PERSONALISING DOSING OF BIOLOGIC THERAPIES IN INFLAMMATORY ARTHRITIS TO MAXIMISE COST-BENEFIT

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Biology, Medicine and Health

2022

Dr Stephanie F Ling

School of Biological Sciences

Division of Musculoskeletal and Dermatological Sciences

Contents

Word count: 63,948

List of tables

List of figures

List of boxes

List of abbreviations

2C-DAS28	Two-component Disease Activity Score of 28 Joints
2-DE	Two-dimensional gel electrophoresis
-2LL	Minus two times the logarithm of the likelihood
ΔDAS28	Change in DAS28 after treatment
σ	Error model standard deviation
ω	Residual estimate standard deviation
A1AG1	α-1-acid glycoprotein 1, also known as ORM1
A1AT	α-1 antitypsin
A2AP	α-2-antiplasmin
ACPA	Anti-citrullinated protein antibody
ACR	American College of Rheumatology
ADAb	Anti-drug antibody
ADIPOQ	Adiponectin
Adj	Adjusted
ADME	Absorption, distribution, metabolism, excretion
AFM	Afamin
AGT	Angiotensinogen
AHSG	Alpha-2-HS-glycoprotein, also known as fetuin-A
AIC	Akaike information criterion
AKT1	RAC-α serine/threonine-protein kinase
ALDOA	Fructose-bisphosphate aldolase A
ALPL	Alkaline phosphatase, tissue-nonspecific isozyme
AMBP	Protein ABMP
AMPN	Aminopeptidase N
Anti-CarP	Anti-carbamylated protein
Ap4A	Diadenosine tetraphosphate
APO	Apolipoprotein
AS	Ankylosing spondylitis
ASPH	Aspartyl/asparaginyl β-hydroxylase
AUC	Area under the curve
AUC_{50}	Cumulative AUC level that produces half of E_{max}
B3AT	Band 3 anion transport protein
BAV	Bicuspid aortic valve
BIRC2	Baculoviral IAP repeat-containing protein 2
bDMARD	Biologic disease-modifying anti-rheumatic drug
BLVRB	Flavin reductase (NADPH)
BMI	Body mass index
BRAGGSS	Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate
BRAGGSS-PD	BRAGGSS Personalised Dosing sub-study
BSRBR-RA	British Society for Rheumatology Biologics Register for Rheumatoid Arthritis
BSV	Between-subject variability
C	Complement component
C1QC	Complement C1q subcomponent subunit C
C4BP	C4b-binding protein
C11orf54	Ester hydrolase C11orf54
CAIA	Collagen antibody-induced arthritis
CALD1	Caldesmon
CALM2	Calmodulin-2
CAMK1	Calcium/calmodulin-dependent protein kinase type 1

CAND1	Cullin-associated NEDD8-dissociated protein 1
CANX	Calnexin
CASP10	Caspase-10
CCL	C-C motif chemokine
CD	Cluster of differentiation
CDR	Complementarity-determining region
CI	Confidence interval
CE	Capillary electrophoresis
CEMIP	Cell migration-inducing and hyaluronan-binding protein
CF	Complement factor
CFHR	Complement factor H-related protein
CfMR	Centre for Musculoskeletal Research
CHI3L1	Chitinase-3-like protein 1, also known as YKL-40
CHIP	E3 ubiquitin-protein ligase CHIP
CL	Drug clearance
CLTC	Clathrin heavy chain 1
CLU	Clusterin
C_{max}	Maximum plasma concentration of a drug
COL6A	Collagen α-(VI) chain
COTL1	Coactosin-like protein
COVID-19	2019 coronavirus disease
CP	Caeruloplasmin
CPN2	Carboxypeptidase N subunit 2
CPSF6	Cleavage and polyadenylation specificity factor subunit 6
CR1	Complement receptor type 1
CRF	Case report form
CRP	C-reactive protein
csDMARD	Conventional synthetic disease-modifying anti-rheumatic drug
CTS	Cathepsin
CV	Coefficient of variation
DAS	Disease Activity Score
DAS_0	Baseline DAS28
DAS28	Disease Activity Score of 28 Joint Counts
DCP2	m7GpppN-mRNA hydrolase
DDA	Data-dependent acquisition
DDX6	Probable ATP-dependent RNA helicase DDX6
DIA	Data-independent acquisition
DNA	Deoxyribonucleic acid
DNAH3	Dynein axonemal heavy chain 3
DNAJB1	DnaJ homologue subfamily B member 1
DSC3	Desmocollin-3
DSTN	Destrin
DV	Observed concentrations
DYNC1CI1	Cytoplasmic dynein 1 light intermediate chain 1
EGF	Epidermal growth factor
EHD1	EH domain-containing protein 1
ELANE	Neutrophil elastase
ELISA	Enzyme-linked immunosorbent assay
E_{max}	Maximum drug effect
ESI	Electrospray ionisation
ESR	Erythrocyte sedimentation rate
EQ-5D	EuroQol Five-Dimension Scale
eQTL	Expression quantitative trait locus/loci

EULAR	European League Against Rheumatism
EZR	Ezrin
F12	Coagulation factor XII
FcγR	Fc-gamma receptor
FcRn	Neonatal Fc receptor
FDA	United States Food and Drug Administration
FDR	False discovery rate
FETUB	Fetuin-B
FGA	Fibrinogen α chain
FH	Fumarate hydratase, mitochondrial
FLII	Protein flightless-1 homologue
FLNB	Filamin-B
FLS	Fibroblast-like synoviocytes
FN1	Fibronectin
FOXO1	Forkhead box protein O1
FTL	Ferritin light chain
FUBP1	Far upstream element-binding protein 1
GARS	Glycine—tRNA ligase
GC	Gas chromatography
GLU2B	Glucosidase 2 subunit β
GM-CSF	Granulocyte-macrophage colony-stimulating factor
GNPTG	N-acetylglucosamine-1-phosphate transferase
GPS	Gray platelet syndrome
GRB7	Growth factor receptor-bound protein 7
GRP78	78 kDa glucose-regulated protein
GSN	Gelsolin
HADS	Hospital Anxiety and Depression Scale
HAPLN1	Hyaluronan and proteoglycan link protein 1
HAQ	Health Assessment Questionnaire
HC	Healthy control
HDGF	Hepatoma-derived growth factor
HNRNPA1	Heterogeneous nuclear ribonuclear A1-like 1
HPR	Haptoglobin-related protein
HPX	Haemopexin
HRG	Histidine-rich glycoprotein
HSP	Heat shock protein
HSPG2	Basement membrane-specific heparan sulfate proteoglycan core protein
IA	Inflammatory arthritis
IBD	Inflammatory bowel disease
IC	Immune complex
IC_{50}	Drug concentration leading to 50% decrease of baseline outcome measure
ICF	Informed consent form
ID	Identifier
IEF	Isoelectric focusing
IFN	Interferon
Ig	Immunoglobulin
IGF1	Insulin-like growth factor I
IGFBP2	Insulin-like growth factor binding protein 2
IKKA	Inhibitor of NF-κB kinase subunit α, also known as CHUK
IL	Interleukin
IL1-RA	Iinterleukin-1 receptor antagonist
IL1RAP	Interleukin-1 receptor accessory protein
ILF3	Interleukin enhancer-binding factor 3

IM	Intramuscular
IPQ	Illness Perception Questionnaire
IPRED	Individual-predicted measurements
IQR	Interquartile range
ITIH1	Inter-α-trypsin inhibitor heavy chain H1
iTRAQ	Isobaric tag for relative and absolute quantitation
IV	Intravenous
IWRES	Individual weighted residual distribution
JIA	Juvenile idiopathic arthritis
ka	Elimination rate constant of a one compartment model
k_{in}	Rate constant for synthesis
k_{out}	Rate constant for degradation
KEGG	Kyoto Encyclopedia of Genes and Genomes
KRT	Keratin
LASP1	LIM and SH3 domain protein 1
LBP	Lipopolysaccharide-binding protein
LC	Liquid chromatography
LCN2	Neutrophil gelatinase-associated lipocalin
LD	Linkage disequilibrium
LDHB	L-lactate dehydrogenase B chain
LMNB1	Lamin-B1
Lp(a)	Lipoprotein a
LRRFIP1	Leucine-rich repeat flightless-interacting protein 1
LRT	Likelihood ratio test
LTBP1	Latent-transforming growth factor β-binding protein 1
LYZ	Lysozyme C
mAb	Monoclonal antibody
MALDI	Matrix-assisted laser desorption/ionisation
MAP2K3	Dual specificity mitogen-activated protein kinase kinase 3
MAPK14	Mitogen-activated protein kinase 14
MASP1	Mannan-binding lectin serine protease 1
MCI	Mild cognitive impairment
MCID	Minimally clinically important difference
MCP	Monocyte chemoattractant protein
MDH2	Malate dehydrogenase, mitochondrial
MDM2	E3 ubiquitin-protein ligase Mdm2
MICE	Multiple imputation by chained equations
MIF	Macrophage migration inhibitory factor
MINPP	Multiple inositol polyphosphate phosphatase 1
MMP	Matrix metalloproteinase
MMCE	Mean misclassification error
MPO	Myeloperoxidase
MRI	Magnetic resonance imaging
MRM	Multiple reaction monitoring
MRPL19	39S ribosomal protein L19, mitochondrial
MS	Mass spectrometry
MSN	Moeisin
MYD88	Myeloid differentiation primary response protein MyD88
MYLK	Myosin light chain kinase, smooth muscle
m/z	Mass-to-charge
NAA25	N-α-acetyltransferase 25, NatB auxiliary subunit
NAD	Nicotinamide adenine dinucleotide
NAMPT	Nicotinamide phosphoribosyltransferase

NAXE	NAD(P)H-hydrate epimerase
NF	Nuclear factor
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
NK	Natural killer
NOAR	Norfolk Arthritis Register
NPDE	Normalised prediction distribution error
OA	Osteoarthritis
ODE	Ordinary differential equations
OGN	Mimecan
OmpT	Outer membrane protease T
OR	Odds ratio
ORM	Orosomucoid
PADI	Protein-arginine deiminase
PAFA	Platelet-activating factor acetylhydrolase
PAGE	Polyacrylamide gel electrophoresis
PARK7	Parkinson disease protein 7
PBMC	Peripheral blood mononuclear cell
PBPK	Physiologically-based pharmacokinetics
PCA	Principal component analysis
PCR	Polymerase chain reaction
PD	Pharmacodynamics
PDIA6	Protein disulfide-isomerase A6
PEA	Proximity Extension Assay
PF4	Platelet factor 4
PFKP	ATP-dependent 6-phosphofructokinase, platelet type
PHIP	PH-interacting protein
PIK3CD	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit δ isoform
PIS	Patient information sheet
PK	Pharmacokinetic
PKM	Pyruvate kinase PKM
PKP3	Plakophilin-3
PLG	Plasminogen
PLTP	Phospholipid transfer protein
PON1	Serum paraoxonase/arylesterase 1
popPK	Population pharmacokinetic
PPIA	Peptidyl-prolyl cis-trans isomerase A
PPV	Positive predictive value
pQTL	Protein quantitative trait locus/loci
PRDX3	Thioredoxin-dependent peroxide reductase, mitochondrial
PRED	Population-predicted measurements
PRO	Vitamin K-dependent protein
PRS	Polygenic risk score
PRS6A	26S proteasome regulatory subunit 6A, also known as PSMC3
PsA	Psoriatic arthritis
PSM	Peptide-spectrum match
PTGDS	Prostaglandin-H2 D-isomerase
PTHrP	Parathyroid hormone-related peptide
PTM	Post-translational modification
PTPN6	Tyrosine-protein phosphatase non-receptor type 6
PWRES	Population weighted residual distribution
PYGL	Glycogen phosphorylase, liver form

QC	Quality control
RA	Rheumatoid arthritis
RACK1	Receptor of activated protein C kinase 1
RANBP1	Ran-specific GTPase-activating protein
RBP4	Retinol-binding protein 4
RCT	Randomised clinical trial
RDX	Radixin
REC	Research ethics committee
RETN	Resistin
RF	Rheumatoid factor
RIA	Radioimmunoassay
RMSE	Root mean square error
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
RPN2	Ribophorin II
RRBP1	Ribosome-binding protein 1
RSU1	Ras suppressor protein 1
RUV	Residual unexplained variability
RSE	Relative standard error
S100A8/9	Calprotectin
S100A8	Protein S100-A8
S100A9	Protein S100-A9
S100A12	Protein S100-A12
SAA	Serum amyloid A protein
SAEM	Stochastic approximation expectation maximisation
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SBDC	Stoller Biomarker Discovery Centre
SC	Subcutaneous
sCD	Soluble cluster of differentiation protein
SCFD1	Sec1 family domain-containing protein 1
SCX	Strong cation exchange
SD	Standard deviation
SDS	Sodium dodecyl sulfate
SE	Standard error
SEPP1	Selenoprotein P
SF-36	Medical Outcomes Survey 36-item Short Form
SHH	Sonic hedgehog protein
SLC4A1	Band 3 anion transport protein
SLE	Systemic lupus erythematosus
SMOTE	Synthetic minority oversampling technique
SNP	Single-nucleotide polymorphism
SOD	Superoxide dismutase
SpA	Spondyloarthritis
SRM	Selected reaction monitoring
SSc	Systemic sclerosis
STUB1	E3 ubiquitin-protein ligase CHIP, also known as CO7
SWATH-MS	Sequential window acquisition of all theoretical mass spectra
$t_{1/2}$	Half-life
TAGLN2	Transgelin-2
TBG	Thyroxine-binding globulin
TCPH	T-complex protein 1 subunit η, also known as CCT7
TFRC	Transferrin receptor protein 1
TGF	Transforming growth factor

TLR	Toll-like receptor
TMDD	Target-mediated drug disposition
TNC	Tenascin
TNF	Tumour necrosis factor
TNFi	Tumour necrosis factor inhibitor
TNFRSR1A	Tumour necrosis factor receptor superfamily member 1A
TOF	Time-of-flight
TPM3	Tropomyosin α-3 chain
TPP2	Tripeptidyl-peptidase 2
tsDMARD	Targeted small molecule disease-modifying anti-rheumatic drug
TTR	Transthyretin
UBE2E1	Ubiquitin-conjugating enzyme E2 E1
UGGT1	UDP-gluocose:glycoprotein glycosyltransferase 1
UK	United Kingdom
UniProt	Universal Protein Resource
USA	United States of America
V/V_D	Volume of distribution
VAS	Visual analogue scale
VCL	Vinculin
VDBP	Vitamin D binding protein
VGFR1	Vascular endothelial growth factor receptor 1
VIM	Vimentin
VPC	Visual predictive check
v/v	Volume per volume
VWF	von Willebrand factor
WHO	World Health Organisation
w/v	Weight per volume
XML	Extensible markup language
XRCC	X-ray repair cross-complementing protein
YWHAH	14-3-3 protein η
ZNF169	Zinc finger protein 169

Abstract

Background

The disease course of rheumatoid arthritis (RA) varies widely between patients, and various therapeutic options are available. However, none are universally effective, all have a risk of side-effects and currently, all are prescribed on a "trial and error" basis, based on escalating cost and not precision medicine targeted to patient endotype. Previous work has shown that circulating drug concentration levels of tumour necrosis factor inhibitors (TNFi, a class of drug used to treat RA and other autoimmune conditions) are associated with response to treatment. This thesis hypothesised that biological factors, such as protein expression, contribute to variability in circulating drug levels and treatment response to biologic agents in patients with RA.

