NIHR150393: Exploring and grouping COVID-19 pharmaceutical interventions to determine mechanisms of action and endotypes of response

This project is funded by the NIHR Efficacy and Mechanism Evaluation (EME) Programme (NIHR150393). The views expressed in this protocol are those of the author(s) and not necessarily those of the NIHR, NHS England or the Department of Health and Social Care

1 Research design

This research project is designed to identify the biological mechanisms through which novel COVID-19 drugs provide therapeutic benefit using data obtained from the ACCORD clinical trial(1). We will compare three novel therapeutics (Bemcentinib, Tozorakimab (previously known as MEDI3506) and Zilucoplan) to the Standard of Care (SoC) for severe COVID-19 patients using multiple data types (transcriptomics, proteomics, viral load, and pharmacokinetic drug levels). Each COVID-19 drug will be compared to the SoC of care individually before comparing the identified biological mechanisms comparatively to compare the mechanisms and conditions being modulated favourably by these drugs. Additionally, we will explore biological mechanisms and differences and relationships to effect size of clinical response as measured by time to improvement or discharge. This will better help us understand how the novel drugs work and how to select efficient treatments for a particular patient

1.1 Study Population

The inclusion criteria for the ACCORD trial were:
Adults (>= 18 years) with SARS-CoV-2 infection confirmed by laboratory tests and/or point of care tests. Patients must be scored between Grade 3 (hospitalised – mild disease, no oxygen therapy) to Grade 5 (hospitalised - severe disease, non-invasive ventilation or high flow oxygen) on ACCORD's 9 point WHO ordinal scale from clinical improvement (OSCI) where Grade 0 is uninfected and 8 is death. Overall, in the ACCORD study 167 subjects were randomised, 14 subjects withdrew and 25 subjects died. In the additional Bemcentinib study, a further 115 were randomised – 60 to SoC and 55 to Bemcentinib in an identical study design.

1.2 Sampling.

Samples from patients were collected in line with the sample collection schedule below:

Table showing the sample collection schedule for the ACCORD COVID-19 clinical trial.

	Baseline								
Day	**1**	**3**	**4**	**5**	**8**	**10**	**11**	**15**	**29**
Biological sample type									
Blood (for RNA)	X			X		X			
Blood (for serology)	X							X	
Blood (for protein)	X			X		X			
Oropharyngeal swab (for viral load)	X	X		X	X		X	X	X
Pharmacokinetic drug concentration									
Bemcentinib	X		X[a]		X			X	
Tozorakimab	X[b]							X[c]	
Zilucoplan	X			X		X		X	

Adapted from (1) to show only the samples collected to generate the datasets relevant to this current project. X = sample collected. All patients were sampled on day 1, which is considered baseline prior to the first dose of the assigned treatment arm or standard of care. Each drug had slightly different pharmacokinetic (PK) sampling regimes and were mostly collected pre-dose. Where samples for PK were collected other than pre-dose, collection information is detailed below:

[a] Sample collected pre-dose and 6 hours post-dose (Day 4).
[b] Sample collected pre-dose and 15 minutes post-dose (Day 1).
[c] Sample collected pre-dose and 15 minutes post-dose (Day 15).

1.2.1 Datasets

The datasets generated from the above sampling protocol of the ACCORD trial include transcriptomics, proteomics, viral load, serology, pharmacokinetic drug concentration and clinical laboratory datasets. Further information about the generation and processing of each dataset is outlined below:

1.2.1.1 Transcriptomic dataset (RNA Sequencing)

Sequencing was performed on patient blood samples taken on days 1 (at baseline prior to treatment), day 5 and day 10 of admission to hospital. A total of 286 samples were sequenced using an Illumina NovaSeq 600 platform. Total RNA was prepared from blood and depleted of highly abundant transcripts for alpha and beta haemoglobin and ribosomal RNA. No more than 50 samples per 'lane' were multiplexed to maximise sequence read depth to ensure a minimum sample read depth of 100 to 150 M. Quality control (QC) of raw read data was performed using FastQC version 0.11.9 and compiled and visualised using MultiQC version 1.12. Samples with <20 million total reads and less than 50% of the total read count were excluded from further analysis. After MultiQC analysis, fourteen sample were removed that failed quality control, leaving 272 high-quality samples available in the transcriptomic dataset for downstream analysis.

