Genomic, Proteomic and Phenotypic Biomarkers of COVID-19 Severity:

Publication Status:
Published online: 09/11/2023

DOI:
10.2196/50733

Document Version
Author Accepted version

Genomic, Proteomic and Phenotypic Biomarkers of COVID-19 Severity: Protocol for a Retrospective Observational Study

Running title: COVRES – Biomarkers of COVID-19 severity

Andrew English[1,2], Darren McDaid[1], Seodhna M. Lynch[1], Joe McLaughlin[1], Eamonn Cooper[1], Benjamin Wingfield[1], Martin Kelly[5], Manav Bhavsar[5], Victoria McGilligan[1], Rachelle E Irwin[3], Magda Bucholc[1], Shu-Dong Zhang[1], Priyank Shukla[1], Taranjit Singh Rai[1,] Anthony J Bjourson[1], Elaine Murray[1], David S Gibson[1]* and Colum P. Walsh[3,4]*

1. Personalised Medicine Centre, School of Medicine, Ulster University, C-TRIC Building, Altnagelvin Hospital, Glenshane Road, Londonderry/Derry, BT47 6SB **2.** School of Health and Life Sciences, Middlesbrough Tower (M4.14), Teesside University, TS1 3BA **3**. Genomic Medicine Research Group, School of Biomedical Science, Ulster University, Cromore Road, Coleraine BT52 1SA **4** Department for Cell- and Neurobiology, Biomedical and Clinical Sciences Division, Faculty of Medicine, Linköping University SE-581 83 Sweden **5.** Western Health Social Care Trust, Glenshane Rd, Londonderry BT47 6SB.

*Corresponding Authors Emails:

David Gibson d.gibson@ulster.ac.uk

Colum Walsh colum.walsh@liu.se;

Abstract word count: **420**

Manuscript word count: 3329 **(excluding references and tables)**

Tables - **2**; Figures – **2**

Abstract

Background - Health organisations and countries around the world found it difficult to control the spread of the coronavirus disease 2019. To minimise future impact on the UK National Health Service and improve patient care, there is a pressing need to identify individuals who are at higher risk of being hospitalised as a result of severe COVID-19. Early targeted work was successful in identifying angiotensin-converting enzyme-2 receptors and type-II transmembrane serine protease dependency as drivers of severe infection. Although a targeted approach highlights key pathways, a multi-omics approach will provide a much clearer and more comprehensive picture of severe COVID-19 aetiology and progression.

Objective - The Covid-19 Response Study (COVRES, NCT05548829) aims to carry out an integrated multi-omic analysis to identify biomarkers in blood and saliva that could contribute to host susceptibility to SARS-CoV-2 and development of severe COVID-19.

Methods - The COVID-19 Response (COVRES) study aims to recruit n=1000 people who recovered from SARS-CoV-2 infection in both community and hospital settings on the island of Ireland. The protocol below describes the retrospective observational study component carried out in Northern Ireland (NI; *cohort a*); the Republic of Ireland cohort will be described separately. For all NI participants (*n*=519), SARS-CoV-2 infection has been confirmed by RT-qPCR. A prospective *cohort b* of n=40 patients are also being followed up at 1, 3, 6 and 12 months post-infection to assess longitudinal symptom frequency and immune response. Data will be sourced from whole blood, saliva samples, and clinical data from the electronic care records, general health questionnaire, and using a GHQ-12 mental health survey. Saliva and blood samples were processed to extract DNA and RNA prior to whole genomic sequencing, RNA sequencing, DNA methylation, microbiome, 16S rRNA gene sequencing, and proteomic analysis performed on plasma. Multi-omic data will be combined with clinical data to produce sensitive and specific prognostic models of severity risk.

Results - An initial demographic and clinical profile of the NI *cohort a* has been completed: *n*=249 hospitalised and *n*=270 non-hospitalised patients were recruited, 64% were female, the mean age was 45

years. High levels of comorbidity were evident in the hospitalised cohort, with cardiovascular disease and metabolic and respiratory disorders being the most significant (P<0.001), grouped according to International Classification of Diseases 10 codes.

Conclusion – This study will provide a comprehensive opportunity to study multi-omic mechanisms of COVID-19 severity in re-contactable participants.

Trial Registration - The trial has been registered as an observational study on clinicaltrials.gov as NCT05548829; http://clinicaltrials.gov/ct2/show/NCT05548829. An outline of the trial protocol is included as a SPIRIT checklist.

Key words

COVID-19, clinical research, multi-omics, comorbidity, severity, Electronic health record.

Introduction

COVID-19 has a wide spectrum of clinical severity, with ~60% of cases thought to be asymptomatic or mildly symptomatic and ~5% being critically ill [1]. Severe infection is characterised by respiratory and multiorgan failure [2]. There are several known demographic risk factors such as age, male sex, diabetes mellitus and obesity [3], and recently high-risk genes and genetic variation have gained a lot of attention [4-8]. Identifying further biomarkers that reflect the pathophysiology of the disease and aid clinical staff in recognising severity is critical [9]. This would also help in the development of clinical management systems that can improve patient outcomes [10]. Early work focused on easily accessible laboratory indices such as elevated C-reactive protein and D-dimer among others which have been helpful in early management of high-risk patients [9, 11]. These biomarkers are commonly recorded in the Electronic Care Records (ECR), a technological development that allows the exchange of health information electronically facilitating effective diagnosis, reducing medical errors, and providing safer care and research. [12]. The limitations of routine laboratory biomarkers are well documented however [13].