Methods

A population pharmacokinetic (popPK) study was carried out in patients with RA starting either Amgevita or Benepali, which are biosimilar agents for the TNFi agents adalimumab and etanercept, respectively. Model parameter estimates from the popPK study were used to simulate altered dosing intervals of these drugs. Proteomics data was obtained on all patients in the popPK study, as well as an additional cohort of patients with RA starting etanercept, using Sequential Window Acquisition of All Theoretical Mass Spectra (SWATH-MS). Protein expression was regressed against RA clinical outcome measures to determine any associations between protein expression and treatment response. Protein expression was also analysed alongside paired genotype data to determine whether any protein quantitative trait loci (pQTLs) existed. Significant pQTLs were then used to construct a polygenic risk score (PRS) for treatment non-response.

Results

16 patients were recruited to the popPK study; PK parameters were successfully estimated and used to simulate the effect of altered dosing intervals. SWATH-MS was used to generate proteomics data in serum samples from 180 selected patients commencing etanercept recruited to the Biologics in RA Genetics and Genomics Study Syndicate, a prospective multi-centre UK-based observational cohort. Proteomics analysis identified 52 proteins associated with RA clinical outcome measures. A pQTL analysis was carried out using 147 patients from the etanercept sub-cohort. 104 pQTLs were identified, 14 of which overlapped with significant proteins from the regression analysis. A PRS was generated using significant pQTLs, but was not found to be statistically significantly predictive of poor treatment response.

Conclusions

The popPK study has provided proof-of-concept for future personalised dosing trials in patients with RA starting TNFi. This thesis has identified several proteins associated with RA clinical outcome measures that also have a genetic basis. Findings require external validation with replication studies in an independent cohort, but once confirmed, this could pave the way for future biomarker and/or drug target development.

Lay abstract

Rheumatoid arthritis (RA) is a long-term disease affecting the joints. It causes pain, stiffness and swelling. In severe cases, it can damage joints and, if not treated in a timely and effective manner, can lead to lasting disability.

Patients with active RA can be treated with drugs called tumour necrosis factor inhibitors (TNFi). These work well in up to 60 – 70% of patients. However, TNFi are expensive and incur significant cost to the National Health Service (NHS). The aim of this work is to better select patients more likely to respond to TNFi and avoid over-treating patients unlikely to respond.

Two types of TNFi drug are being studied in this thesis: adalimumab (also known as Amgevita) and etanercept (also known as Benepali). All patients receive the same dose: an injection under the skin every 14 or 7 days, respectively. Previous research has shown that after receiving these drugs, the amount detected in the blood varies amongst patients. Patients with lower levels of drug in their blood are less likely to respond to treatment.

The relationship between the level of drug found in each patient's blood and their response to treatment has not been thoroughly explored. It is possible that changing the frequency of injections could influence drug levels in the blood and hence, the likelihood of a patient responding to treatment.

As part of the work for this thesis, I measured levels of Amgevita and Benepali in the blood of patients first starting these drugs. I then used statistical methods to analyse the change in drug levels over time. Using this information, I developed a formula to predict how other patients will hypothetically respond to treatment using different injection frequencies of Amgevita and Benepali. In future, this will allow the frequency and dose of the drug given to patients to be tailored to treat their arthritis as quickly and effectively as possible, while also reducing the risk of side-effects.

I also investigated factors that might cause drug levels to vary between patients. I measured the amounts of naturally-occurring proteins in the blood of patients with RA. Proteins in the blood can bind drugs and reduce their effectiveness, and this could explain why different patients respond differently after being given the same dose of a drug. I used mathematical techniques to determine whether a link exists between any of these proteins and a patient's response to a drug.

I hope that as a result of this research, it will be possible to individualise the frequency that each patient receives their drug, but this will first need to be tested in clinical trials. It may also be possible to predict whether a patient is likely to respond to a specific drug, based on the levels of different proteins measured in their blood. If they are unlikely to respond, they can instead be better treated with an alternative.

TNFi are prescribed in many long-term conditions, not just RA. The findings of this thesis could allow patients with a range of conditions to be treated more effectively. By tailoring each patient's treatment, their time in hospital will be reduced and their outcomes improved. At the same time, the financial and time pressure on the NHS will be reduced, freeing up capacity that will benefit patients elsewhere in the healthcare system.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree of qualification of this or any other university or other institute of learning.

Copyright statement

(i) The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and she has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

(ii) Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

(iii) The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

(iv) Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University's policy on Presentation of Theses.

Dedication

For my parents

and

For Jack

Acknowledgements

Alison Ahmed
Mohammed Shafi Ahmed
Sarah Ashton
Anne Barton
Ivona Baricevic-Jones
James Bluett
Jon Boden
Laura Bromilow
Tom Dumont
Carlo Ferrazano
Ge Xiangyu
Bethany Geary
Oliver Gough
Li Zhaoxun
Tony Kanal
Dua Lipa
Jack Looman
Stephen McDonald
Ruairí McErlean
Amirah Binti Mohammad Ariff
Nisha Nair
Jo-Ann Nicholson
Kayode Ogungbenro
Rebecca Pad
Darren Plant
Vanessa Reid
Anthony Robinson
Rowena Seaton Kelly
Shi Chenfu
Mehreen Soomro
John Spiers
Michael Stadler
Gwen Stefani
Megan Sutcliffe
Georgia Swain
Melissa Tordoff
Rachel Wray
Yap Chuanfu
Adrian Young

The Author

I graduated from the University of Liverpool with MBChB (Commendation) in 2011. During my undergraduate medical studies, I undertook an intercalated degree and was awarded an MRes in Clinical Sciences in 2010. As part of this degree, I undertook a clinical study involving the recording and analysis of crackles in idiopathic pulmonary fibrosis, a laboratory project involving flow cytometry of neutrophils from systemic sclerosis patients, and an epidemiology project analysing factors associated with increased disease activity in patients with early inflammatory arthritis. After graduation, I was awarded an Academic Foundation Programme post at University Hospital Aintree in Liverpool, where I was able to continue my epidemiological work on an early inflammatory arthritis cohort during my early clinical training. I moved to Manchester in 2013, when I was awarded an NIHR Academic Clinical Fellowship in Rheumatology. During this Fellowship, I trained in genetic epidemiology techniques and I carried out analysis on an early inflammatory arthritis cohort involving genetic associations with sub-components of disease activity scores. I obtained the MRCP (UK) postgraduate examination in 2015. In 2018, I was awarded an NIHR Manchester Biomedical Research Centre Clinical Research Fellowship in order to undertake this PhD. I obtained the MRCP (Rheumatology) Specialty Certificate Examination in 2021. I returned to full-time clinical practice at the beginning of January 2022.

Publications arising during the course of this PhD

Jude EB, **Ling SF**, Allcock R, Yeap BXY, Pappachan JM. Vitamin D deficiency is associated with higher hospitalisation risk from COVID-19: a retrospective case-control study. J Clin Endocrinol Metab. 2021. Doi: 10.1210/clinem/dgab439.

Maciejewski M, Sands C, Nair N, **Ling S**, Verstappen S, Hyrich K *et al*. Prediction of response of methotrexate in patients with rheumatoid arthritis using serum lipidomics. Sci Rep. 2021; 11(1): 7266. Doi: 10.1038/s41598-021-86729-7.

Ling SF, Broad E, Murphy R, Pappachan JM, Pardesi-Newton S, Kong M-F, Jude EB. High-dose cholecalciferol booster therapy is associated with a reduced risk of mortality in patients with COVID-19: a cross-sectional multi-centre observational study. Nutrients. 2020; 12(12): 3799. Doi: 10.3390/nu12123799.

Ling SF, Bluett J. Pharmacogenetics of methotrexate response in rheumatoid arthritis: an update. Pharmacogenomics. 2020; 21(1): 3-6. doi: 10.2217/pgs-2019-0154.

Ling SF, Nair N, Verstappen SMM, Barton A, Zucht HD, Budde P *et al*. Proteomic analysis to define predictors of treatment response to adalimumab or methotrexate in rheumatoid arthritis patients. Pharmacogenomics J. 2020; 20(3): 516-23. doi: 10.1038/s41397-019-0139-4.

McDonald S, Reed R, Baricevic-Jones I, BRAGGSS, **Ling S**, Plant D *et al*. Can serum interleukin-17 and interleukin-10 levels predict response to biologic treatments in patients with rheumatoid arthritis? Rheumatology (Oxford). 2019; 58(1): 1872-3. doi: 10.1093/rheumatology/kez147.

Hensor EMA, MKeigue P, **Ling SF**, Colombo M, Barrett JH, Nam JL *et al*. Validity of a two-component imaging-derived disease activity score for improved assessment of synovitis in early rheumatoid arthritis. Rheumatology (Oxford). 2019; 58: 1400-1409. doi: 10.1093/rheumatology/kez049.

Corzo P, Pros A, Martinez-Llorens J, Molina L, **Ling SF**, Balcells E. Isolated DLCO/VA reduction in systemic sclerosis patients: a new patient subset? Clin Rheumatol. 2018 Dec; 37(12): 3365-3371. doi: 10.1007/s10067-018-4342-5.

CHAPTER ONE: INTRODUCTION

Summary of chapter contents:

Rheumatoid arthritis (RA) is a chronic autoimmune arthritis that also has multisystem involvement. Active disease of the joint is mediated by inflammation at the synovium; the aim of pharmacological therapy is to control this inflammation before joint damage and subsequent irreversible disability occurs[1]. There is a wide variability in disease course between individual patients with RA, and whilst various different therapeutic agents exist to treat active inflammation, none are effective in all patients, all have risk of side-effects and currently, are all prescribed on a "trial and error" basis in clinical rheumatology practice[2]. Treatment is usually initiated with conventional synthetic disease-modifying anti-rheumatic drugs (csDMARDs), with patients who do not respond to these agents who have moderate-to-high disease activity being escalated onto more costly biologic medications (biologics, bDMARDs) or targeted synthetic medications (tsDMARDs)[3 4].

Biologics are prescribed in several chronic autoimmune conditions, as well as in RA, and prescription of these agents pose a significant cost burden on the National Health Service (NHS) in England[5]. Adalimumab and etanercept are tumour necrosis factor inhibitors (TNFi, a class of bDMARD) commonly used in the treatment of RA; they are the first and third agents, respectively, costing the most to the NHS in England, albeit for all indications, not just RA. In 2017/8, the total cost of adalimumab to the NHS in England was £494.5m and the total cost of etanercept was £219.8m[5]. By the following financial year 2018/19, £109.7m and £36.1m had been saved by switching to biosimilar versions of these respective

medications, but these agents still represent a significant cost to the NHS[6]. Biosimilar medicines are biological agents with similar properties to an originator medicine which has come off patent; these properties include biological activity, efficacy, safety and quality[7]. Biosimilars tend to be at a lower cost than the originator drug, and prescription of biologic agents is anticipated to increase as reduced cost will make access to these drugs less prohibitive. In fact, guidance from the United Kingdom (UK)-based National Institute for Health and Care Excellence (NICE) has recently been updated to lower the disease activity threshold at which bDMARDs can be commenced in patients with RA[8], which reflects the lower cost of biosimilar versions of these agents.

Patients receive a standardised dosing regimen of biologics such as adalimumab and etanercept, yet in up to 40% of patients with RA, inflammation remains inadequately controlled, either due to primary inefficacy or loss of response[9 10]. There is large variability in circulating drug levels between patients, and it has previously been reported that drug levels correlate with subsequent clinical improvement[11]. Some of this effect is accounted for by the presence of neutralising anti-drug antibodies (ADAbs), which occur with monoclonal TNFis such as adalimumab. However, the presence of ADAbs does not entirely explain poor response to a drug or variation in drug levels observed. Furthermore, no neutralising monoclonal antibodies have yet been reported with etanercept, so other factors must contribute to variability in drug levels.

Multiple factors are likely to influence TNFi pharmacokinetic (PK) parameters, such as drug clearance (CL), including disease/treatment-related variables, heritable factors and potentially other demographic patient-specific variables. Improved knowledge of the contribution of additional factors underpinning drug level differences and disease activity will be required to better understand the dose-concentration-response relationship in patients with inflammatory conditions such as RA that are treated with adalimumab and etanercept. Ultimately, a better understanding of the variables determining drug concentration could lead to a greater personalisation of dosing to optimise efficacy.

All TNFi agents are administered at a set dose and a set interval between doses. Given that patients respond heterogeneously to TNFis, it could be hypothesised that different patients might require higher or lower doses to achieve therapeutic levels at initiation of drug. Analysis of drug levels over time in a set of real-world RA patients can be used to develop a population (pop)PK model, which will aid understanding of how patients respond to drugs

at initiation. This popPK model can be used to simulate alternative dosing intervals, which could eventually lead to a clinical trial of personalised dosing intervals in the future. Patients receiving their drug at wider time intervals have obvious cost-saving benefits for the NHS. In addition, patients who require narrower dosing intervals may also benefit, as they might be less likely to experience disease flares and further encounters with healthcare services prior to achieving disease control following the introduction of a bDMARD; this also holds the potential for NHS cost savings. Finally, higher drug levels of TNFi agents have been shown to increase the risk of infection in patients with RA[12], so personalised dosing intervals could potentially minimise exposure and reduce this risk.

