	2012		Yes			
Wales	Other	**SDEC: SAIL Dementia e-Cohort**	March 2019		Yes			
Wales	Other	**WASD: Welsh Ambulance Services NHS Trust**	2013		Yes			
Wales	Other	**WDSD: Welsh Demographic Service Dataset**	1990		Yes			
Wales	Other	**WRRS: Welsh Results Reporting Service**			Yes			
Northern Ireland		**TBC**			Expected TBC			

[15] Additional approval process required for this dataset (4-6 week lead time). Analysts requiring access must be ONS Safe Researcher Training certified and have a valid Accredited Researcher (AR) number.