Early work also implicated angiotensin-converting enzyme 2 (ACE2) receptors, type II transmembrane serine protease (TMPRSS2) in viral entry [14, 15]. A recent genome wide association study (GWAS) of 2000 critically ill patients [5] identified dipeptidyl peptidase 9 (DPP9), antiviral restriction enzyme activators OAS1, OAS2, OAS3 and tyrosine kinase 2 (TYK2). To date single-omic approaches have been used to identify genomic markers of COVID-19 severity [5, 16, 17] [18]. Here we seek to utilise multi-omic analysis using two tissue types (blood and saliva), in combination with comprehensive electronic care records and self-reported data to build one of the most extensive pictures yet.

Study aims and overview

The Covid Response Study (COVRES, NCT05548829) aims to carry out an integrated multi-omic analysis of factors contributing to host susceptibility to SARS-CoV-2 among a patient cohort of 1000 people from the geographically isolated island of Ireland. Due to differences in site, governance, and timelines the protocol below describes the study to be carried out in Northern Ireland (NI-COVRES) by Ulster University and the Western Health and Social Care Trust only; the Republic of Ireland component (Trinity College Dublin/St. James Hospital Dublin) will be described separately.

Figure 1 shows an overview of the main stages and timeline with data for each participant (n=519) on : i) Disease status ii) Genome iii) Transcriptome iv) Proteome v) Methylome iv) Microbiome iiv) Immune response iiiv) Patient history ix) Mental health x) Electronic care record and prospectively on *n*=40 at 1, 3, 6 and 12 months post positive PCR to assess persistent inflammatory and immune responses.

Methods

Status and timeline of the study

The main retrospective *cohort a* recruitment commenced in December 2020 and was completed in March 2021 except for the prospective *cohort b* (ongoing), integration of ECR record data was completed in January 2022, at time of writing and omics samples are being processed (Figure 2).

Ethical approval

Standard operating procedures (SOPs) and participant response questionnaires included SOPs for saliva sample kit preparation, blood collection and processing, downstream sample processing, website

management, data protection, and participant contact. The COVRES study was subsequently approved by the Health and Care Research Wales Ethics service on the 14th of July 2020 (REC ref 20/WA/0179).

Social Media outreach

Social media content (Twitter, Facebook) and webpage visuals were designed, with input from recovered patients, by the project Principal Investigators including a range of infographics and short explanatory texts. Information was circulated to local and national news outlets (TV, radio, newspaper) across Northern Ireland for recruitment purposes. Interested patients contacted the research team and were sent a patient information sheet. Appointments were then organised at least 24 hours later to gain informed consent and samples.

Participant recruitment with inclusion and exclusion criteria

Inclusion criteria: patients had to be >18 years of age but could have any body mass index (BMI) or ethnic origin. Exclusion criteria: patients were excluded if <18 years of age and if any intellectual disabilities were present. Hospitalisation status was determined if a patient attended/was admitted to hospital within 14 days of positive PCR result. Patients were also classified based on the World Health Organisation (WHO) scale [19] which reflects severity over the duration of the patient's infection regardless of hospitalisation status. For example, a patient may have an overall WHO score of 5 and be classified as non-hospitalised as they attended hospital >14 days from their positive PCR result. After receiving a Participant Information Sheet (PIS), patients interested in participating gave written informed consent and were enrolled. A self-report questionnaire established demographic information, lifestyle choices, family history of clinical disorders and COVID-19 severity and symptoms. This was followed by a general health questionnaire (GHQ-12) to help ascertain the patient's mental health after COVID-19 infection (Figure 2). This data was securely digitalised onto a bespoke database CovresNIdb generated on the REDCap platform [19] to comply with the terms of the ethical approval, human tissue act, and general data protection regulations (GDPR). This process is being repeated for *cohort b* (prospective) n= 40 with stricter timelines followed (1, 3, 6 and 12 month).

Biological sample processing

The Western Health and Social Care Trust (WHSCT) recruitment team coordinated sample collection appointments at hospital wards, Clinical Translational Research and Innovation Centre (C-TRIC) clinic rooms or home visits. Participants and related study code numbers were predetermined dependent on hospitalisation and logged in encrypted clinical data sheets on a secure server to ensure full data traceability. All whole blood and saliva processing carried out includes recruitment numbers, samples collection types, sample processing and downstream analysis: *n* numbers refer to patient numbers for specific omics analyses.

Isolation was carried out in a Category III containment hood with full PPE. Samples were not deactivated upon receipt or prior to processing. Participants provided 3x 10 ml of whole blood and 2x saliva samples of approximately 2 ml each (Figure 2). Blood was extracted using 21G Vacuette® safety needles (Greiner Bio-One Ltd, Gloucestershire) into 3x10 ml EDTA coated Vacuette® tubes and centrifuged at 4000 rpm (4 ºC) for 15 minutes. The buffy coat was extracted, washed, and stored for RNA sequencing (Figure 2). All samples were frozen at –80 ºC; time to freezer was <2 hrs and none showed signs of haemolysis. Saliva was collected using 1xDNA Genotek (DNA Genotek, Ottawa) Oragene DNA (OG-500) and 1x RNA (CP-190) collection tube per participant), samples were considered deactivated once lysed. Peripheral blood mononucleocyte cells (PBMCs) were isolated using the ficoll gradient separation methods as per [20].