1.2.1.2 Proteomic dataset (Luminex)

A Bio-Plex Pro Assay kit was utilised to assess the concentrations of cytokines, chemokines and growth factors in the study population from samples taken on days 1 (pre-treatment), 5 and 10 of admission. The 27 cytokines, chemokines and growth factors measured are: IL-1β, IL-1ra, IL-2, IL-4, IL-5, IL-6, IL-7, CXCL8/IL-8, IL-9, IL-10, IL-12p70, IL-13, IL-15, IL-17A, Eotaxin, Basic FGF, G-CSF, GM-CSF, IFN-γ, IP-10, CCL2/MCP1, CCL3/MIP-1α, CCL4/MIP-1β, PDGF-BB, CCL5/RANTES, TNF-α and VEGF. Following analysis of each batch, 27 reports (one for each analyte) containing the assay results (in pg/ml) were generated using the xPONENT for MagPix software and saved within the study specific electronic file. A CSV file will be generated and exported for further analysis.

1.2.1.3 SARS-CoV-2 viral load dataset

RNA was extracted using the QIAmp viral RNA mini kit (Qiagen). PCR for SARS-CoV2 targets was performed using the CE-marked COVID-19 genesig Real-Time PCR assay (Primerdesign).

1.2.1.4 SARS CoV-2 serology dataset

Anti-SARS-CoV2 IgG antibodies in serum were measured using the Elecsys Anti-SARS-CoV-2 nucleoprotein assay (Roche)(2)

1.2.1.5 Pharmacokinetic drug concentration dataset

A pharmacokinetic dataset is available for each subject who received at least 1 dose of experimental drug and in whom drug levels were detectable. PK assessments were specific to the candidate agent and delivered by the pharmaceutical partners using validated assays for drug registration.

1.2.1.6 Clinical Laboratory variables dataset

Standard clinical laboratory safety testing was performed on all subjects, including haematology, coagulation and clinical chemistry readouts which are available for analysis of response.

2 Analysis plan

2.1 Missing data, interpolation, and imputation

Missing data will be minimised, with outcomes collected after premature withdrawal of treatment. The primary clinical estimate will be analysed using a treatment policy strategy. If an OSS is missing on a particular study day, the unscheduled visits will be considered to see whether there was an unscheduled visit on that day. If so, the score from the unscheduled visit will be used. For the purposes of the primary analysis patients who died or were transferred will not be censored at day of death/transfer, time to clinical response will instead be censored as 29 days. Patients are considered to be in the study until Day 29 but a score of 8 will be imputed for the OSS for those who died from their date of death and for all subsequent visits. Moreover, if a patient withdraws and then dies and we have a record of their death then a score of 8 will be imputed from their date of death onwards for all analyses. For those who transferred, additional data post transfer has been collected and will be used. If a patient is discharged, has no Day 29 value but has a day 60 value they will be considered Alive for the mortality analysis and Last observation carried forward (LOCF) will be used to impute Day 29. If a patient has discharged and no additional data are available, then they are assumed alive at Day 29 and LOCF will be used to impute Day 29 value. If the OSS is missing on the date of discharge, this will be imputed to 2 (most conservative value possible for discharged patients). If the OSS is ≥ 3 on the date of discharge, then the score will be imputed to 2. If a patient is subsequently readmitted and has ordinal scale data, then this data will be used. However, if data has not been entered upon readmission, then the patient will be considered a non-responder for the primary endpoint. Once they have been discharged then if OSS is missing it will be imputed as 2 for subsequent days. Reported ethnicity will be compared with ancestral ethnicity inferred using genetics data (Peddy software) and mapping of genotypes to gnomAD v2.1.1 data. Inferred (genetic ancestry) will be preferred over reported ancestry where discrepancies occur. Sex will be calculated counting genotypes on the X chromosome and be used preferentially where gender may not match biological sex.

2.2 Comparison and confounding of baseline clinical characteristics

Baseline clinical characteristics, including age, sex, ethnicity, significant comorbidities, laboratory results and clinical observations will be recorded, assessed and compared between patients in different treatment arms using R (v 4.1.2). Patient statistics tables will be calculated to test for any significant confounding variables between blinded groups. Principal component analysis of normalised expression values will be correlated with axes of variation e.g. ethnicity, to identify clusters affecting gene expression.

Differential gene expression analysis will control for baseline factors that may skew analysis by updating the statistic model design in DESeq2. Differential splicing analysis will be subject to weighting by multinomial covariate-balanced propensity scores to control for baseline confounding factors, as implemented by the R package WeightIt version 0.12.0, using the “just-identified” approach. Proteomics models will incorporate weighted baseline clinical data to avoid bias.