This work centres on the exploration of factors underpinning drug CL and treatment response, as well as whether these factors can be integrated into popPK modelling and simulation of personalised dosing intervals. The future aim is to maximise both treatment response to bDMARD agents, as well as their cost-effectiveness.

1.1. Clinical features of RA

RA is a chronic autoinflammatory condition associated with autoantibodies to immunoglobulin (Ig) G (i.e. rheumatoid factor, RF) and citrullinated proteins (i.e. anti-citrullinated protein antibodies, ACPA)[13]. Both disease pathogenesis and manifestations are heterogeneous. The disease predominantly affects the joints, hence its nomenclature as an arthritis, but the condition also has systemic manifestations. In addition, a proportion of patients with RA will be seronegative for the above autoantibodies, introducing another element of heterogeneity to the patient population. Furthermore, previous studies have demonstrated that patients with similar clinical phenotypes of RA have varied infiltrates and cytokine/gene profile expression at the synovium[14-16]. This disease heterogeneity means that RA patients can be classified into different endotypes; a disease endotype represents a subtype of a condition that can be defined by distinct functional and/or physiological mechanisms. Karsdal *et al* proposed that different RA endotypes are likely to respond in different ways to therapies with varying modes of action, lending RA to a precision medicine approach[17]. However, current RA treatment is not targeted by endotype, which could account for varied treatment response in this patient cohort.

Diagnosis of the condition is led by the rheumatologist and no set diagnostic criteria exist; instead, classification criteria can inform clinical diagnosis and aid stratification of patients with similar characteristics for the purpose of clinical research. Classification criteria have

evolved over several years, moving from those defined by the American College of Rheumatology (ACR) in 1987[18] to the joint ACR/European League Against Rheumatism (EULAR) criteria 2010[19] in more recent years.

The cardinal clinical feature of RA is synovial inflammation causing joint swelling, usually accompanied by early morning joint stiffness and tenderness on palpation. The pattern of joints affected by RA is different to those of other inflammatory arthritides; these include: the metacarpophalangeal and proximal interphalangeal joints of the hands and feet, the wrists, the elbows, the shoulders, the knees and the hips[20]. Whilst these joints include those in the periphery, RA is striking in its avoidance of the distal interphalangeal joints and the axial skeleton, with the exception of the atlanto-axial joint in the cervical spine. RA is aggressive in its destruction of joints, with subsequent inflammatory breakdown of cartilage and damage to articular and periarticular bone if left untreated.

Extra-articular manifestations of RA occur alongside systemic and articular inflammatory response. Active disease within the joints and beyond is associated with an increased acute-phase response and raised inflammatory markers such as erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP)[20]. Systemic inflammation can manifest in the eyes, lungs, heart and other organs, as well as rheumatoid nodules, vasculitis and secondary Sjögren's syndrome. In addition, RA is associated with increased cardiovascular mortality and morbidity[21], as well as with interstitial lung disease[22]. The aetiology of extra-articular manifestations is unknown, although there is an association with high RF titres[23], smoking[24], early disability[24], age[25] and comorbidity[25]. The frequency of extra-articular manifestations of RA do appear to have decreased in incidence over time[26], and this could be related to modern practices of treating RA in a timely manner in order to control disease activity as quickly as possible. However, while extra-articular manifestations have declined, the incidence and prevalence of RA remain stable[26].

1.2. RA prognosis and the role of autoantibody measurement

Multiple concurrent processes are involved in RA pathogenesis, including T cell autoreactivity and formation of autoantibodies[27]. Two of these autoantibodies are the aforementioned RF and ACPA; both are almost ubiquitously tested in clinical practice as they aid both diagnosis and determining prognosis, and as such, testing of both form part of both European and British guidelines for management of RA[28 29]. When considered in the healthy population, ACPA was found to be measured as positive in 0.8% of a Dutch

population of 40,136 people[30]. RF has been estimated as more prevalent than ACPA, ranging from positive in 5% of healthy 50-year-olds to 10-25% of healthy 70-year-olds[31].

Seropositivity for both ACPA and/or RF is associated with a poorer prognosis in the long-term in patients with RA[32], and this fits in with the finding that increased RF titres are associated with extra-articular manifestations as well[23]. A primary care inception cohort of inflammatory arthritis patients recruited between 1990 and 1994 demonstrated that treatment was not as beneficial in ACPA-positive patients versus seronegative patients[33]. Furthermore, RF and ACPA seropositivity are also associated with reduced response to TNFi[34]. Given this association with worse long-term disease outcomes and more difficulty in achieving disease control, it is vital to treat seropositive RA patients as quickly as possible with rapid escalation of treatment to prevent adverse sequelae.

Whilst ACPA and RF are known to correlate with long-term outcomes in RA, conflicting evidence also exists in the literature. In a UK-based multi-centre study involving both early and established RA patients, ACPA seropositivity was found to be associated with improved treatment response and with reduced odds of poor treatment response[35]. A subsequent larger study within the same cohort found that patients defined as having "poor prognosis" prior to commencing treatment on bDMARDs (defined as ACPA ± RF seropositivity in the presence of radiographic erosions) had lower disease activity after treatment[36]. One explanation for these findings could be that patients with known ACPA/RF seropositivity are treated more aggressively because of prior knowledge of these poor prognostic indicators.

Understandably, ACPA and RF are the most investigated autoantibodies in RA, and are even included in classification criteria to aid diagnosis. However, various other autoantibodies have been identified, and some even are associated with prognosis. For example, anti-carbamylated (anti-CarP) antibodies have been found to be associated with increased disease severity in RA patients seronegative for ACPA[37]. The search for predictors of prognosis and/or treatment response in RA patients is ongoing, and of particular interest would be in patients who are ACPA/RF seronegative. Many other autoantibodies exist, both citrullinated and non-citrullinated, yet their role in RA, its subtypes and correlation with treatment response have not yet been explored. Most published work has utilised the commercially-available CCP2 ELISA to determine ACPA seropositivity, but this excludes other citrullinated autoantibodies not detected by the assay. The first generation ACPA assay, CCP1, contained only citrullinated peptides derived from human filaggrin; the second

generation CCP2 assay also includes further epitopes that mimic true conformational epitopes, selected from pre-existing libraries generated of citrullinated peptides[38]. Therefore, by no means do either CCP1 or CCP2 assays cover the full spectrum of citrullinated peptides that could potentially be expressed in patients with RA. Third generation CCP3 assays provide further citrullinated peptide coverage still, but the use of these assays is currently not as widespread as CCP2.

1.3. Measures of treatment response in RA

As mentioned previously, the principles of RA treatment centre on controlling RA-associated inflammation to prevent joint damage and other sequelae of the disease. The aim is to achieve clinical remission or low disease activity; this section outlines some of the methods used to measure disease activity and hence, treatment response to RA medication. If a patient is deemed not to have responded to a drug, treatment is escalated from csDMARD(s) to bDMARD or tsDMARD in order to "treat to target," the target being a measure of RA disease activity that is below a pre-defined level. One such measure of disease activity that is commonly employed in the UK is the Disease Activity Score (DAS)[39].

Prior to the development and implementation of the DAS, initiation of medication for RA was largely based on individual clinicians' assessment of disease activity, and this was shown to have a wide variation between rheumatologists[40]. In an attempt to formalise disease activity assessment as well as provide a standardised tool to compare efficacy across clinical trials, van der Heijde *et al* developed the DAS from a prospective cohort of 113 patients, and this was based on the decisions made by each patient's treating physician on treatment escalation in order to achieve disease control[39]. Treatment escalation in this study from 1990 reflected available csDMARD therapeutic options at that time:

- Step one: hydroxychloroquine or sulfasalazine.
- Step two: intramuscular (IM) gold.
- Step three: D-penicillamine, azathioprine or methotrexate.

Prescribing practices have significantly evolved, particularly since the development and introduction of bDMARDs and tsDMARDs into common prescribing practices, yet treatment decisions continue to be made using the DAS. The DAS was developed as a multi-variable score as the authors attributed multiple variables to contribute to active RA, and the final variables selected were: the Ritchie articular index (a measure of joint tenderness)[41], 44 swollen joint count (indicating active synovial inflammation, synovitis), ESR (indicating

systemic inflammation) and patients' assessment of global health using a visual analogue scale (VAS), in order of descending importance.

The DAS has been further simplified to 28 joint counts (DAS28, both tender and swollen joints) for clinical utility, and was found to be as valid as the more extensive 44-joint DAS[42]. Subsequent definitions have substituted CRP for ESR[43] or have shown that VAS may be omitted[42], although different definitions cannot be used interchangeably[44-46].

However, the DAS has been shown to have limitations. DAS28 remission (DAS28 < 2.6) is the recommended primary outcome for trials of agents other than non-steroidal anti-inflammatory drugs (NSAIDs)[47], yet Cohen *et al* reported that a proportion of patients in sustained DAS remission studied over 5 years demonstrated evidence of radiographic progression, indicating ongoing disease activity[48]. Data from the Norfolk Arthritis Register (NOAR), a primary care inception cohort study, showed that HLA-DRB1 risk haplotypes for RA (defined by amino acids at positions 11, 71 and 74) were predictive of radiographic damage[49]. A valine amino acid at position 11 was found to be associated with the SJC and acute phase reactant components of the DAS28, but not tender joint count (TJC; VAS was not assessed in this study)[50]; these findings were validated in the independent Early RA Study cohort. Similarly, Baker *et al* reported that of the original DAS components, only swollen joint count (SJC) and acute-phase reactant (CRP/ESR), were independently associated with synovitis detected using magnetic resonance imaging (MRI)[51] i.e. only the objective components of the DAS were related to radiographic evidence of synovitis.

TJC is a subjective measure of disease activity as it is based on patient reporting during examination and can be falsely raised in conditions such as osteoarthritis and fibromyalgia (i.e. not caused by an underlying inflammatory condition), but it is given the strongest weighting in the composite DAS28. Hensor *et al* sought to overcome this subjective element by re-weighting the DAS28 to include only SJC and CRP; this was based on ultrasound synovitis data from multiple RA cohorts and was validated in the independent NOAR cohort[52]. This two-component DAS28 (2C-DAS28) showed superior association with both acute radiographic synovitis, as well as long-term radiographic damage, when compared to the conventional four-component DAS28.

Despite these limitations, the DAS28 remains the mainstay of clinical decisions regarding dose and treatment escalation. Other disease activity assessment scores have been developed

(such as the Simplified Disease Activity Index, SDAI[53], and the Clinical Disease Activity Index, CDAI[54]), but criteria determining treatment response, such as those developed by EULAR[55 56], utilise the DAS28.

The DAS28 provides a contemporaneous measure of a patient's disease activity at any one time, but response criteria have been developed to assess the effectiveness of medication over time. The EULAR response criteria[55 56] are summarised in Table 1.1. This measure was developed for use in medication trials to determine efficacy of the studied agent, and not for routine clinical decision-making. Another commonly used set of criteria for measuring treatment response in trials are the ACR improvement criteria[57].

Table 1.1. The EULAR response criteria using the DAS28 (adapted from Fransen and van Riel[58]).

DAS28 at endpoint	**Improvement in DAS28 from baseline**		
	≥1.2	>0.6 and ≤1.2	≤0.6
≤3.2	Good	Moderate	None
>3.2 and ≤5.1			
>5.1			

It is important to measure response to treatment in patients with RA to decide whether to escalate therapy or not, particularly given the high cost of many second-line and beyond agents. However, measurement of disease activity (and subsequently treatment response) is imperfect and may not reflect underlying inflammation and biological processes. Disease activity measurements also have an element of subjectivity, due to interpretation by both patient and assessor. Unfortunately, it is not practical to utilise more objective measurements of synovitis in the day-to-day clinic; time pressures preclude routine ultrasound scanning of all patients having a disease activity assessment, and magnetic resonance imaging (MRI) scanning patients regularly would be costly and time-consuming. While currently used disease activity measures are imperfect for correlation with active synovitis, they provide the best approximation of disease course and treatment response to the physician.

This is not the only imperfect facet of the treatment and management of RA: as mentioned in the previous section, the prediction of prognosis and treatment response has not been fully explored or optimised. Testing of ACPA and RF is well-documented and researched, but measurement does not result in all patients being streamlined onto the most appropriate

treatment. Up to 40% of patients starting bDMARDs for RA still have uncontrolled disease (due to both primary and secondary inefficacy)[9 10], so there is clearly a disconnect between known prognostic markers and personalisation of treatment. However, it should be noted that ACPA/RF seropositivity were not included as covariates in the trial conducted by Weinblatt *et al*[9], and only RF seropositivity (and not ACPA) was included as a covariate in the analysis of the study by Finckh *et al*[10], so serology may not have contributed to treatment non-response.

Given that prior knowledge accepts that RF, ACPA and anti-CarP antibody seropositivity is associated with different RA endotypes, it is reasonable to hypothesise that there could be other serological markers that could predict disease prognosis and/or treatment response to certain therapeutic agents. Along with more representative measurement of disease activity and hence, treatment response, a better understanding of the utility of precision medicine in RA also requires development.

1.4. Predictors of treatment response in RA

Just as prognostic markers have been identified in RA patients, so too have clinical and serological predictors of treatment response to bDMARDs.

1.4.1. Clinical predictors of treatment response

Clinical predictors of treatment response can include patient characteristics, behaviours and modifiable risk factors. Some of these modifiable risk factors have been demonstrated to be associated with treatment response. For example, cigarette smoking has been established as a risk factor for developing RA[59], and it has also been shown to affect response to bDMARDs. Results from a Swedish registry of bDMARD use in RA patients found that current smoking status was a negative predictive factor for achieving EULAR response (although it is not clear whether this includes good response, or both moderate and good response)[60]. The same study also found that current smoking status was associated with poorest drug survival in the cohort studied.

A large multi-centre observational UK-based study using the British Society for Rheumatology Biologics Register for RA (BSRBR-RA) examined RA patients refractory to bDMARDs, which the authors defined as switching to a third class of bDMARD[61]. Patient-related factors at baseline associated with disease refractory to bDMARDs included female gender, younger age, shorter disease duration, higher patient VAS, higher Health

Assessment Questionnaire (HAQ) score, current smoking status, obesity and worse social deprivation. Furthermore, another study from the BSRBR-RA found that symptoms of depression (as identified from patient reporting of a history of depression or baseline questionnaires, being either the Medical Outcomes Survey 36-item Short Form (SF-36) or the EuroQol five-dimension scale (EQ-5D)) with reduced odds of achieving good EULAR response after 12 months of treatment[62].