Immune assays

Whole blood was analysed at 1 and 3 months post positive PCR test. Using the FACSAria III high speed cell sorter (Becton Dickinson, Oxford, UK, software version 9) with an 85 µm nozzle fitted, whole blood and PBMC samples were stained for T, B and NK cell populations using CD45 PerCP-Cy5.5, CD3 FITC, CD8 APC-Cy7, CD4 PE-Cy7, CD19 APC and CD16/CD56 PE (BD) before erythrocyte lysis by PharmLyse (BD) according to manufacturer's instructions. T cell subpopulations were measured using two defined panels- Panel 1: CD3 FITC, CD4 PE-Cy7, CD8 BV605, CD30 APC, CD45RA V450, CD45RO BV786, CD183 BB700; Panel 2: CD3 FITC, CD4 PE-Cy7, CD8 BV605, CD69 APC, CD45

V450, CD127 BV786, CD152 BB700, CD25 R718 and FoxP3 PE. Cell-surface staining was performed prior to fixing, permeabilizing and FoxP3 labelling using the Transcription Factor Buffer Set (BD Pharmingen).

DNA isolation

Saliva samples (WGS, methylome, microbiome) were incubated for 2 h at 56 °C, followed by DNA isolation using PrepIT L2P (DNA Genotek, Canada). DNA from whole blood (methylome) was isolated using the DNA Blood 200 360 prefilling H96 Kit (CMG-717, Perkin Elmer, UK) and 200 µl of whole blood on the Chemagic 360 system (Perkin Elmer, UK) was used. Microbial DNA was extracted from saliva aliquots using a modified protocol from Teng et al (2018) [21] using the DNeasy Blood and Tissue kit (Qiagen, UK). All Extracted DNA was evaluated using the Qubit® 3.0 fluorometer (Thermo Scientific, UK) and Nano Drop 1000 spectrophotometer (Thermo Scientific, UK) and if to be sequenced, using the Invitrogen™ Quant-iT™ PicoGreen™ dsDNA Assay Kit (P7589) on the Hamilton Microlab Star before storage at -80 °C.

RNA isolation

RNA from saliva was isolated using the Oragene RNA purification protocol and Qiagen RNeasy micro kit (Qiagen, UK), RNA from whole blood using the Chemagic 360 system (Perkin Elmer, UK) with Chemagic RNA Tissue 360 H96 Kit (CMG-1212). Purity and quantity were assessed as above for DNA but with Invitrogen Quant-iT RiboGreen Assay Kit (R11490). Integrity (RIN) was determined using the Agilent 4200 TapeStation and RNA ScreenTape (5067-5366), before storage at -80 °C prior.

Clinical data

Self-reported data on physical and mental wellbeing

All participants completed two surveys as part of the trial. The General Health Questionnaire (GHQ-12) is a self-administered 12 item screening tool designed to detect current mental state disturbances in primary care settings, a score of ≥ 2 indicates a disorder. The Health and Lifestyle questionnaire (HLQ) is a survey tool designed by UU to capture key health-related data not present on the ECR. Fields included; COVID-19 risk factors, medications, comorbidities, hospitalisation information, symptoms at admission,

lab tests, family history, drinking status and occupation. The same protocol is being followed for all prospective appointments (ongoing).

Clinical database development

The participants' consent forms, as well as data from the self-reported questionnaires, but with all Personally Identifiable Information (PII) removed by the project's data controller as per GDPR guidelines, was also recorded into the CovresNIdb database. Data were subjected to quality control by two independent researchers against the original sources. The same protocol is being followed for all prospective appointments (ongoing).

Electronic Care Records

In addition to the self-reported data, consent was also given by each patient to enrich the database by accessing their Northern Ireland Electronic Care Records information (NIECR) held by the NHS. PCR positive dates, severity (hospitalised due to COVID-19 infection, or recovered from COVID-19 infection at home), lab results (full blood count, blood pressure, lipids, CRP, GFR, troponin), treatment administered, drugs prescribed within the last six months and co/multimorbidity's held on record for each patient were recorded.

Omics analyses

Genome

Whole genome library preparation was performed using the Illumina TruSeq PCR Free Library Prep protocol (20015963) with an input amount of 1 µg on a Hamilton NGS Star robotic workstation, Quality assessed using Roche KAPA Library Quantification Kit (7960298001) before pooling and sequencing (150 bp paired end (PE)) on an Illumina NovaSeq 6000 instrument using NovaSeq 6000 S4 Reagent Kit v1.5 (20028312), mean coverage of 30X as described previously [22]. Sequences are being uploaded to the European Genome-phenome Archive (EGA)

Methylome

Methylation analysis was performed on DNA samples from saliva (*n*=450) and whole blood (*n*=40) using the Illumina Infinium Methylation EPIC largely as described previously [22]. Data was adjusted for

known epigenetic covariates and surrogate variable analysis was performed via the *sva* inference module [23]. Our in-house developed tool CandiMeth [24] will be employed to streamline methylation analysis for gene lists of interest.

Transcriptome

RNA-Sequencing library preparation used the Illumina TruSeq Stranded Total RNA Library Prep Globin kit (20020612) with an input amount of 100 – 1000 ng. Library preparation was automated and processed using a Hamilton NGS Star and quality was assessed using the Roche KAPA Library Quantification Kit (7960298001) and GX Caliper HS Assay (CLS760672, 760517), run on Roche Lightcycler 480 II and Perkin Elmer LabChip GX Touch analysers, respectively. Libraries were pooled and sequenced (75bp PE) on an Illumina NovaSeq 6000 instrument using NovaSeq 6000 S2 Reagent Kit v1.5 (20028314) targeting 50M paired reads. Raw data (BCL format) were demultiplexed and converted to FASTQ format using BCL2FastQ (Illumina). Adapters were trimmed using Skewer [25] and QC assessed using FASTQC. STAR [26] was used to align reads to the reference genome (GRCh38/hg38) as well as to the transcriptome (GENCODE v. 25). The quality of the RNA alignment was assessed using Picard QC. Gene and isoform quantification will be performed using RSEM [27] with prospective patient (1 and 3 month) T-cell receptor sequencing completed following flow cytometry..