2.3 Transcriptomic analysis

The focus of this study will be the analysis of the transcriptomics dataset comprising patient data taken at three time points for three novel drugs and the SoC trial arm. The RNA-Sequencing samples have been generated and the raw FASTQC data have been downloaded from the ACCORD clinical trials platform. These data have been subject to quality control as previously described in the methods section. To avoid any unintentional bias, the groups will be blinded so that the bioinformatician will not know which patient group or SoC arm is which. The novel drug and SoC arms will be given a random identifier such as A, B, C or D for the duration of this analysis.

2.3.1 Differential Gene Expression analysis

Firstly, we will conduct outlier analysis using data visualization methods such as Principal Component Analysis, Boxplots, Hierarchical Clustering and IQR/Median plots. After testing for potential outliers, non-specific filtering will be conducted on our dataset to remove any very lowly expressed genes. Differential expression testing will be conducted in R version 4.1.2 using the Bioconductor package DESeq2(3). The data will be normalized with the DESeq2 mean of ratios method and baseline characteristics will be controlled for using the function design(model). We will first compare, for each drug arm, differentially expressed genes at pre-treatment versus post-treatment to understand the biological pathways driving drug response. We will modify our analysis as results evolve to test for differences in gene expression and identify transcriptomic gene signatures for genes and/or pathways driving differences between responders and non-responders for each drug. We will then compare gene expression between each treatment and SoC. We will also build smaller models comparing Differentially Expressed Gene (DEG) analysis between drug groups.

2.3.2 Weighted Gene Correlation Network Analysis (WGCNA)

The normalized dataset will undergo variance stabilizing transformation (vst) using the DESeq2 package, to be further explored using Weighted Gene Correlation Network Analysis (WGCNA)(4). WGCNA is an unsupervised analysis method which clusters genes into defined modules of highly interconnected genes based on the similarity of gene co-expression profiles. These modules can be assessed for correlations with specific clinical traits; specifically, we will assess whether any gene module is associated with one of the blinded treatment arms. Additionally, for each gene module, 'hub' genes can be identified, which are genes with high connectivity, and therefore high importance, within a co-expression module. Hub gene analysis and visualization of gene networks will be carried out using Cytoscape(5). Identification of these hub genes may reveal the key molecular drivers within treatment arms or relationships with clinical response.

2.3.3 Gene pathway analysis

Following the identification of DEGs, gene modules of interest and hub genes, Gene Ontology and Pathway Analysis will be performed on each of these lists to establish which pathways are impacted in each comparison. The novel pathways for each arm comparison will guide our work towards specific mechanisms of action in COVID-19 patients. Mechanisms that vary between the arms of this study will also be identified by performing Gene Set Variation Analysis (GSVA)(6).This method generates a score for each pathway within the chosen pathway dataset, allowing for comparison of the pathways scores for each sample across the condition arms. This will be done using the linear modelling package limma. We will use the most recent release of pathway databases for this analysis, such as KEGG or Wikipathways.

2.3.4 Differential splicing analysis

High quality data passing multiQC analysis will be retained for downstream differential splicing expression analysis. STAR version 2.7.1 software will be used to generate an index file mapped to the Human Genome Reference GRCh38 primary assembly and the GENCODE v34 annotation and generated using STAR's genome generate function with the argument --sjdbOverhang 149 and all other settings as default. Individual FASTQ files will be subsequently aligned to GRCh38 using STAR. Resultant unsorted BAM files will be sorted using Samtools version 1.16. Splicing junction files will be generated using RegTools version 0.5.2, and resultant files will be inputted into LeafCutter version 0.2.9 software for differential splicing analysis. Leafcutter only supports comparison of two groups and therefore differential splicing will be compared for each treatment group pre- and post-treatment and then between treatment and SoC. We will implement MAGIQ version 2.3 software for multigroup differential splicing analysis to compare each drug arm.

In summary, this transcriptomic analysis plan will interrogate our transcriptomic dataset on a pathway level as well as on an individual gene basis and enable a full characterisation of the molecular mechanisms and drivers of severe COVID-19 disease response to novel drugs.