One mechanism for reduced treatment response that has been explored is adherence to medication. The World Health Organisation (WHO) defines adherence as: "The extent to which the patient's behaviour… [for example] taking medication… corresponds with agreed recommendations from a health-care provider"[63]. Hence, one could make the assumption that reduced adherence to drug could result in reduced treatment response. Bluett *et al* found that in a prospective cohort of 392 patients with RA on a mixture of bDMARDs, 27% self-reported non-adherence to medication; non-adherence was associated with worse treatment outcomes at 6 months[64]. Furthermore, in a systematic review of RA patients on methotrexate (a csDMARD), non-adherence to drug as defined through patient self-reporting was associated with reduced treatment response, as well as radiographic evidence of joint erosions[65].

On a similar note, Jani *et al* found that after 3 months of treatment with adalimumab, both ADAbs and low drug levels were significant predictors of poor EULAR response at 12 months[11]. In the same study, etanercept levels were not associated with EULAR response following adjustment for population covariates. A subsequent study of RA patients starting certolizumab (a bDMARD) found that drug levels were associated with achieving EULAR response at 12 months, and the presence of ADAbs was significantly associated with reduced drug levels over a 12-month follow-up period[66]. Following adjustment for confounders, ADAb levels and patient self-reporting of adherence to drug were associated with certolizumab drug levels.

1.4.2. Serological predictors of treatment response

Whilst ACPA and RF are known prognostic indicators in RA, they have also been reported to be predictive of treatment response in several studies. For example, seropositivity for RF and ACPA have been found to be associated with reduced response to TNFi drugs[34]. Conversely, RF seronegativity was associated with non-response to methotrexate in a cohort of bDMARD-naïve patients[67]. However, there is evidence that autoantibody reactivity exists

in seronegative RA patients as well. Vordenbäumen *et al* identified reactivity to three autoantibodies in two independent cohorts, being N-acetylglucosamine-1-phosphate transferase, gamma subunit (GNPTG), heterogeneous nuclear ribonucleoprotein A1-like 1 (HNRNPA1) and insulin-like growth factor binding protein 2 (IGFBP2)[68]; although these autoantibody reactivities were not compared to treatment outcome measures, this demonstrates proof-of-concept of the wide scope for autoantibody discovery beyond ACPA and RF.

As part of preliminary work to this thesis, in a cohort of 286 RA patients starting either adalimumab or methotrexate, a proportion of ACPA-negative patients were found to be seropositive for citrullinated autoantibodies that were not otherwise detected on a commercial CCP2 assay (Axis-Shield Diagnostics Ltd, Dundee, UK)[35]. This assay is used to determine whether patients are seropositive for ACPA. Autoantibodies to citrullinated cleavage and polyadenylation specificity factor subunit 6 (CPSF6) were associated with worsening DAS28 at three and six months after initiation of treatment; CPSF6 is involved in the maturation of pre-mRNA into functional mRNA[69]. Conversely, autoantibodies to citrullinated DnaJ homologue subfamily B member 1 (DNAJB1) were associated with improved DAS28 at three and six months; DNAJB1 is involved in the heat shock response and interacts with heat shock protein (HSP) 70[70]. Citrullination is a post-translational modification constituting the conversion of the amino acid arginine to citrulline, and the citrullinated forms of autoantibodies described above may represent finer specificities of ACPA outwith the scope of commercial assays, such as CCP2. These finer ACPA specificities appear to be associated with treatment response, and suggest that discovery studies may identify other antibodies or proteins that may predict response to certain therapeutic agents.

Proteomic analysis of samples can be carried out using a variety of detection and analysis techniques, which will be detailed in Section 1.5. Each strategy has its strengths and weaknesses and can be applied in a variety of situations, as appropriate. Rapid development of mass spectrometry (MS) technology in the last two decades has been instrumental in a plethora of protein studies that have been published, meaning that studies to identify protein biomarkers of treatment response in RA are still a developing field, given how recently some MS techniques have been developed and validated. Discovery proteomics studies for biomarkers of treatment response are attractive, given that proteins are stable and easily quantifiable, they can be measured in a variety of biological components (e.g. blood and its

components, synovial fluid, saliva, cerebrospinal fluid etc), and many established NHS laboratory techniques routinely use proteomics methods such as enzyme-linked immunosorbent assays (ELISAs) to guide clinical management. Additionally, proteins capture post-translational modifications (PTMs), giving an advantage over other methods, such as genetics or transcriptomics. The next section will address the relative advantages and disadvantages of various proteomics techniques, in order to then outline the state of research in biomarkers of treatment response in patients with RA.

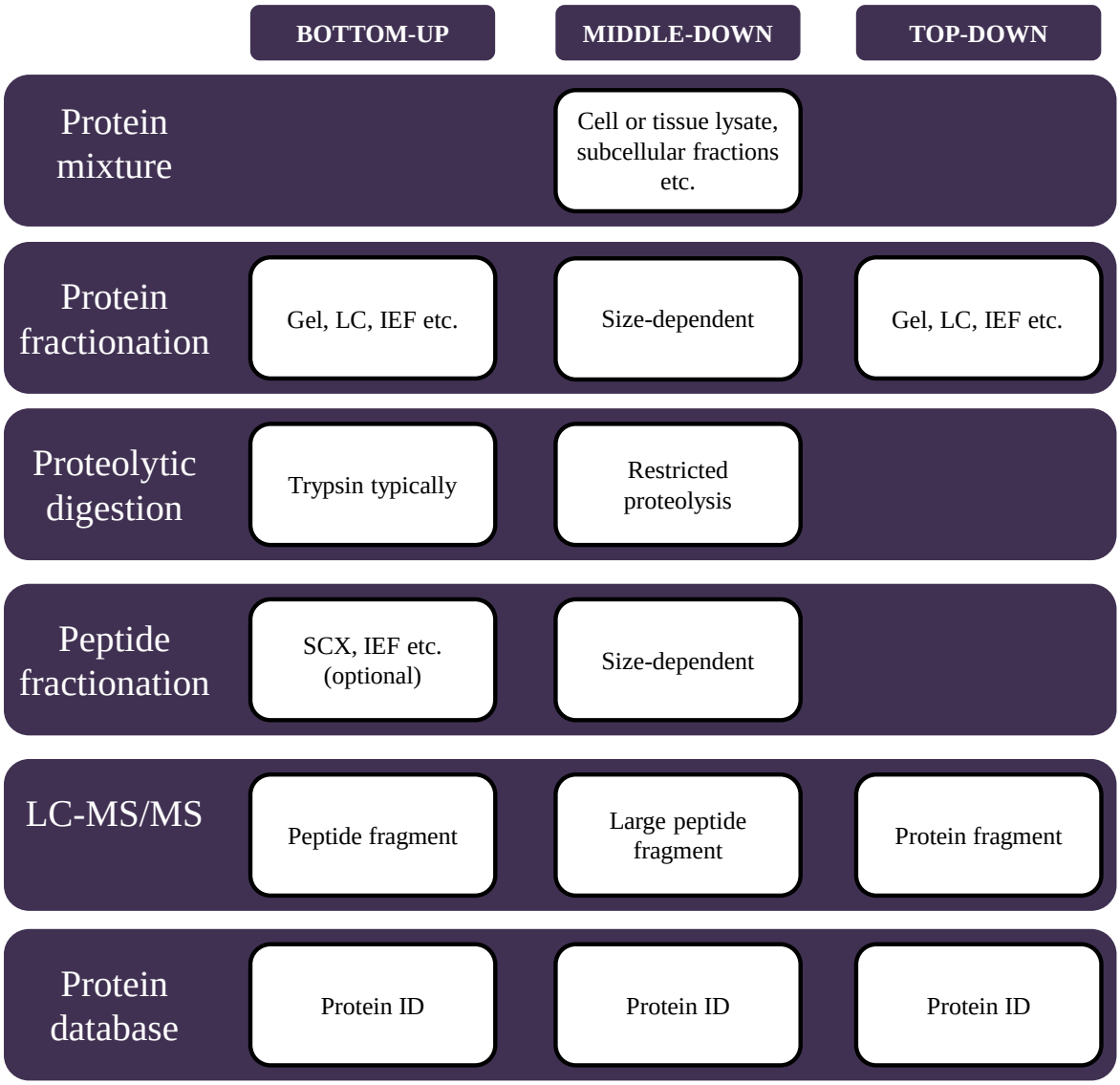
1.5. Proteomics as a discipline for biomarker discovery

The term "proteomics" was originally coined in 1997 as a portmanteau of the words "protein" and "genomics"[71]. The discipline of proteomics is concerned with identification of the components of the proteome and the function of each form of a protein. The field has developed exponentially due to advances in both laboratory techniques and technologies, as well as methods of bioinformatics data analysis[72].

Methods such as ELISAs and multiplexed immunoarrays (e.g. Luminex) are widely employed in proteomics studies. These methods rely on candidate protein studies with targets that have been pre-identified either through hypothesis-driven selection or a discovery study. Because only a limited number of proteins can be tested for during any one ELISA experiment and the focus of this PhD is on high-throughput discovery proteomics, only these latter techniques will be outlined here. However, the ELISA method has been briefly mentioned here because many shotgun proteomics studies subsequently use ELISAs and similar assays in the validation of the most associated proteins identified from MS. Further downstream, as previously alluded to, ELISAs may also play a role in measurement of biomarkers in patients prior to commencing a therapeutic agent, or early on in treatment, in order to predict future efficacy.

Strategies for large-scale proteomics analyses can be divided into "bottom-up," "top-down" and "middle-down" methods[72]. The bottom-up approach is concerned with lysed protein peptides, whereas intact proteins are measured in top-down proteomics. Middle-down methods are a combination of top-down and bottom-up methods: larger peptides are analysed than in bottom-up methods as a result of limited proteolysis or more selective proteases. These strategies are summarised in Figure 1.1.

Figure 1.1. Proteomics strategies: bottom-up versus middle-down versus top-down. Adapted from Zhang *et al*[72]



ABBREVIATIONS: Liquid chromatography (LC), isoelectric focusing (IEF), strong cation exchange (SCX). **LEGEND:** Bottom-up approaches analyse proteolytic peptide products. Middle-down approaches analyse larger peptides created by limited proteolysis or more selective proteases. Top-down approaches analyse intact proteins. One or more protein or peptide fractionation techniques may be utilised prior to MS analysis and database searching.

1.5.1. Top-down proteomics

Top-down proteomics is concerned with the analysis of whole, intact proteins, so it is suited to detecting PTMs and determining protein isoforms. PTMs are crucial in the function and antigenicity of proteins[73], and the three PTMs most commonly associated with RA are glycosylation, carbamylation and citrullination[74]. Citrullinated antigens have been found to

be associated with increased antibody reactivity in patients with RA[75], and as previously discussed, ACPAs are almost universally tested to confirm diagnosis and inform prognosis of RA[76].

Top-down proteomics has proved successful in quantifying proteins to over 200kDa in size[77]. For example, Tran *et al* were able to identify more than 3,000 protein species created by PTMs, ribonucleic acid (RNA) splicing and proteolysis using a four-dimensional separation system[78]. However, the top-down approach does have limitations in comparison with bottom-up methods, as the workflow for fractionation, ionisation and gas-phase fragmentation can be complicated and time-consuming, leading to a lower throughput than bottom-up methods[72]. These methods are essential for ensuring accurate quantification of proteins, but limit the capacity of sample analysis.

1.5.2. Multiplexed protein quantification techniques

A form of top-down proteomics exists that does not rely on MS technology, namely proprietary multiplexed protein quantification techniques. These techniques rely on binding to whole proteins and subsequent quantification via various technologies. Protein quantification can be as accurate as using an ELISA, and multiplexing means that sometimes thousands of known proteins can be measured at once in a single sample.

SomaLogic, a biotechnology company based in Boulder, Colorado in the United States of America (USA), has developed the SomaScan® Platform, a highly multiplexed platform for protein quantification, aimed at discovery proteomics[79]. To date, the current version of the SomaScan® Assay can measure around 7,000 human protein analytes. The technology is based on a new class of modified aptamer, Slow Off-rate Modified Aptamer (SOMAmer), which allows the development of high-affinity aptamers for most known protein targets[80]. In short, proteins in complex biological substances, such as plasma, are bound to SOMAmers via a series of reactions. Due to the unique nucleotide sequences assigned to each SOMAmer, subsequent binding with specific hybridisation probes leads to corresponding deoxyribonucleic acid (DNA) aptamer concentrations that can be very accurately quantified on a DNA microarray, with highly reproducible results. While not ubiquitous, uptake has been high, with over 300 peer-reviewed papers written using the SomaScan® Platform[81]. The platform remains proprietary, and sample processing is only conducted by SomaLogic in the USA.

Olink Proteomics, based in Uppsala, Sweden, also provide a range of proprietary multiplexed platforms for human protein biomarker discovery; Olink platforms have been utilised for almost 600 peer-reviewed publications[82]. Samples can be sent for processing in either Uppsala, Leeds, UK, or Boston, Massachusetts, USA. Up to 1,536 unique human proteins can be quantified, but more targeted panels of as few as 48 proteins can be processed. Olink platforms rely on Proximity Extension Assay (PEA) technology, where two paired antibodies bind to the target protein simultaneously, leading to hybridisation of the matching DNA oligonucleotides[83]. Once double-stranded DNA has been formed following antibody binding to the protein of interest, this can then by amplified via polymerase chain reaction (PCR) and then quantified. Because of this additional amplification step, the readout signal is strong, giving assay sensitivity similar to an ELISA. Olink platforms are comparable to SomaScan®, with good intra-sample reproducibility between assays carried out in a validation study of 4,998 healthy controls (HCs)[84].