Microbiome

16S rRNA gene amplicons for sequencing by Illumina MiSeq system (Illumina, USA) were prepared using the V3 and V4 region as described in Klindworth et al (2013), with sequencing performed in-house.

Proteome

Protein analysis of 400 plasma samples (baseline) (186 non-hospitalised, 214 hospitalised), 40 prospective (20 non-hospitalised, 20 hospitalised; 1 and 3 month), was outsourced to OLINK proteomics (OLINK, Uppsala, SW) using the Explore® 384 Inflammation panel (Protein Proximity Extension assay). EDTA plasma samples were thawed at room temperature (20°C) and 45 µl of each plasma sample was (at random) pipetted into a LightCycle® 480Multiwell Plate 96-well white PCR plates (Roche Molecular Systems Inc, Charles Avenue, Burgess Hill, West Sussex, UK; Product no. 04729692001) with 8 x wells

left empty on each plate for internal controls to be added at OLINK. Samples were inactivated as per OLINK's protocol and shipped on dry ice (CO2, -78°C). Only samples above 0.2 Normalised Protein Expression (NPX) and samples that deviate less than 0.3 NPX passed QC

The MSD plasma multi-Spot assay system comprising V-PLEX COVID-19 serology panel 11, 'total IgG' and 'ACE2 neutralisation' assays were used to determine viral variant prevalence. Samples were prepared at 1:10 (ACE2) and 1:5000 (neutralisation) for specific assays, then treated essentially as in [28].

The Roche COBAS Elecsys, SARS-CoV-2 spike (S) protein receptor binding domain (RBD) assay was used to determine SARS-CoV-2 antibody presence. as per manufacturer's instructions.

Statistics

Univariate and multivariate analysis

Only patients from *cohort a* who had their BMI recorded on the database (n=507) were selected for the odds ratio analysis. We considered the following risk factors: gender, age, BMI, and disease subgroups. First, univariate analyses (Table 1, fishers exact test) were performed to identify risk factors associated with COVID-19 severity. P-values for univariate analyses were generated using Fisher's exact test comparing frequencies of each potential risk factor between non-hospitalised and hospitalised participants. Variables with a p value < 0.001, i.e. gender, age <50 years, >50 years, cardiovascular, respiratory, endocrine, and musculoskeletal comorbidities, were considered clinically relevant and entered into the multivariable logistic regression model (Table 2). This and further analysis is being undertaken on Base-R software (version 4.2.2) using the Visdat library.

Demographics Table

The demographic table below (Table 1) of COVRES data (n = 519) was generated using IBM SPSS Statistics for Windows, version 27 (IBM Corp., Armonk, N.Y., USA)' [29]. Statistical analysis for the contingency table was undertaken using Fishers exact two-sided test to obtain required P-values and confidence rates were set at 95 %.

Bioinformatic analyses

Bioinformatic analyses will focus on using computational approaches to identify genomic, transcriptomic, proteomic and clinical correlates of severity. Planned analyses primarily include the identification of clinical features, gene variants (host)/eQTLs, transcriptomics signature, cytokine profiles associated with disease severity, as well as the differential methylation among the host genomes of the severity groups.

Variant calling will use mathematical models from the Best Practices Genome Analysis Toolkit. Data is being stored according to genomic position in the Genuity Science Genomically Ordered Relational Database (GORdb) to facilitate rapid access by the Clinical Sequence Analyzer™ user interface and Sequence Miner visualisation software's. Initial data processing for methylome analysis will be carried out in *GenomeStudio* (Illumina version 3.2) prior to import of idat files to the RnBeads package [30] using RStudio (version 2022.02.0+443) on the R platform (version 4.1.2). Quality control will be performed using the *greedycut* algorithm, involving the removal of probes with missing values and poor quality. For RNA-seq, gene and isoform quantification will be performed using RSEM [27] before further analysis is carried out. 16S analysis has been previously described (see above) and OLINK data will be processed in R as per standard pipelines. Whole genome sequencing and transcriptomics data are to be deposited in the EGA [EGAS pending] and shared as a collaboration with the International Covid19 Host Genetics Initiative.

Results

Retrospective cohort demographics (*cohort a*)

The main demographic features are summarised in Table 1. As expected, there was a significant difference in mean age between hospitalised and non-hospitalised patients, as well as gender (both $P<.001$). Age bias was also evident, with 63 % of under 50 years in the non-hospitalised ($P<.001$) and 72 % of over 50 years in hospitalised, ($P<.001$). As expected, comorbidity incidence was higher in the hospitalised subgroup, with auto-immune ($n=12$; $n=26$, $P<.001$), metabolic ($n=33$; $n=94$, $P<.001$), respiratory ($n=39$; $n=83$, $P<.001$), cardiovascular ($n=32$; $n=100$, $P<.001$) and musculoskeletal ($n=23$;

278 *n*=58, *P*<.001) disorders of note. There was no difference between cohorts for gastrointestinal disorders
279 (Table 1).
280
281 ## Prospective cohort demographics (*cohort b*)
282 Data collection is ongoing for 40 participants which are being followed up over 12 months: *n*=20 (8
283 female) hospitalised; *n*=20 non-hospitalised (12 female), gender distribution is not significantly different
284 between subgroups (*P*<.21), but average age is (hospitalised; 52 years, non-hospitalised; 45 years,
285 *P*<.001) Table 2). Contrasting with initial recruitment, within this follow up *cohort b* there was no >50
286 age bias (*P*<.20), there was also no significant difference in vaccination status between hospitalisation
287 subgroups (*P*<.55) and only cardiovascular disease as a comorbidity was more prevalent in hospitalised
288 patients (*P*<.02), though numbers are small.