2.4 Proteomics Analysis

Following quality control and quantitation of the 27 proteins (undertaken at the Medicines Evaluation Unit as part of the ACCORD trial) we will compare differences between the drugs and SoC arms. The dataset will be visualised using PCA plots with phenotype information overlaid to identify any unknown batches. If detected, batches will be corrected before or accounted for in the statistical modelling. Statistical analysis will be conducted to identify statistical differences between the different groups within this study (Each therapeutic compared to SoC and responders vs non-responders within each therapeutic arm). Shapiro wilk tests will be performed on each variable to assess the normality of the data and if it is so then t tests will be used but more likely we will use Mann-Witney U Tests between these groups. Following this we will also perform multiple testing corrections to the resulting p values.

2.5 Other Factors

Viral Load, SARS-CoV-2 Serology, Pharmacokinetic drug concentration, patient demographics and clinical severity (baseline OSCI score) will be included in the pathway analysis model. These factors will be coded as numerical values and input as clinical traits for analysis of correlation with eigengene values of gene cluster modules within the WGCNA work stream.

2.6 Machine Learning

Machine learning integrates statistical and other mathematical techniques of function approximation and optimization to extract useful information from large and complex datasets. It offers algorithm for computational modelling to extract nonlinear relationship between covariates and responses, quantifying uncertainties in predictions, integrating multi-modal data and for dealing with changing environments such as drifts in population statistics. In this project, building on recent cutting-edge developments in the subject, with some of our own work in particular, we will use machine learning approaches to go beyond classic uses of statistics in bioinformatics for the analysis of the data. The primary goal will be the integrative analysis of gene expression observed at the transcriptomic and proteomic levels, as well as measurements of viral loads and responses to various treatments. We are mindful that the problem at hand is different from challenges such as automatic machine translation or computer vision used for self-driving cars which are characterized by the availability of very large corpora of data (in their millions) used in training very large models (sometimes in the billions of free parameters). The challenges here will be with relatively small amounts of data, carrying substantial uncertainty at levels of quantification and/or measurement and probably imbalanced when stratified by outcomes. Thus, we emphasize that the use of machine learning (expanded in the paragraphs below) is not to build predictive black-box models, but rather extracting useful relationships via the integration of the various sources of data mentioned above. To this end we will use machine learning for: (i) analysis by dimensionality reduction (e.g. Principal Component Analysis and other two dimensional visualization techniques that are robust; (ii) integrative analysis of multi-modal gene expression data (i.e. those genes whose transcript and protein level expressions are observed); (iii) formulation of outlier detection problems (as opposed to classification and regression problems to characterize measured responses; (iv) explore non-linear relationships by means of small-scale or highly regularized neural network models. The four parts will be heavily interlinked (e.g. dimensionality reduction may uncover small subspaces that can be used for efficient training of predictive models) and are not proposed as four sequential work-packages.

2.6.1 Dimensionality Reduction:

As the available transcriptome is of very high dimensions (of the same number of genes in the genome) in comparison to proteome measurements (cytokines), we will use extensive dimensionality reduction and feature selection techniques to extract a meaningful subspace for integrative analysis. Information to be explored at the transcriptome level to match what is being measured at the protein level is not merely the expression levels of the matching genes. Strongly correlated mRNA expressions along specific pathways carry more information than simply those of the matching genes. Additionally, we will use statistically meaningful features that can be derived from the gene sequence following recent work inspired by natural language modelling. This combination of differential expression, sequence features and pathway analysis will give a subspace representation at the transcript level and its correlation (or lack thereof) with protein level measurements will elucidate useful gene expression level information that is not observable from transcript levels alone.

2.6.2 Integrative Analysis:

In a data matrices of gene expressions of individuals at the levels of transcript and proteins, we would expect correlations, both along the axes of subjects (patients of similar characteristics) and genes (acting along the same pathway or regulating one another). Mathematically, such matrices are known as low rank matrices and information-bearing subspaces may be extracted from them by methods known as low rank approximation methods (the popular principal component analysis –PCA) is among the simplest forms of this family. We will build on recent work in analysing such multi-view gene expression data by robust low rank projections(7,8) where we developed efficient algorithm for such analysis and demonstrated their use in single-cell genomic data analysis.

2.6.3 Outlier Analysis:

While much work in the application of machine learning to real-world problems is about the accuracy of prediction (regression or classification), careful error analysis and detection of outliers in errors carry useful information. In recent work(9) we carried out precisely this at the interface between transcriptomic and proteomic level measurements of gene expression. We showed that when one regresses protein expression levels on transcript levels and several proxies for translation rates (mostly derived from sequence, e.g. codon adaptation index), one obtains a higher level of accuracy in predicting protein levels than what is explained by mRNA levels alone. We further showed, by a machine learning formulation designed to extract outliers, that large errors are overrepresented in proteins that are post-translationally regulated. In this part of the analysis, we will apply the above technique to identify putative post-translational regulation of the relevant proteins and their stratification with respect to treatment and viral loads.