While both SomaScan® and Olink platforms provide excellent and reproducible protein quantification, they remain costly and can only be carried out in a handful of proprietary laboratories. Furthermore, protein quantification is relative, as opposed to absolute, so comparisons can only be made between proteins quantified using the same proprietary platform. As such, MS likely represents a more accessible means of quantitative proteomics, as mass spectrometers are readily available for use in the majority of biomedical research institutions. However, it should be noted that MS also does not provide absolute protein quantification, although this could be achieved if robust internal standards are employed for quantification of specific proteins of interest. Furthermore, only the proteins specifically selected in each biomarker discovery panel are quantified, rather than all proteins in a sample, which theoretically could lead to missing measurements. Therefore, bottom-up proteomics methods, which will be discussed in Section 1.5.3, are still likely to represent a more comprehensive means of mapping the human proteome as part of discovery studies. A number of other proprietary multiplexed platforms exist, for example, based on bead technology (e.g. FLEX® [Millipore Corporation, Billerica, Massachusetts, USA], MagPlex™ microspheres, xMAP® [both Luminex Corporation, Austin, Texas, USA]), and these also have similar benefits and drawbacks as the SomaScan® and Olink platforms, but are not as widely used in proteomics studies.

1.5.3. Bottom-up proteomics

Bottom-up proteomics relies on the identification of the peptide products of proteolysis. Peptide fragments can be fractionated, ionised and fragmented more easily than whole, intact proteins, so mass spectra generated from peptide fragments are more practical and accurate to interpret. Ultimately, peptides give an indirect measurement of proteins in a sample, but this information can then be used to infer protein quantification.

When bottom-up methods are employed on a mixture of proteins (as opposed to purified samples), they are termed "shotgun proteomics," a term derived from shotgun DNA sequencing[85]. Typically, shotgun proteomics techniques begin with the enzymatic digestion of proteins in a mixture, followed by separation via either liquid chromatography (LC), gel electrophoresis or isoelectric focusing (IEF). Finally, identification of peptides is carried out via tandem MS, a process by which peptide ions undergo two or more stages of MS, separated either by time or space[86]. An optional additional step prior to MS is peptide fractionation e.g. using strong-cation exchange (SCX) or IEF. By comparing the tandem mass spectra with theoretical mass spectra from an *in silico* protein library, peptides can be identified and proteins inferred by the assignment of peptide sequences to proteins[72]. However, redundant and homologous peptide sequences may lead to misidentification, as in many instances, a set of peptides may represent multiple proteins (known as degenerate peptides), the so-called "protein inference problem"[87].

Stable isotope labelling can be employed alongside shotgun proteomics techniques[88]. Whilst labelling leads to more precise and accurate quantitation in comparison to label-free strategies, labelling procedures are costly and convoluted. Sample numbers are limited when isotope labelling techniques are used, and these techniques are not compatible with all experimental designs. Label-free methods are more commonly used; they are easy to apply with no sample limitations, although precision and accuracy are reduced in comparison to isotope labelling methods.

Shotgun proteomics is termed an "untargeted" strategy for proteomic analysis; "targeted" approaches (such as selected reaction monitoring, SRM, plural multiple reaction monitoring, MRM) rely on manual selection of target peptides and their corresponding transition ions[89]. In MRM, peptide fragments undergo electrospray ionisation during a first stage of MS (MS1), where selection of the intact analyte (parent ion) takes place[90]. The parent ion is then fragmented with gas ions during a second MS stage (MS2), where a specific fragment of the

parent ion is then selected. MS1 and 2 comprise an SRM assay. The detection of a parent (or precursor)-ion and product-ion pair is termed a "transition," and various methods aim to detect multiple transitions. This process relies on both pre-selection of specific peptides of interest for detection during the MS experiment, as well as knowledge of product-ion characteristics, and this can be carried out either *in silico* or from previous MS experiments. MRM is considered to possess superior quantification accuracy and dynamic range in comparison to conventional label-free methods[91], but throughput is substantially limited by the workflow because all transition ions must be individually queued for measurement. The optimum balance in an MS experiment consists of a long ion dwell time (i.e. the time spent acquiring each transition during MS cycling) with a short cycle time (i.e. the time spent acquiring data points across the LC peak). MRM's utility is when there are only a limited number of targeted proteins of interest, but MRM is considered the gold standard protein quantification method due its accuracy and enhanced dynamic range.

1.5.3.1. SWATH-MS

Sequential window acquisition of all theoretical mass spectra (SWATH-MS) is a technique developed by Gillet *et al* that provides an alternative data-independent strategy for interrogating proteomic samples[92]. SWATH-MS enables the detection and measurement of any protein of interest in a sample by comparing fragment ion spectral libraries to complete fragment ion maps obtained via data-independent acquisition (DIA). Conventional tandem MS techniques (i.e. conventional shotgun proteomics) rely on data-dependent acquisition (DDA), where an initial survey identifies specific peptide ions to take forward to a second phase of MS. However, with DIA, all ions within a pre-specified mass-to-charge (*m/z*) region are processed in the second step of MS[93].

SWATH-MS relies on the establishment of a spectral library before analysis commences so that targeted data extraction can be carried out. This is usually generated via DDA shotgun proteomics, and ideally, on the same machine used for SWATH acquisition, in order to obtain accurate peptide retention times[94]. Predicted fraction ion spectra can also be calculated using computational methods, although these are not preferred as they are thought not to match as well as spectra generated using DDA[95]. There are also online repositories of consensus spectral libraries with readily available data online, such as the SWATHAtlas database[96].

In the original Gillet experiment, a fast, high-resolution quadrupole-quadrupole time-of-flight (TOF) mass spectrometer was used to acquire data. 32 isolation windows of 25-Da were predefined, then samples were subjected to repeated cycling through each window during MS; these windows were defined as "swaths"[92]. In each swath, ionised peptides are fragmented systemically and in an unbiased fashion. In this initial experiment with SWATH-MS, every analyte in a single sample injection with a pre-determined composition was detected.

Other established shotgun proteomics techniques tend not to capture all peptide fragments in an experiment, due to a combination of more abundant peptides suppressing the signal of the less abundant, and due to the vast quantities of different peptides in any given sample. This demonstrates the utility of SWATH-MS, in that it creates a record of *all* peptides in a sample, which can then be re-analysed again and again *in silico*, for example, using a different peptide spectral reference library. In addition, extracted fragment ions in the Gillet study were specific enough to identify peptides over a dynamic range of four orders of magnitude, mitigating some of the issues of the protein inference problem[92].

In comparison to MRM, SWATH-MS techniques were shown to be comparable in terms of reproducibility and accuracy, but with superior proteome coverage and throughput[92]. Furthermore, SWATH-MS spectral maps constitute a permanent fragment ion spectral record for all precursors within specific acquisition settings for each sample i.e. mass, hydrophobicity. This enables re-examination of data sets *in silico* at a future date for any new proteins of interest following a first-pass biological review of the data. Huang *et al* have subsequently shown that SWATH-MS techniques can be applied on a whole-proteome scale[97]. Using a complex mouse cell lysate sample, 3,600 proteins were identified and quantified without sample pre-fractionation.

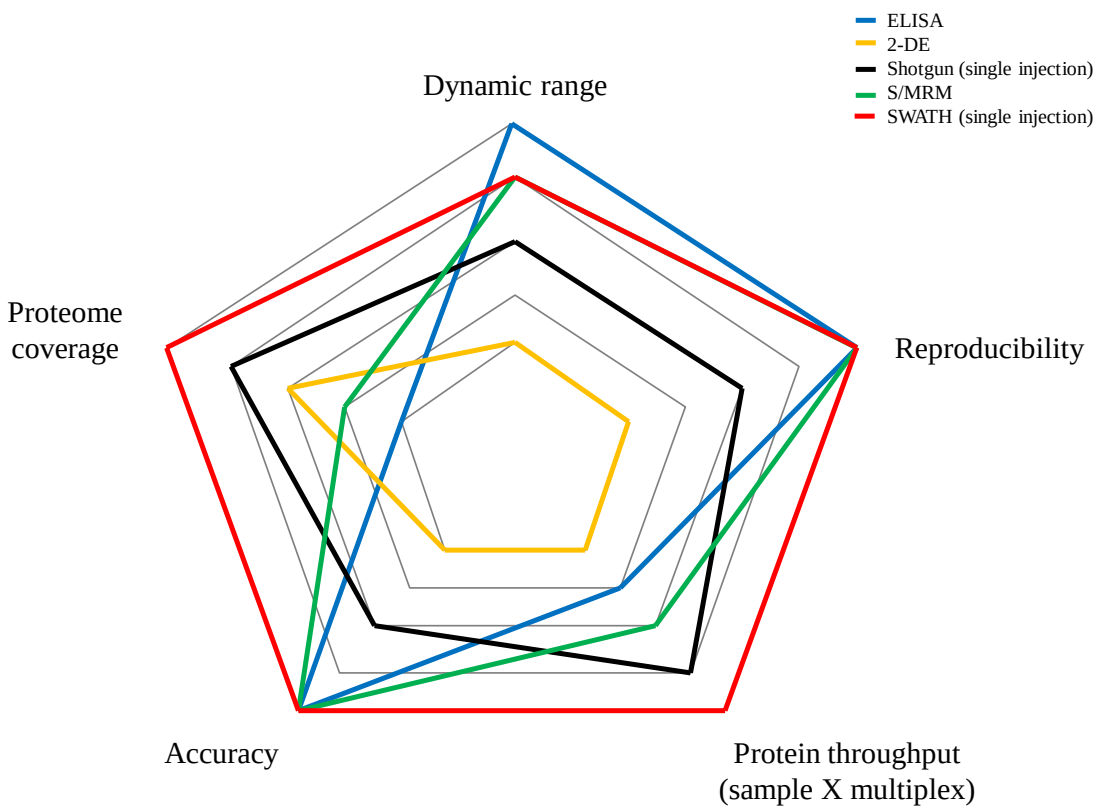
More recent studies have demonstrated the utility of SWATH-MS in analysis of clinical samples, and not just in animal and HC studies. For example, Harrison *et* al were able to use SWATH-MS data to identify biomarkers in plasma samples predictive of aortic diameter and future aortic aneurysm risk in patients with bicuspid aortic valve (BAV)[98]. In 8 BAV patients with aortic aneurysm and 7 BAV patients without, 4 plasma proteins were identified as potential biomarkers for monitoring maximum aortic diameter and 12 as potential biomarkers for future aneurysm risk in non-aneurysmal BAV patients. Despite the limited sample size, these 16 proteins were found to be significantly predictive of clinical outcomes,

meaning that a more focused validation study can be carried out in the future. Another study investigated the plasma proteome of patients with gray platelet syndrome (GPS), a haematological disorder that has been difficult to study due to its rarity. 51 proteins had altered expression in 11 patients with GPS compared with 13 HCs, and gene ontology analysis determined that these proteins had a pro-inflammatory and hepatic signature, which had not been hitherto seen[99].

However, given that proteomics techniques rely on the quantification of large numbers of peptides/proteins at any given time, these techniques can be very prone to missing data. A review of gel-based proteomics methods estimated missing values in between 10-50% of samples, with the proportion of peptides/proteins with at least one missing value ranging between 70-90%[100]. As with all proteomics techniques, SWATH-MS is not immune to this, but it does have a lower proportion of missing values compared with DDA methods[101]. However, this can be mitigated somewhat by statistical imputation of missing values, although this process has not been standardised within the field of proteomics[102].

SWATH-MS appears to be a good strategy for potential biomarker discovery, given its accuracy, dynamic range and ability for acquired data to be re-interrogated *in silico*. In addition, the technique can be used with multiple biological fluids, such as serum, plasma and synovial fluid. Furthermore, given that theoretically all peptides that are present in a biological sample are captured in SWATH proteome maps, this method represents excellent value for money, as sample processing costs are comparable to commercially-available multiplexed methods with more limited proteome coverage than SWATH (previously discussed in Section 1.5.2). SWATH's performance against other bottom-up proteomic techniques is summarised in Figure 1.2.

Figure 1.2. Technical performance of MS-based bottom-up proteomics methods, comparing proteome coverage, accuracy, dynamic range, protein throughput (sample X multiplex) and reproducibility. Adapted from Shao *et al*[103].



ABBREVIATIONS: Enzyme-linked immunosorbent assay (ELISA), selected/multiple reaction monitoring (S/MRM), sequential window acquisition of all theoretical mass spectra (SWATH-MS), two-dimensional gel electrophoresis (2-DE).

LEGEND: Axes represent the magnitude of each variable. All comparisons are qualitative.

1.5.4. Middle-down proteomics

As previously discussed, both top-down and bottom-up strategies of proteomic analysis have strengths and limitations. Wu *et al* proposed a hybrid approach based on peptides sized up to 20kDa, which they termed "middle-down" proteomics[104]. Middle-down proteomics relies on size-dependent protein fractionation using methods such as continuous tube-gel electrophoresis, coupled with restricted proteolysis using outer membrane protease T (OmpT) as a protease. The authors were able to identify 3,697 unique peptides from 1,038 high-mass HeLa proteins. Additionally, they were able to separate closely-related protein isoforms and detect numerous PTMs using this strategy. Middle-down approaches are likely to be utilised increasingly, given their ability to detect PTMs from larger fragments without the technical considerations of top-down analysis of whole proteins. However, middle-down

techniques are not currently in wider use as workflows are still being optimised and validated; in particular, techniques are still under development to digest proteins down to peptides to the optimal size of 3.5-10kDa[105].

1.5.5. Bioinformatics considerations and identification of proteins from peptide fragments

Given that proteomics analyses are increasingly generating more and more complex and comprehensive datasets, bioinformatics plays a crucial role in accurate and rational interpretation of studies. Once peptide spectra have been identified, they have traditionally been matched to sequences in a database using four basic approaches: descriptive, interpretative, stochastic and probability-based matching[106]. These are referred to as "spectrum-centric" analyses.

Descriptive models (e.g. using the SEQUEST program[107]) compare a mechanistic prediction of peptide fragmentation in a tandem mass spectrometer with an experimental mass spectrum; statistical techniques such as correlation analysis can be used to determine the degree of agreement. Interpretative models (e.g. using the Peptide Search program[108]) involve the use of either a manual or automated interpretation of a partial peptide sequence from a tandem mass spectrum for a database search. In stochastic modelling (e.g. using the SCOPE program[109]), basic probabilities of fragment ion matches are generated from training sets of spectra with known sequence identity; statistical limits are applied to the measurement and fragmentation process to determine the likelihood that the match is correct. The relationship between tandem mass spectra and peptide sequences are determined in statistical and probability models (e.g. using the Mascot program[110]), then the probability and significance of peptide identification can be extracted.