289 **Table 1 COVRES *cohort a* demographic information.**
290 P value calculated using 2-sided Fisher's exact test between non-hospitalised versus hospitalised. < .05 set
291 as statistical significance. n = 519, * continuous variables used a 2-sided t-test.

Cohort demographics		**Non-hospitalised (n=270)**	**Hospitalised (n = 249)**	**Total (n = 519)**	**P value**
Gender:					
	Female, n (%)	184 (64.3)	102 (35.7)	286 (55.1)	< .001
	Male, n (%)	83 (36.7)	143 (63.3)	226 (43.5)	< .001
	Other, n (%)	3 (42.9)	4 (57.1)	7 (1.3)	.715
Age at diagnosis:					
	Mean (Std.Dev.)	45.4 (13)	56.5 (12.7)	50.7 (14	*< .001
	Under 50 years old, n (%)	169 (62.6)	67 (26.9)	236 (45.5)	< .001
	Over 50 years old, n (%)	101 (37.4)	182 (73.1)	283 (54.5)	< .001
Disease subgroup*:					
	1. Autoimmune, n (%)	12 (4.4)	26 (10.4)	38 (7.3)	.011

2. Metabolic, n (%)	33 (12.2)	94 (37.8)	127 (24.5)	<.001
3. Respiratory, n (%)	39 (14.4)	83 (33.3)	122 (23.5)	< .001
4. Cardiovascular, n (%)	32 (11.9)	100 (40.2)	132 (25.4)	< .001
5. Cancer, n (%)	7 (2.6)	21 (8.4)	28 (5.4)	.003
6. Gastrointestinal, n (%)	13 (4.8)	21 (8.4)	34 (6.6)	.111
7. Musculoskeletal, n (%)	23 (8.5)	58 (23.3)	81 (15.6)	< .001

*Disease subgroup key:

1. Autoimmune or rheumatic disease including: rheumatoid arthritis, systemic lupus erythematosus, and multiple sclerosis.
2. Metabolic or endocrine disease including: thyroid conditions, hypercholesterolaemia or other hyperlipidaemia, gout, diabetes, and kidney disorders.
3. Respiratory disorder Chronic lung diseases including: chronic obstructive pulmonary disease, asthma (moderate-to-severe), interstitial lung disease, cystic fibrosis, sleep apnoea and pulmonary hypertension.
4. Cardiovascular system disorders including: angina, hypertension, stroke, peripheral vascular disease, balloon angioplasty or percutaneous coronary intervention, atrial fibrillation, venous thromboembolism, anaemia, and chronic cardiac disease other than hypertension.
5. Cancer including: leukaemia, lymphoma, malignant solid tumour, and to include current, past and remission.
6. Gastrointestinal disorders including: gallbladder, liver disease, pancreatic disease, and Inflammatory bowel syndrome.
7. Musculoskeletal disease including: osteoarthritis and ankylosing spondylitis, excludes subgroup 1 conditions.

Table 2: COVRES2 cohort multivariate analysis, hospitalised versus non-hospitalised analysing risk factors for COVID-19 severity

Inclusion criteria for analysis: participants (n=40) were required to have a BMI score recorded.

Prospective Omicron cohort demographics	**Non-hospitalised (n = 20)**	**Hospitalised (n = 20)**	**Total (n = 40)**	***P*-value**
Female, n (%)	12 (60)	8 (40)	22 (50)	0.206
Age at diagnosis: Mean (Std.Dev.)	45.2 (13.5)	52 (17.2)	48.6 (15.6)	***<0.001**
Over 50 years old, n (%)	7 (35)	11 (55)	18 (45)	0.204
Vaccine status , n (%)	19 (95)	18 (90)	37 (92.5)	0.548
Comorbidity:				
1. Autoimmune, n (%)	3 (15)	6 (30)	9 (22.5)	0.256
2. Metabolic, n (%)	4 (20)	9 (45)	13 (32.5)	0.91
3. Respiratory, n (%)	2 (10)	7 (35)	9 (25.5)	0.58
4. Cardiovascular, n (%)	2 (10)	11 (55)	13 (32.5)	**0.02**
5. Cancer, n (%)	0 (0)	4 (20)	4 (10)	0.35
6. Gastrointestinal, n (%)	2 (10)	4 (20)	6 (15)	0.376
7. Musculoskeletal, n (%)	3 (15)	7 (35)	10 (25)	0.144