2.6.4 Nonlinear Models:

The relationship between gene expression levels observed at the mRNA and protein levels is likely to be nonlinear due to the various different forms of regulations observable in biology (post transcriptional, post translational etc.). As such, it would be plausible to hypothesize that simple techniques such as linear regression and logistic classifiers may not be able to capture these relationships. The use of artificial neural networks is motivated by this observation and has proved successful in an impressive range of applications. However, there is an underlying compromise to be struck: that between power of a computational model and the available data (scarce and noisy). In this part we will establish if nonlinear models based on ANNS are applicable for modelling transcript → protein relationships and the specific drug response of interest. Our initial expectation, building on (9) is that there isn't much to be gained over and above simple (parsimonious) models, but this is to be empirically verified in this dataset.

2.6.5 Using natural language processing to address the issue of the small sample size

Recent inspirations from natural language processing, adopted to biological data analysis, suggest how machine learning models trained on large datasets could be adopted to solve problems that are characterised by small amounts of data. This is because the statistical and

functional relationships we wish to extract via the modelling exercise have similarities to what is acquirable from prior knowledge of other archived data. For example, Bepler & Berger (2021) and Rives et al. (2021), show how models trained on very large protein sequence data capture evolutionary and structural constraints of biology and can be used in subsequent benchmark downstream tasks such as predicting homologous relationships and secondary structures, outperforming models that are trained on the available small datasets alone(10,11). Here, we expect to build on these developments and use existing multi-omics data archived in public repositories to derive usable prior knowledge in models which could help alleviate the issues arising from the small amounts of data we have. Additionally, we will use leave-one-out cross validation, whereby models are trained repeatedly, leaving only a single datum out, and quantifying any inference we draw as averages over the left-out data. This way, statistically meaningful confidence can be derived on any inference we make -- the best one could do when in a limited data setting

References:

1. Wilkinson T. ACCORD: A Multicentre, Seamless, Phase 2 Adaptive Randomisation Platform Study to Assess the Efficacy and Safety of Multiple Candidate Agents for the Treatment of COVID-19 in Hospitalised Patients: A structured summary of a study protocol for a randomised. (2020). *Trials*. 21(1). doi:10.1186/s13063-020-04584-9

2. Ainsworth M. Performance characteristics of five immunoassays for SARS-CoV-2: a head-to-head benchmark comparison. (2020). *Lancet Infect Dis*. 20(12):1390–400. doi:10.1016/S1473-3099(20)30634-4

3. Love MI. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. (2014). *Genome Biol*. 15(12):550. doi:10.1186/s13059-014-0550-8

4. Langfelder P. WGCNA: An R package for weighted correlation network analysis. (2008). *BMC Bioinformatics*. 9(1):559. doi:10.1186/1471-2105-9-559

5. Shannon P. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. (2003). *Genome Res*. 13(11):2498–504. doi:10.1101/gr.1239303

6. Hänzelmann S. GSVA: Gene set variation analysis for microarray and RNA-Seq data. (2013). *BMC Bioinformatics*. 14(1):1–15. doi:10.1186/1471-2105-14-7/FIGURES/7

7. Shetta O. Robust subspace methods for outlier detection in genomic data circumvents the curse of dimensionality. (2020). *R Soc open Sci*. 7(2):190714. doi:10.1098/RSOS.190714

8. Shetta O. Convex Multi-View Clustering Via Robust Low Rank Approximation with Application to Multi-Omic Data. (2021). *IEEE/ACM Trans Comput Biol Bioinforma*. doi:10.1109/TCBB.2021.3122961

9. Gunawardana Y. Outlier detection at the transcriptome-proteome interface. (2015). *Bioinformatics*. 31(15):2530–6. doi:10.1093/BIOINFORMATICS/BTV182

10. Bepler T. Learning the Protein Language: Evolution, Structure and Function. (2021). *Cell Syst*. 12(6):654. doi:10.1016/J.CELS.2021.05.017

11. Rives A. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. (2021). *Proc Natl Acad Sci U S A*. 118(15). doi:10.1073/PNAS.2016239118/-/DCSUPPLEMENTAL

Protocol change log

Date	Version number	Amendment
17/04/2023	1	Original created
21/04/2023	1.1	Added funder information