Peptide sequences from tandem mass spectra can be identified accurately by most database searching algorithms (such as the examples given above), but spectra of poorer quality, spectra containing inconsistent fragmentation processes or spectra with peptides of low-abundance proteins prove more demanding to analyse[106]. One strategy could be to remove poor-quality spectra and eliminate duplicates to reduce the volume of peptides analysed and create a pool of unique spectra[111-113]. It has been suggested that spectra should be searched with at least two different algorithms to account for differing selectivities (e.g. SEQUEST and Mascot). Unassigned spectra can be searched for modified amino acids, analysed by automated or manual sequence tagging, and automated or manual *de novo* analysis can

finally be used for any residual unassigned spectra[106]. Furthermore, several algorithms have been developed to assess peptide database searches, utilising both filtering and statistical methodology[114-117].

Methods have also been developed to solve the protein inference problem. Earlier algorithms relied on simple heuristics to assign proteins containing a specified number of known peptide sequences[116]. Ma *et al* were able to refine this process by deriving a minimum protein list from peptide sequences filtered to a specified false discovery rate; this highly discriminant protein filtration algorithm was thus more stringent in detecting false-positive proteins[118]. The method most commonly used is pseudo-probabilistic: degenerate peptides are divided amongst all corresponding proteins, then a minimum protein list that can account for all peptide assignments is generated using the expectation-maximisation algorithm[119]. More recently, in the last decade or so, others have developed algorithms to counter the protein inference problem using meticulous probabilistic modelling[120].

Since the more widespread uptake of high-throughput high-sensitivity techniques such as SWATH-MS, "peptide-centric" techniques of protein identification and quantification have been developed. These analyses test directly for the presence and absence of query peptides[121], and can be carried out using a wide range of software, for example, OpenSWATH, SWATH 2.0, Skyline and Spectronaut[122]. In addition, DIA-Umpire software combines both peptide- and spectrum- centric protein analysis in a hybrid approach[123]. All five of these software packages were compared in a benchmarking exercise, which demonstrated very similar protein identification and reproducible quantification, demonstrating that all the packages tested are equally reliable[122]. This further demonstrates the benefit of label-free data-independent techniques such as SWATH-MS, given the utility and availability of protein analysis software to be used in conjunction with MS data acquisition.

Bioinformatics considerations are less of an issue when using multiplexed methods, such as the SomaScan® and Olink platforms, as these have been incorporated during assay development.

1.6. Proteins associated with treatment response to biologic agents in patients with RA: findings from previous studies

Proteomic analysis of samples can be carried out using a variety of detection and analysis techniques, as outlined above. Each strategy has its strengths and weaknesses and can be applied in a variety of situations, as appropriate. The rapid development of MS technology in the last decade or so has been instrumental in the burst of protein studies that have been published. However, given the evolving nature of MS studies, research to identify protein biomarkers of treatment response in RA are still a developing field.

1.6.1. Proteomic studies of treatment response to infliximab

Infliximab was one of the first TNFi agents to be developed and approved for treatment of RA in the UK. It is administered every 8 weeks by intravenous infusion, but since the development of subcutaneous (SC) biologic agents that can be self-administered by patients in their own homes, it has been prescribed less frequently for RA in the UK. Nevertheless, insights into treatment response to this agent may be applicable to other members of the TNFi class that are in more popular usage. In addition, infliximab remains widely prescribed in the treatment of inflammatory bowel disease (IBD)[124 125].

In a study of RA patients commencing infliximab, Sekigawa *et al* used 2D LC-tandem MS to identify 21 proteins with increased expression in serum or plasma following infliximab treatment, and one protein (desmocollin-3, DSC3) with reduced expression[126]. Identified proteins were related to the tumour necrosis factor (TNF)-mediated pathway for nuclear factor (NF)-κB activation, as well as relating to articular cartilage metabolism and regeneration. However, this study only included 10 patients, and the comparison in protein levels was within-subject, before and after treatment. A similar study of serum from 33 patients with RA using isobaric tag for relative and absolute quantitation (iTRAQ) labelling and nano-LC-tandem MS identified 71 differentially expressed proteins in RA patients pre- and post-treatment with infliximab[127]. Leucine-rich α-2 glycoprotein (LRG) was identified as a biomarker of disease activity and was proposed for use in monitoring of treatment response. Again, these findings were within-subject, with no control group.

Another study identified apolipoprotein A-I (APOA1) and platelet factor 4 (PF4) as biomarkers for infliximab response, using a combination of sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE, a technique for peptide and protein separation) and surface enhanced laser desorption/ionisation (SELDI)-TOF MS (a technique of MS)[128].

This study had improved power, included 60 patients, and comparisons were made between two groups of responders versus non-responders to treatment, providing more robust findings than previous studies. These two proteins were significantly associated with treatment response to infliximab, and findings were confirmed following further purification and analysis of these proteins using SDS-PAGE.

1.6.2. Proteomic studies of treatment response to etanercept

A small number of proteomic studies of treatment response to etanercept in RA patients have been carried out. High serum levels of monocyte chemoattractant protein (MCP)-1 and epidermal growth factor (EGF) were associated with response to etanercept in a study of 33 French patients, in whom 12 cytokines were measured using a commercially available proteomic array[129]. Treatment response was defined within-subject, pre- and post-treatment. Hueber *et al* used a similar approach on RA patient serum with antigen microarrays, ELISAs and a multiplex bead assay (FLEX®)[130]. Patients starting etanercept were recruited from three cohorts based in the USA, Sweden and Japan and a signature consisting of 24 protein biomarkers were found to correctly predict both treatment response and non-response to etanercept in all three cohorts, again based on within-subject pre- and post-treatment findings. This was based on model development in a training cohort, followed by validation in a holdout cohort. Biomarkers were biologically feasible, as they had already been pre-selected due to their previously known roles in immunity and in RA. A later study using SDS-PAGE and nano-LC-MS identified two proteins in serum, vitamin K-dependent protein S (PROS) and E3 ubiquitin-protein ligase CHIP (CHIP), as predictive of treatment response to etanercept (within-subject) at 6 months in a cohort of 22 RA patients; this was validated in an independent cohort of 16 RA patients using targeted ELISA[131].

These studies have all been on a small scale and only one of the above studies used a validation cohort to replicate findings. No comparisons were made between RA patients and HC or other disease controls, so it is difficult to ascertain whether protein expression was specific to the RA disease state, or a physiological finding that could also be found in HC. Nonetheless, these could still represent biomarkers of therapeutic response.

1.6.3. Proteomic studies of treatment response to other biologic agents

There have been a limited number of proteomic studies on RA patients established on other bDMARD agents, in addition to infliximab and etanercept. Fabre *et al* sought to identify biomarkers predictive of treatment response to rituximab (a bDMARD that acts as an anti-

cluster of differentiation (CD) 20 antibody that causes B cell depletion) at pre-treatment using a commercially available protein biochip array (Investigator Evidence array, Randox, Montpellier, France) to test the sera of 46 patients with RA[132]. They were unable to find any baseline features predictive of response at 3 months within subjects, but responders and non-responders did have unique cytokine profiles at 90 days, which led the authors to suggest that these cytokine profiles might be useful in monitoring rituximab therapy contemporaneously, given that rituximab can take up to 16 weeks to show effect. In a study of patients with RA who had failed TNFi and been escalated to rituximab, Nguyen *et al* used nano-LC tandem MS to identify 43 proteins enriched in sera of responders to rituximab and 15 proteins enriched in sera of non-responders[133]. Following validation with ELISA, haptoglobin, calprotectin (S100A8/9), lipoprotein a (Lp(a)), C4b-binding protein (C4BP) and serum amyloid A protein (SAA) were found to have lower concentrations in treatment responders, whereas alpha-2-HS glycoprotein (AHSG, also known as fetuin-A) and thyroxine-binding globulin (TBG) had increased expression in responders.

Murota *et al* identified 33 proteins elevated in the serum of 28 patients with RA when compared to HCs using the SomaScan® assay, measuring 1,128 serum proteins[134]. Analysis focused on proteins associated with matrix metalloproteinase (MMP) 3, as this is known to have a role in degradation of cartilage in the joints of patients with RA[135]. Interleukin (IL)-16 was identified as being the MMP3-associated protein most correlated with treatment response. Validation with ELISA was carried out in a cohort of patients on a mixture of either methotrexate or of either of the biologics tocilizumab (a humanised monoclonal antibody against interleukin IL-6R), abatacept (a fusion protein that causes T cell inhibition) or infliximab, and decreasing IL-16 levels were found to be an effective clinical parameter for predicting treatment response. Patients receiving abatacept and tocilizumab were pooled together in analysis, although patients on methotrexate monotherapy and infliximab were analysed separately. Similarly, Tesitsma *et al* used a multiplexed immunoassay (xMAP®) to measure 85 proteins in the sera of patients with RA commencing either methotrexate or tocilizumab monotherapy or tocilizumab *plus* methotrexate as part of the U-Act-Early trial[136]. C-C motif chemokine (CCL) 18, CCL20 and soluble IL-2 receptor α were associated with achieving sustained drug-free remission in the tocilizumab + methotrexate arm, but there were no other significant protein associations in the other arm. No HCs were used to determine differential expression of proteins from patients with RA.

In a study of 250 Dutch RA patients on the TNFi agents adalimumab, etanercept, golimumab, infliximab and certolizumab, Cuppen *et al* utilised the xMAP® platform and were unable to identify any proteomic predictors of treatment response, although 12 proteins were differentially expressed in the discovery cohort of 65 patients[137]. This lack of a significant biomarker or biomarker panel may be due to a reduction in power caused by pooled analysis of patients on different TNFi treatments to one another in an already modestly-sized cohort. Furthermore, proteins with non-parametrically distributed expression were excluded from analysis, as a single partial least squares model was chosen, so this could have excluded potential biomarkers even prior to analysis in order to use data that fit with the chosen modelling technique; instead, a statistical method that can be used with non-parametric data or even normalisation of protein expression during pre-processing could have been used.

1.7. Multi-omics approaches to assessment of treatment response to bDMARDS in patients with RA

Current studies are now moving towards a multi-omic approach to addressing disease pathophysiology and assessing disease outcomes i.e. the integration of genetic, genomic, proteomic, metabolomic and lipidomic etc. data in order to achieve more granular understanding of a disease state. For example, Sun *et al* incorporated protein expression data and genetic data from 3,301 healthy blood donors to demonstrate protein quantitative trait loci (pQTL) overlapping with both gene expression quantitative trait loci (eQTL) and loci associated with specific diseases[84]. This study in a large number of healthy patients has paved the way for more integrative multi-omics studies in disease states, and such studies have started to populate the field in RA.

One such study in RA patients was carried out by Tasaki *et al*, and studied the effect of treatment with either methotrexate, infliximab or tocilizumab in 49 patients with RA, with a HC comparator group of 42 participants[138]. Using a combination of transcriptomic, proteomic (SOMAScan™) and immunophenotyping methods, the authors were able to demonstrate that treatment altered the molecular profile of RA cases versus HCs. Furthermore, molecular profiling was able to identify patients with a phenotype associated with long-term remission from RA. 255 serum proteins were identified that were associated with drug-naïve status in patients with RA, including up-regulation of CRP and complement component 3 (C3). However, despite RA patients being on three different medications, each with different modes of action, RA patients were pooled together heterogeneously for the

purposes of measurement and analysis. Furthermore, statistical modelling to derive an RA diagnostic model was carried out using a heterogeneous set of patients with modest numbers, namely 45 patients with RA, 30 patients with primary Sjögren's syndrome and 35 HCs.

Similarly, Farutin *et al* integrated whole-blood mRNA, plasma protein, glycopeptide and cell type-specific measurements in order to define a molecular signature to define response to treatment response to MTX plus either of the TNFi adalimumab or infliximab[139]. Despite interrogation of multiple types of biological samples from the same 76 patients, the most significant findings were that innate immune cells were increased at baseline in treatment responders and adaptive immune cells were increased in non-responders. Again, patients on different treatments were pooled, although all treatment consisted of TNFi. No HCs were included.

More recently, multi-omic studies have also included analysis that applies machine learning algorithms in order to derive meaningful clinical applications from the complicated, large biological datasets generated. Mellors *et al* carried out a study integrating clinical data, gene expression data in whole blood and RA-associated single-nucleotide polymorphism (SNP) transcriptions in 143 RA patients starting either adalimumab, certolizumab, etanercept, golimumab or infliximab recruited from the CERTAIN trial[140]. Gene expression biomarkers of treatment non-response to etanercept, adalimumab and infliximab were first derived using a discovery cohort of 58 female RA patients from two study cohorts (Autoimmune Biomarkers Collaborative Network, ABCoN, and the Brigham and Women's Hospital RA Sequential Study, BRASS). A random forest algorithm was used to develop a model predictive of treatment response and then validated in a further 175 patients with a positive predictive value (PPV) of 89.7% and specificity of 86.8%. Of note, no other machine learning algorithms apart from random forest were trained and validated, so it is unclear whether optimum model performance was achieved. Members of the same group subsequently used 345 RA patients from the prospective CERTAIN study and 146 patients from the prospective NETWORK-004 study, both based in the USA, in order to also define a molecular signature of treatment response to various different TNFi: adalimumab, etanercept, infliximab, certolizumab and golimumab[141]. A random forest machine learning algorithm was used to rank protein-coding RNA transcripts, which were then mapped to known proteins from a previously defined RA interactome. It is unclear whether patients from the first study by Mellors *et al*[140] were included in this new analysis, and could lead to potential bias due to data leakage from reuse of the same patients.

Tao *et al* studied 80 RA patients starting adalimumab or etanercept, and carried out differential gene expression and methylation analyses between patients who responded to treatment and those who did not[142]. Profiling was carried out on peripheral blood mononuclear cells (PBMCs), monocytes and CD4+ T cells. Once transcription and epigenetic signatures of treatment response were defined, a random forest algorithm was used to train separate machine learning models for adalimumab and etanercept in order to define treatment response; no other machine learning algorithms apart from random forest were utilised and compared, so it is unclear if optimum models have truly been derived from the data available. However, models were externally validated with nine patients who originally did not achieve treatment response and were switched to the other agent, and models were able to predict treatment response to an accuracy of 77.8% in the epigenetic model and 88.9% in the transcriptomic model.