Discussion

The COVRES study provides a novel opportunity to identify multi-omics biomarkers from blood and saliva indicative of COVID-19 severity in Northern Ireland and may provide unique insights into disease mechanisms and identify potential therapeutic targets. The maximum recruitment number (*n*=519) was reached and various analyses are ongoing, including whole genome and RNA sequencing, proteomic, microbiome and methylation experiments. We have also collected detailed medical data using NIECR that will be used to enrich the biomarker data [31]. All *cohort a* participants were recruited over a four month period (1ST December 2020- 31st March 2021) during a pandemic peak, allowing homogeneous data collection from the same viral variant (B.1.1.7) (Figure 1). Compared to other work, the COVRES study has a high participant number and uses a significantly wider biomarker identification approach. This was achieved while significant pandemic restrictions were in place and was only possible due to our local health trust (WHSCT, NHS) collaboration, which facilitated patient access and enabled recording of laboratory parameters that have not been possible in other studies [32]. A recent multi-omics COVID-19 study utilised proteomics and metabolomics to screen thirteen samples at two time points and found 10 significant proteins, 32 significant peptides, and 5 metabolites that were dysregulated in severe patients [33]. Recruitment for this study also occurred in early 2021, but the small scale (*n*=13) brings into question the generalisability of the findings. Another multi-omics study based in the USA sampled one hundred and twenty-eight individuals between 6th April 2020 and 1st May 2020 and conducted follow-up until June 2020. The authors quantified transcripts, proteins, metabolites, and lipids and made associations to clinical outcomes [34]. Links were made between platelet function, blood coagulation and endotheliopathy and a severe COVID-19. Our study builds on these smaller studies and may offer increased stastistical power and potential to validate or compare markers identified.

The COVRES study was designed to recruit hospitalised (*n*=250) COVID-19 patients, classified as having severe infection and non-hospitalised (*n*=250) COVID-19 patients, classified as having mild infection, within three months of sampling. It is worth noting that the recruitment of non-hospitalised COVID-19 patients makes this cohort particularly valuable, most trials have only involved either patients

who have been admitted to hospital or those who have not [35, 36], and few have investigated earlier stages of the disease process such as pre-exposure, or post-exposure and outpatient treatment.

To maximise impact and benefit to the scientific and healthcare communities, this study was designed to be cross-border covering both NI and RoI. The global drive to identify clinical biomarkers of COVID-19 severity has led to many clinical studies and trials that have varied methodology, in terms of different control groups, follow-up periods, omics of interest and lab methodologies [37-39]. Studies have also been carried out in different geographical regions without any standardised operating procedures and have been powered according to different endpoints [40]. This variation makes reproducibility questionable, and it is difficult to apply findings across geographic regions and variant time periods. To align with as many studies as possible COVRES participants have been classified according to the WHO [41], and to facilitate cross-border collaboration we coordinated with Trinity College Dublin. We also plan to share our Whole Genome Sequencing (WGS) data with the EGA for the advancement of science and improved public health outcomes.

The recruitment of non-hospitalised and hospitalised COVID-19 patients in NI is a main strength of the COVRES study and adds novelty to existing research regarding COVID-19 severity with the majority recruiting patients based on a positive PCR test regardless of hospitalisation. Gender and age matching was considered but an exact match not achieved due to the complexities and limitations COVID-19 presented in terms of patient access [5]. The mean age of the hospitalised COVRES subgroup was 56.5 years (Table 1), slightly younger compared to a large UK wide observational study [42] (*n*=20,908 hospitalised) which had a mean age of 62 years. There was no difference between gender (male 49 %, female 51 %), compared to our 43.5 % male. Another smaller (*n*=429) UK study found the average age of hospitalised COVID-19 patients to be 70 years and a male bias 57 % which is close to our study. Corresponding with our study, they also found the average BMI to be 28 kg/m^2 (overweight - obese) and highly comorbid (Table 2), with the most common comorbidities being Type-2 diabetes, hypertension, and respiratory disorders [42, 43]. The previous study (Ken-Dror, et al., 2020) (51) in England is a good

comparison for COVRES NI as the recruitment protocols and cohort demographics are similar. The similarities in the data are promising and may indicate that our findings could be useful to the wider UK.

Limitations

COVRES participants were all sampled at a single time point, limiting our ability to assess genomic, proteomic, and immune biomarkers as the disease progresses. Future work will focus on obtaining follow-up samples to enable longitudinal analysis and assess prognostic capability of markers of interest. Manual data input at some points increases the risk of human error [44]: although quality control checks were carried out between two WHSCT staff members there is inherent risk of incorrect data.

It also needs to be considered that the COVRES cohort represents a COVID-19 population recruited in NI and the demographics show a low representation of ethnic minorities, therefore data may not be able to be generalisable beyond Caucasian Irish/UK populations.

Conclusion

The COVRES Study offers a novel opportunity to study multiomics mechanisms of COVID-19 severity in re-contactable participants. This research has the potential to impact COVID-19 clinical decision making and therapeutic development. Our WHSCT and industry collaborators enabled rapid and effective recruitment, allowing us to reach our goal of n=500, and begin analysis pipelines immediately. We hope that this paper will not only demonstrate the effectiveness of the study methodology but will also raise awareness of the availability of this cohort to researchers in the field and promote future collaboration.

List of abbreviations

ACE2 - angiotensin-converting enzyme 2

BMI - body mass index

COVRES - Covid Response Study

CTRIC - Clinical Translational Research and Innovation Centre

CVD – Cardiovascular disease

DDP9 - dipeptidyl peptidase 9

ECR - Electronic Care Record

EGA - European Genome-phenome Archive

GDPR - general data protection regulations

GHQ-12 - general health questionnaire

GWAS - genome wide association study

NIECR - Northern Ireland Electronic Care Record

PII - Personally Identifiable Information

PIS - Participant Information Sheet

SOP - Standard operating procedures

TMPRSS2 - type II transmembrane serine protease

TYK2 - tyrosine kinase 2

UU – Ulster University

WHO - World Health Organisation

WHSCT - Western Health and Social Care Trust

383

384 Declarations

385 Ethics approval and consent to participate - The COVRES study was approved by the Health and Care
386 Research Wales Ethics service on the 14th of July 2020 (REC ref 20/WA/0179). All participants provided
387 informed consent toparticipate.

388 Consent for publication – All relevant permissions were obtained during the consent process.

389 Availability of data and materials - Data and meta data will be stored according to UU policy on data
390 management and sharing. Data will be available via Ulster University's Research Data Repository and in
391 accordance with their Research Data Management Policy. Any personal/identifiable information will be
392 redacted, data queries will be addressed on an individual basis by the research team. Genomics data are
393 being deposited with the EGA and will beavailable on request.

Completing interests – Authors declare no completing interests

Funding - This work was supported by a research grant to AJB (PI) and coauthors from the Department for the Economy Northern Ireland under its contribution to Science Foundation Ireland's COVID-19 Rapid Response Call (Phase 2) and Opportunity-Led Research Awards to TSR and DSG from HSC R&D Division, Public Health Agency (COM/5631/20).

Generative AI – Generative AI was not used for any portion of the manuscript writing

Author contributions – English, A - Investigation, formal analysis, writing, resources, supervision; McDaid, D - Formal analysis, data curation; Lynch, S. - Formal analysis, Investigation; McLaughlin, J. - Investigation, formal analysis, writing; Cooper, E. - Formal analysis, data curation; Wingfield, B. - Formal analysis, data curation; O'Kane, M. - Conceptualisation, project administration; Kelly, M. - Conceptualisation, project administration; Bhavsar, M. - Conceptualisation, project administration; McGilligan, V. - Conceptualisation, project administration, supervision; Irwin, R. - Formal analysis, data curation; Bucholc, M. - Formal analysis, data curation; Zhang, S. - Formal analysis, data curation; Shukla, P. - Formal analysis, data curation; Rai, T. - Conceptualisation, project administration, supervision, funding acquisition; Bjourson, A. - Conceptualisation, project administration, supervision, funding acquisition, resources; Murray, E. - Conceptualisation, project administration, supervision, funding acquisition; Gibson, D. - Conceptualisation, project administration, supervision, funding acquisition, resources, writing; Walsh, C. - Conceptualisation, project administration, supervision, funding acquisition, writing.

Acknowledgements - We would like to extend our sincere gratitude to the participants of this study for their invaluable contribution and to the Western Health Social Care Trust (WHSCT), Northern Ireland, for facilitating access to patients and clinics, Dr Maurice O'Kane for enabling WHSCT lab completion of Roche anti-SARS-CoV-2 antibody assays and Genuity Science (Ireland) Ltd for their funding in-kind for completion of the WGS and methylation analysis.

References

1. Heneghan, C., J. Brassey, and T. Jefferson, *COVID-19: What proportion are asymptomatic.* Centre for Evidence-Based Medicine, 2020. **1**.
2. Bhargava, A., et al., *Predictors for severe COVID-19 infection.* Clinical Infectious Diseases, 2020. **71**(8): p. 1962-1968.
3. Jordan, R.E., P. Adab, and K. Cheng, *Covid-19: risk factors for severe disease and death*. 2020, British Medical Journal Publishing Group.
4. Zeberg, H. and S. Pääbo, *A genomic region associated with protection against severe COVID-19 is inherited from Neandertals.* Proceedings of the National Academy of Sciences, 2021. **118**(9).
5. Pairo-Castineira, E., et al., *Genetic mechanisms of critical illness in Covid-19.* Nature, 2021. **591**(7848): p. 92-98.
6. Mosharaf, M., et al., *Computational identification of host genomic biomarkers highlighting their functions, pathways and regulators that influence SARS-CoV-2 infections and drug repurposing.* Scientific reports, 2022. **12**(1): p. 1-22.
7. Singh, H., R. Nema, and A. Kumar, *Genomic, proteomic biomarkers and risk factors associated with COVID-19*, in *Advanced Biosensors for Virus Detection*. 2022, Elsevier. p. 95-111.
8. Lynch, S.M., et al., *Role of senescence and aging in SARS-CoV-2 infection and COVID-19 disease.* Cells, 2021. **10**(12): p. 3367.
9. Ponti, G., et al., *Biomarkers associated with COVID-19 disease progression.* Critical reviews in clinical laboratory sciences, 2020. **57**(6): p. 389-399.
10. Hsu, H., et al., *Clinical informatics during the COVID-19 pandemic: Lessons learned and implications for emergency department and inpatient operations.* Journal of the American Medical Informatics Association, 2021. **28**(4): p. 879-889.
11. Tjendra, Y., et al., *Predicting disease severity and outcome in COVID-19 patients: a review of multiple biomarkers.* Archives of pathology & laboratory medicine, 2020. **144**(12): p. 1465-1474.
12. Tayefi, M., et al., *Challenges and opportunities beyond structured data in analysis of electronic health records.* Wiley Interdisciplinary Reviews: Computational Statistics, 2021. **13**(6): p. e1549.
13. Pourbagheri-Sigaroodi, A., et al., *Laboratory findings in COVID-19 diagnosis and prognosis.* Clinica Chimica Acta; International Journal of Clinical Chemistry, 2020. **510**: p. 475.
14. Singh, H., et al., *ACE2 and TMPRSS2 polymorphisms in various diseases with special reference to its impact on COVID-19 disease.* Microbial Pathogenesis, 2021. **150**: p. 104621.
15. Strope, J.D. and C.H. Chau, *TMPRSS2: Potential biomarker for COVID-19 outcomes.* Journal of clinical pharmacology, 2020.
16. Fernández-Pato, A., et al., *Plasma miRNA profile at COVID-19 onset predicts severity status and mortality.* Emerging microbes & infections, 2022. **11**(1): p. 676-688.
17. Russell, C.D., et al., *Tissue Proteomic Analysis Identifies Mechanisms and Stages of Immunopathology in Fatal COVID-19.* American journal of respiratory cell and molecular biology, 2021(ja).
18. Spick, M., et al., *An integrated analysis and comparison of serum, saliva and sebum for COVID-19 metabolomics.* Scientific reports, 2022. **12**(1): p. 1-12.
19. Harvey, L., *REDCap: web-based software for all types of data storage and collection*. 2018, Nature Publishing Group. p. 625-625.