Luque-Tevar *et al* measured a proprietary panel of 27 cytokines known to be associated with a systemic inflammatory response (Bio-Plex), oxidative stress parameters, NET-osis-derived products and microRNA in 104 RA patients starting a combination of the TNFi infliximab, etanercept, adalimumab, golimumab and certolizumab, as well as in 29 HCs[143]. Amongst several analyses, three different regularised logistic regression machine learning algorithms were used to train models predictive of treatment response, and a combined model utilising both clinical and molecular predictors was shown to have a superior area under the curve (AUC) in receiver operating characteristic (ROC) curve analysis compared to clinical or molecular models alone. Again, treatment response was defined from heterogeneous treatment groups in both training and validation cohorts of patients. However, a mixed clinical and molecular model demonstrated an impressive AUC of 0.909, but this needs to be validated, preferably in a prospective clinical cohort, independent from the cohort from which model training and validation was carried out in.

These multi-faceted studies are starting to generate more in-depth understanding of the systemic interactome during the active RA disease state, but as yet, there is little intra-study reproducibility. This could likely be due to many factors, such as timing of blood sampling, sample processing, heterogeneity of RA treatment and deviations in machine learning analyses. Clearly, a multi-omic approach will lead to more comprehensive indicators of treatment response, but current studies are only in their infancy at present.

1.8. Summary of proteomics studies in RA

Bottom-up proteomics studies and multi-omics studies of treatment response in patients with RA are scarce and are largely small-scale, with fewer than 100 participants. Some of these studies have been carried out with expensive proprietary assay-based methods, such as SomaScan® and xMAP®, and may be difficult to reproduce. Furthermore, some findings have relied on pooled analysis of patients receiving treatment with different therapeutic agents, which may weaken associations or even generate spurious ones. Of the studies discussed, very few utilised HCs in order to first separate out proteins associated with the active RA disease state. There is still a huge amount of scope for future discovery studies in patients with RA to attempt to identify biomarkers of treatment response.

Progress has been slower than anticipated in predicting treatment response to therapeutic agents in RA using biomarkers, and this could be because most studies have tried to demonstrate correlation between biomarker levels and clinical outcome measures, including subjective measures, such as the DAS28 (which includes subjective patient-reported components). However, more progress has been made in identifying predictive biomarkers in other fields, where the outcome is also a biological measure, such as the genetic determinants of treatment response to warfarin loading[144]. Previously, the measurement of drug levels in the sera of patients with RA on bDMARDs has shown that drug levels within a known therapeutic range are associated with achieving treatment response[11]. However, circulating drug levels vary widely between patients with similar disease activity who are starting the same drugs, but it has not been fully ascertained what underpins this variability (some of this variability is explained by the pharmacokinetic properties of bDMARDs). Given that drug levels vary between individuals, as well as correlating with clinical outcomes, are objectively measured and are biological measures, then research to identify predictors of drug levels and explore causes of variability between individuals is a potentially useful area of research. Analysis to first generate a model of how drug levels vary on initiation of treatment, which could then also incorporate proteomic data, could be a powerful tool in understanding why some patients achieve therapeutic efficacy and why others do not.

1.9. Basic pharmacology of bDMARDs

Some of the variability in circulating bDMARD levels can be explained by an understanding of the basic pharmacology of these drugs. bDMARD agents are costly, yet demonstrate superior efficacy over csDMARDs alone[145]. These agents are more usually prescribed long-

term for several years in the treatment of chronic conditions and a series of escalating treatment options with combinations of csDMARDs must fail in patients before qualifying for bDMARDs or tsDMARDs in the UK[8 146-149]. It is important to note, however, that the majority of patients respond to csDMARDs and do not require escalation to bDMARD or tsDMARD therapy in the long term.

This next section will highlight some of the basic properties of bDMARD agents, in order to explain some of the variability in treatment response to these agents.

1.9.1. The structure of bDMARDs

Most TNFis (the class of bDMARD being studied in this thesis), bar etanercept, are monoclonal antibodies (mAbs). mAbs are antibodies specific to a single antigen that are produced by immune cells which are clones of an originator parent cell[150]. Of the bDMARDs licensed for use in RA, these drugs target TNF (i.e. the agents infliximab, etanercept, adalimumab, golimumab and certolizumab), T-cells (i.e. abatacept), CD20 (i.e. rituximab) and IL-6 (i.e. tocilizumab). Not all of the above listed agents are mAbs, but of those listed above (apart from certolizumab, which will be discussed later, and infliximab, which is chimeric mouse-human) have the structure of human Ig of the G isotype (IgG), and specifically, derive from IgG1[151]. IgG1 is a large molecular weight (150 kDa) hydrophilic protein, and consists of paired, identical heavy and light variable chains (Figure 1.3.) The Fab domain constitutes the antigen-binding site of the protein. The variable portions (or complementarity-determining regions, CDRs) establish the binding site of each IgG1 protein by forming a complementary structure to the target antigen; variations are caused by changes in the amino acid sequences in each CDR. The crystallisable Fc portion dictates each antibody's effector function, and Fc portions are different in each IgG subtype. The structure of certolizumab differs in that it consists of a Fab fragment conjugated to two polyethylene glycol (PEG) chains (Figure 1.4).

Figure 1.3. Simplified structure of an IgG1protein.

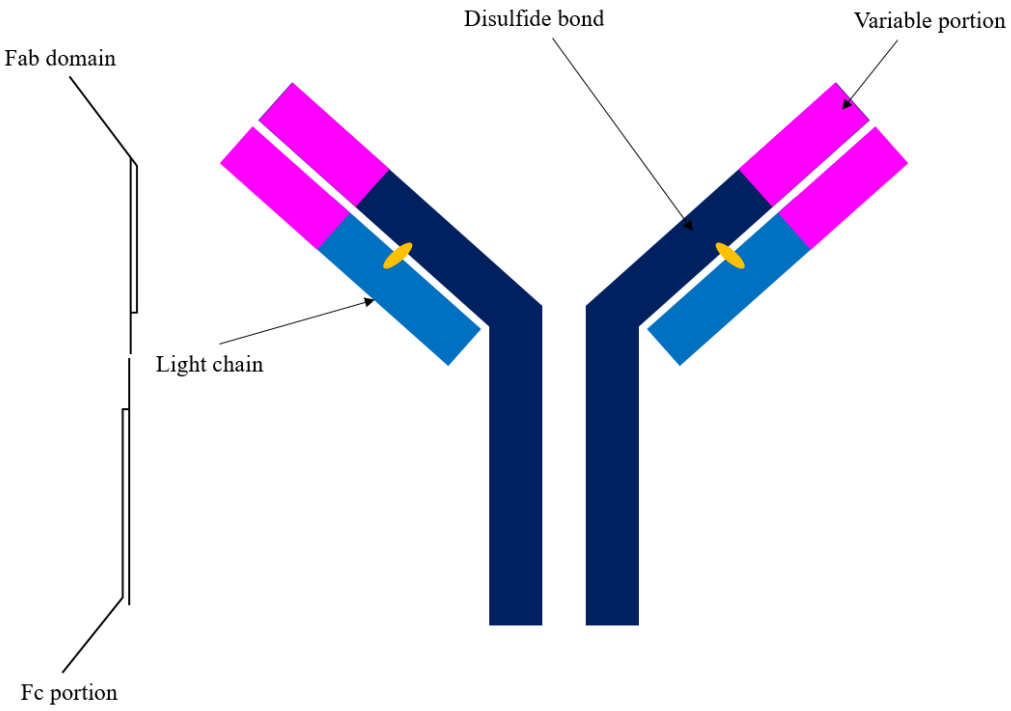
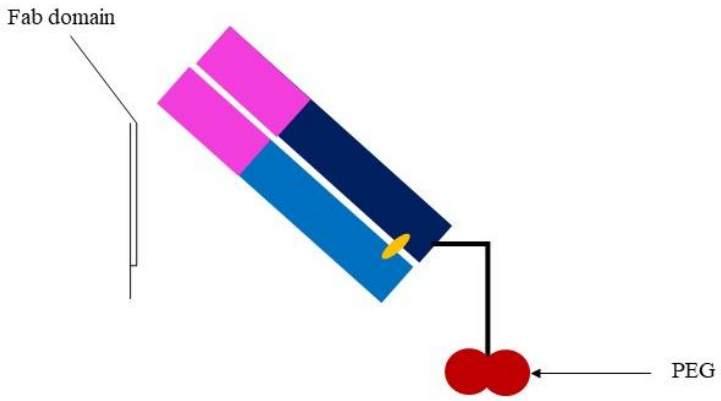


Figure 1.4. Simplified structure of certolizumab.



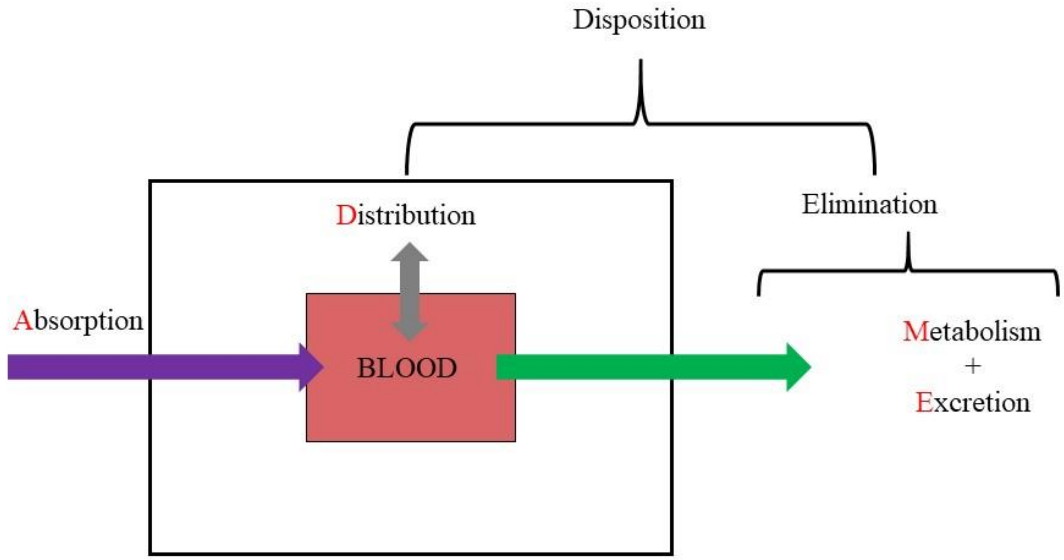
ABBREVIATION: Polyethylene glycol (PEG).

The Fc portion binds to two different receptors in order to enact its various functions. Firstly, via Fcγ receptors (such as FcγIIIA on the surface of natural killer cells), the Fc portion binds to effector systems. It also binds to neonatal Fc receptor (FcRn); this receptor facilitates protection against natural IgG breakdown inside cells, and hence, is involved in the clearance of both natural IgGs and mAbs containing an Fc portion[152]. These properties are vital in understanding the variability in the PK of mAbs, and why certolizumab behaves differently from other mAbs in the same TNFi class, given its absence of an Fc portion.

1.9.2. Basic PK characteristics of bDMARDs

In order to better understand the pharmacological properties of bDMARDs, some basic PK terms will first be defined. PK processes can be considered using the acronym "ADME," which represents absorption, distribution, metabolism and excretion[153]. A simplified diagram of these processes is outlined in Figure 1.5. Absorption constitutes the transfer of a drug from its site of administration (e.g. oral, SC) to systemic circulation; drug levels (which constitute the measurement of unchanged administered compound) can then be measured in blood or plasma. Absorption is measured in terms of its rate and extent, which can be affected by various factors, such as adequate time at absorption site and blood flow. Bioavailability is a component of absorption, and refers to the fraction of the administered dose that proceeds unaltered from the site of administration to systemic circulation. Distribution refers to the reversible transfer of a drug between blood and tissues, and can be affected by both plasma protein binding and tissue binding. Protein binding of a drug is mostly reversible, and bound and unbound drug fractions are always in a state of dynamic equilibrium. Drug metabolism is a process that is more relevant to small molecule drugs (as opposed to bDMARDs) and is also known as biotransformation; enzymatic processes (mostly in the liver, but also in the gut wall) break down drug molecules to facilitate elimination of the drug from the body. Excretion describes removal of the unchanged drug from the body. The two major routes of drug excretion are the kidneys and in bile, but minor routes include in saliva, sweat, tears, expired air and breast milk. The term "elimination" is used to cover the processes of metabolism and excretion combined, and the term "disposition" covers all the processes involved in distribution, metabolism and excretion.

Figure 1.5. An overview of PK processes (ADME; absorption, distribution, metabolism, elimination).



bDMARD agents are large molecules, originally designed for intravenous (IV) administration. However, agents are now overwhelmingly formulated to be administered by the SC route, as these can be self-administered by patients in the community, without requiring a hospital visit for every IV infusion. Despite the majority of bDMARDs now being developed and prescribed via the SC route, the mechanism of systemic absorption via this route has not been fully elucidated. It is thought that absorption predominantly occurs via a combination of diffusion across blood capillary beds and convection through lymphatic vessels[154]. The absorption of bDMARDs via the SC route is often assumed to be first-order i.e. a fixed fraction of drug is absorbed per unit time, and bioavailability ranges between 50-80%[155 156]. In zero-order drug kinetics, a fixed amount (as opposed to fraction) of a drug is absorbed or eliminated per unit time, and the rate of drug absorption/elimination is independent of the available amount or concentration (as opposed to first-order kinetics, where the rate is proportional to the available amount or concentration).

The volume of distribution (V or V_D) is a PK parameter that gives an indirect indication of the degree of tissue distribution of a given drug; the V_Ds of TNFi are modest, ranging between a minimum of 4.5L for infliximab[157] and a maximum of 26.3L for golimumab[158]. The V_D is calculated by dividing the amount of drug in the body by its measured plasma concentration, so this low V_D of bDMARDs can be explained by their large, hydrophilic structure. Theoretically, this causes low tissue penetration and confinement to capillaries and

lymphatic vessels, but mAbs are able to enter cells via two mechanisms: firstly, by fluid phase endocytosis, which tends to take place in endothelial cells, and secondly, by FcγR-mediated endocytosis; FcγRs are proteins expressed on the surface of various immune cells, as well as platelets[154].

Pharmacological models to explain the disposition of mAbs are wide-ranging in complexity and have no consensus; pharmacological models will be discussed in more detail in the next section. However, Fronton *et al* used a pharmacological modelling technique, physiologically-based pharmacokinetic (PBPK) modelling, in order to demonstrate that the disposition of mAbs can be considered on a whole-body level, with tissue distribution being rate-limited by extravasation, and elimination taking place from a variety of tissues as well as plasma[159]. They also found that when common data such as plasma or plasma *plus* tissue drug levels were utilised, it was not possible to determine which tissues were eliminating the drug, so models cannot simply be based on measured data.