20. Golke, T., et al., *Delays during PBMC isolation have a moderate effect on yield, but severly compromise cell viability.* Clinical Chemistry and Laboratory Medicine (CCLM), 2022. **60**(5): p. 701-706.
21. Lapsley, C.R., et al., *Methylome profiling of young adults with depression supports a link with immune response and psoriasis.* Clinical Epigenetics, 2020. **12**(1): p. 1-16.
22. Dhingra, R., et al., *Evaluating DNA methylation age on the illumina MethylationEPIC bead chip.* PloS one, 2019. **14**(4): p. e0207834.
23. Leek, J.T., et al., *sva: Surrogate variable analysis.* R package version, 2019. **3**(0): p. 882-883.
24. Thursby, S.-J., et al., *CandiMeth: Powerful yet simple visualization and quantification of DNA methylation at candidate genes.* GigaScience, 2020. **9**(6): p. giaa066.
25. Jiang, H., et al., *Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads.* BMC bioinformatics, 2014. **15**(1): p. 1-12.
26. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
27. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC bioinformatics, 2011. **12**(1): p. 1-16.
28. Collins, E.S., et al., *A clinically based protein discovery strategy to identify potential biomarkers of response to anti-TNF-α treatment of psoriatic arthritis.* PROTEOMICS–Clinical Applications, 2016. **10**(6): p. 645-662.
29. Statistics, I., *IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.* Google Search, 2013.
30. Müller, F., et al., *RnBeads 2.0: comprehensive analysis of DNA methylation data.* Genome biology, 2019. **20**(1): p. 1-12.
31. Bucholc, M., et al., *Identifying pre-existing conditions and multimorbidity patterns associated with in-hospital mortality in patients with COVID-19.* Scientific Reports, 2022. **12**(1): p. 17313.
32. Sáez, C., et al., *Potential limitations in COVID-19 machine learning due to data source variability: A case study in the nCov2019 dataset.* Journal of the American Medical Informatics Association, 2021. **28**(2): p. 360-364.
33. Suvarna, K., et al., *A multi-omics longitudinal study reveals alteration of the leukocyte activation pathway in COVID-19 patients.* Journal of Proteome Research, 2021. **20**(10): p. 4667-4680.
34. Overmyer, K.A., et al., *Large-scale multi-omic analysis of COVID-19 severity.* Cell systems, 2021. **12**(1): p. 23-40. e7.
35. Caricchio, R., et al., *Effect of canakinumab vs placebo on survival without invasive mechanical ventilation in patients hospitalized with severe COVID-19: a randomized clinical trial.* Jama, 2021. **326**(3): p. 230-239.
36. Salvarani, C., et al., *Effect of tocilizumab vs standard care on clinical worsening in patients hospitalized with COVID-19 pneumonia: a randomized clinical trial.* JAMA internal medicine, 2021. **181**(1): p. 24-31.
37. Chemaitelly, H., et al., *mRNA-1273 COVID-19 vaccine effectiveness against the B. 1.1. 7 and B. 1.351 variants and severe COVID-19 disease in Qatar.* Nature medicine, 2021. **27**(9): p. 1614-1621.
38. Tolaney, S.M., et al., *The impact of COVID-19 on clinical trial execution at the Dana-Farber Cancer Institute.* JNCI: Journal of the National Cancer Institute, 2021. **113**(11): p. 1453-1459.
39. Robishaw, J.D., et al., *Genomic surveillance to combat COVID-19: challenges and opportunities.* The Lancet Microbe, 2021. **2**(9): p. e481-e484.
40. Park, J.J., et al., *How COVID-19 has fundamentally changed clinical research in global health.* The Lancet Global Health, 2021. **9**(5): p. e711-e720.

41. de Terwangne, C., et al., *Predictive accuracy of COVID-19 world health organization (Who) severity classification and comparison with a bayesian-method-based severity score (epi-score).* Pathogens, 2020. **9**(11): p. 880.
42. Goodacre, S., et al., *Characterisation of 22445 patients attending UK emergency departments with suspected COVID-19 infection: Observational cohort study.* PloS one, 2020. **15**(11): p. e0240206.
43. Ken-Dror, G., et al., *COVID-19 outcomes in UK centre within highest health and wealth band: a prospective cohort study.* BMJ open, 2020. **10**(11): p. e042090.
44. Menachemi, N. and T.H. Collum, *Benefits and drawbacks of electronic health record systems.* Risk Manag Healthc Policy, 2011. **4**: p. 47-55.

521