Due to their large size, immunoglobulins and most proteins cannot be eliminated from the body by conventional small molecule methods of excretion i.e. renal and biliary, nor can they be broken down by hepatic metabolism. IgG is eliminated via two pathways: intracellular elimination, which is non-specific, and target-mediated elimination, which is specific to IgG. CL is a steady-state concept, and represents the apparent volume of plasma (or blood or plasma water etc.) that is completely cleared of drug per unit time. It is defined by the rate of elimination divided by the plasma concentration of a drug. Therefore, the total CL of a given mAb is the combination of non-specific (linear) and target-mediated (non-linear) CLs[151].

It is not fully understood how mAbs are cleared from systemic circulation, although it is known that IgG are cleared via intracellular catabolism following uptake into cells, and this constitutes the majority of IgG elimination[160]. IgG enters cells via a range of mechanisms, including endocytosis, internalisation upon Fab-antigen binding at the target cell surface and Fc-FcγR binding at immune cells[161]. If a cell contains FcRn, the Fc portion of a mAb will bind to it at pH <6.5, gaining protection from lysosomal degradation. At neutral pH (7.4), the mAb is released back into systemic circulation or extracellular fluid. This FcRn salvage prolongs the half-life ($t_{1/2}$) of IgGs to approximately 21 days, which is longer than other comparable protein molecules. FcRn saturation is possible, but only when IgG concentration

is high e.g. during treatment with polyclonal Ig[162], but usual therapeutic mAb doses do not cause this to occur. Therefore, non-specific elimination of mAbs is a linear process.

Target-mediated drug disposition (TMDD) pertains to high-affinity binding of a drug (such as a mAb) to its pharmacological target (i.e. the corresponding antigen of the mAb), leading to non-linear drug elimination[163]. In the case of mAbs, binding with its target antigen leads to the formation of a mAb-target complex, which is then eliminated by the immune system; the full mechanism of elimination has not yet been determined, but can be described using a TMDD model[164]. Multiple factors affect target-mediated mAb elimination, such as antigen turnover, saturable and reversible mAb-antigen binding and mAb-antigen complex elimination rates. TMDD increases with the amount of available antigen, but antigen turnover is often unknown. In this instance, the TMDD model can be approximated using PK models that incorporate both linear and non-linear CL (termed Michaelis-Menten kinetics)[165].

Given that understanding of the behaviour of mAbs in the body has not been fully achieved, for example, due to lack of studies of subcutaneously administered bDMARDs, or due to variations in bDMARD disposition, pharmacological understanding of these drugs remains an area of unmet need.

1.9.3. Concepts in pharmacological modelling

While the PK of drugs in the physiological conditions of the body is complicated, pharmacological modelling seeks to generate a simplified explanation of PK processes. Pharmacological models are concise representations of an overall "system" (i.e. the body) that aim to confer knowledge or understanding of a specific drug in that system[166]. As with other mathematical models to explain complex systems, pharmacological models are best assessed on their "fitness for purpose," as opposed to how "correct" or "true" they are. This is summarised in the famous George Box quote: "Essentially, all models are wrong, but some are useful"[167].

In the discipline of clinical pharmacology, modelling and subsequent simulation are established methods for amalgamation of collected data, pre-existing knowledge and known drug mechanisms, which enables ensuing decisions to be made on drug prescribing and

development. Determining the correct dose of a therapeutic agent is essential; as Paracelsus* stated in the early 16th Century: "All things are poison and not without poison; only the dose makes a thing not a poison". This concept has always held true in clinical pharmacology; appropriate selection of dosage and dosing regimen are the cornerstones of ensuring that drugs behave as therapeutic agents, and not as pathological poisons[168]. An example of this is the concept that the therapeutic benefit of bDMARDs in controlling the inflammation of RA must be balanced against the risk of infection from immunosuppression.

There is variability in circulating drug levels of bDMARD agents between patients; this is an example of between-subject variability (BSV) in drug exposure. BSV occurs with all drugs, and also exists in drug response, as well as exposure, such as in the heterogeneous treatment response to bDMARDs[166]. Multiple factors can account for BSV, such as age, biological sex, body weight, genotype, renal and hepatic function and concomitant medications. Through modelling of study subjects, the effect of these covariates on the PK of a drug can be determined and dosing recommendations can be made, thus improving drug safety and efficacy via mitigation of variability in drug exposure.

PK models seek to describe how drug concentrations vary over time. Many PK models incorporate components called "compartments," which each represent a region of the body in which the drug of study is well-mixed and kinetically homogeneous; this means that the drug can be described with respect to a single representative concentration at any given time point[169]. Compartment choice and connection can deeply influence differences between models. Human models tend to employ a central compartment (representing plasma), interlinked with one or two peripheral compartments via rate constants (e.g. k_{12} and k_{21})[170]. ka represents the single elimination rate constant in a one compartment model. A compartment can correspond to an actual physiological space in the body (such as blood or extracellular fluid), but compartments tend to be more abstract concepts that do not map to a physical region in the body.

Whilst PK explains the behaviour of a drug inside the body, pharmacodynamics (PD) explains the effect of the drug on the body and incorporates treatment response outcomes. One example of pharmacological modelling that incorporates both PK and PD is PK/PD (or PKPD) modelling, which can relate the PK of a drug to clinical outcomes[166]. The relationship

* Philippus Aureolus Theophrastus Bombastus von Hohenheim, also known as Paracelcus (c.1493-1541).

between drug concentration and effect can be described as a continuous function (e.g. linear, E_{max} i.e. the maximum effect of the drug, sigmoid E_{max}). PKPD models can also instead take on an exposure-response format, where instead of using concentration as the independent variable, steady-state drug exposure is instead adopted. It is also possible to adopt such models to utilise other measures of exposure, such as AUC and peak plasma concentration, (C_{max}). An example of pharmacological modelling that assesses PK data alone over time is popPK modelling, which will be discussed in the next section.

1.9.3.1. PopPK modelling

Individual PK studies employ detailed drug concentration-time data (measurement over many time points), typically using non-compartmental methods. By contrast, popPK modelling (also known as mixed-effects modelling) employs concentration-time data from multiple individuals and does not require rich or balanced drug concentration sampling. Methods devised by Sheiner *et al* in 1972 enable pooling of sparse sampling data from multiple participants for estimation of population mean parameters, BSV and the effects of any possible covariates on BSV, as well as parameter precision measurement via calculation of standard errors (SE)[171]. In essence, popPK modelling represents a PK study, but carried out in a group of individuals, and it can be used to analyse sparse data, where the standard approach of non-compartmental analysis cannot be used.

PopPK modelling relies on detailed and accurate information on dosing, drug concentration and population covariates[166]. This form of modelling is based on three components: structural models, stochastic models and covariate models. Structural models can be expressed as either algebraic or differential equations. Algebraic equations are the simplest representation of a PK model, and explain the relationship between drug concentration and time. However, some PK systems cannot be summarised in algebraic equations due to their complexity, but instead, they can be stated as differential equations that describe the rate of change of a variable. Stochastic models explain both the variability between subjects, as well as between drug concentration and different time points. Covariate models incorporate pre-identified covariates into popPK models to explain observed differences in parameter estimates between individuals. Model-based approaches for both drug development and maximising therapeutic utility of pre-existing drugs continues to evolve and gain importance in clinical practice.

1.9.3.2. Pharmacological modelling of TNFi drugs in patients with RA

A number of popPK studies have previously been carried out in patients with RA receiving TNFi agents. One of the more commonly studied TNFi agents is etanercept, which is not surprising, given its frequency of prescription due to its early licensing date (1998) and SC form of administration. Lee *et al* studied administration of etanercept 25mg SC twice-weekly in 25 participants and etanercept 50mg once-weekly in 77 participants in a popPK and PD modelling study[172]. Using a one-compartment popPK model, they identified sex and race as significant covariates. Zhou *et al* carried out a popPK analysis using data from a phase 3 trial in RA patients and HCs receiving etanercept 25mg SC twice-weekly, and compared two groups, one receiving MTX and one without MTX[173]. They found that a two-compartment model with first-order elimination and either zero-order input (IV) or first-order absorption (SC) demonstrated superiority over a one-compartment model. Age and body weight were found to affect CL and race affected V_D in the central compartment. However, this study included IV data and etanercept was not dosed using the dosing regimen that is now most commonly employed (50mg SC weekly), so may not reflect current clinical practice. An integrated popPK study using HC and patients with RA, ankylosing spondylitis (AS) and juvenile idiopathic arthritis (JIA) used a two-compartment model to demonstrate that age and body weight affected CL in paediatric subjects, but that no covariates affected the model in adult patients[174]. The study subjects were deliberately heterogeneous, and dosing regimens varied between IV and SC at a range of different doses (but not 50mg SC weekly). Shennak *et al* carried out a PK study comparing Enbrel, the proprietary originator compound of etanercept, with a new etanercept biosimilar YLB113, and they used non-compartmental analysis to carry out PK estimates of a study population of 52 biologically male subjects[175]. This study demonstrated a similar PK profile between patients receiving Enbrel or YLB113.

Using a PK-PD approach, Hsu and Huang used DAS28 as a clinical endpoint in a meta-analysis of multiple published pharmacological studies[176]. They found that a one-compartment model with first-order absorption and elimination best represented their etanercept plasma concentration-time data and an inhibitory E_{max} model was used to characterise the relationship between predicted etanercept cumulative AUC and DAS28. They were also able to simulate a number of alternative dosing regimens which were equally effective to etanercept 25mg twice-weekly in alternative dosing scenarios.

A number of pharmacological studies have also been carried out on adalimumab, the other bDMARD of interest in this thesis. Prior to licensing, the United States Food and Drug

Administration (FDA) carried out a review of adalimumab studies[177]. Interesting findings were that MTX decreases adalimumab CL, and that a popPK study demonstrated that adalimumab increased the CL of MTX by 40%. Body weight > 82kg was found to increase CL, and increasing age was associated with reduced CL. However, a subsequent phase 1 trial demonstrated that repeated administration of adalimumab IV did not have a significant effect on MTX PK and concluded that MTX dose adjustments were not necessary[178]. Ducourau *et al* carried out a PKPD study in 127 samples from 30 RA patients receiving adalimumab 40mg every two weeks SC, which demonstrated both large BSV in response to adalimumab, as well as that the target adalimumab concentration was associated with individual disease activity measured using DAS28[179]. Another PKPD study by the same group studied RA patients receiving adalimumab 40mg SC every two weeks, and they found that a one-compartment first-order absorption model best described the data[180]. Maximal response to the drug was measured in terms of improvement in DAS28 and CRP levels, and simulations showed that a one-off loading dose of 160mg resulted in increased adalimumab concentrations, with maximal response being reached before the second injection.

Other studies have been carried out in RA patients receiving certolizumab[181 182], infliximab[183-185], abatacept[186 187] and golimumab[188-190], with not dissimilar results to those described for etanercept and adalimumab above. All of the studies discussed in this section largely employ trial data, and not real-world drug administration and concentration data, and none have been carried out in biosimilar versions of either etanercept or adalimumab, which are the predominant formulations that are currently prescribed in UK clinical practice. Furthermore, none have simulated alterations in dosing intervals (as these medications are usually prescribed in pre-filled syringes with a set dosage), although some have simulated altered loading doses, which are not currently recommended for patients with RA. Further popPK studies, particularly in biosimilar versions, will lead to improved understanding of these drugs' PK parameters in patients with RA, and will add further to the body of knowledge.

1.10. Chapter summary

Treatment response to therapeutic intervention in patients with RA is heterogeneous and can be measured using a range of different outcomes. Some predictors of prognosis and treatment response have been previously determined, but these factors do not entirely explain this heterogeneity. Various antibodies (such as ACPA and anti-CarP) have been shown to be associated with RA prognosis, so biomarker discovery studies searching for other

antibodies or proteins may fill in some of the knowledge gaps. Studies linking these proteins back to a patient's underlying genetics may additionally aid understanding of disease pathogenesis, as well as contributing to personalisation of therapeutic prescribing in RA. Whilst proteomic studies give a dynamic snapshot of a patient's physiology at the time of study, genetics are more stable and may represent a more practical means of patient classification in the clinic room. Furthermore, PK properties of bDMARD medications used in the treatment of RA have not been comprehensively studied in a real-world environment, and in particular, this has not been carried out in the now more widely-prescribed biosimilar versions of bDMARDs. The remainder of this thesis will seek to address some of the knowledge gaps that have been identified here.

CHAPTER TWO: HYPOTHESIS AND AIMS

Hypothesis:

Biological factors, such as protein expression, contribute to variability in circulating drug levels and treatment response to biologic agents in patients with RA.

Aims:

To identify factors underpinning adalimumab and etanercept PK and clinical response.

Objectives:

1. Define popPK models in patients with RA starting either adalimumab or etanercept biosimilars (Amgevita and Benepali, respectively).
2. Use this model to simulate alterations in dosing intervals in order to achieve maximal effect and/or steady state at an accelerated time point.
3. Carry out detailed protein mapping using SWATH-MS in these patients to ascertain whether any serum protein levels or objective measures of inflammation are predictive of drug concentration.
4. Carry out a discovery SWATH-MS proteomics study in extant samples from patients with RA who have received etanercept as a first-line bDMARD to develop a protein-based classification model of treatment non-response.
5. Carry out pQTL analysis to determine whether serum protein levels are predicted by genetics.
6. Determine whether genetic markers that are correlated with protein levels are predictive of treatment response.

CHAPTER THREE: METHODS

Summary of chapter contents:

3.1. PopPK study of RA patients starting Amgevita or Benepali: the Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate Personalised Dosing sub-study

3.1.1. Ethical approval

Study participants were recruited as part of a sub-study of a pre-existing study: the Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate (BRAGGSS). The popPK study conducted for this thesis was carried out under the BRAGGSS Personalised Dosing (BRAGGSS-PD) sub-study. Both BRAGGSS and BRAGGSS-PD received favourable ethical approval (Research Ethics Committee, REC, reference: 04/Q1403/37). Following an application for a substantial amendment written by the author and supported by the senior study coordinator for BRAGGSS, Sarah Ashton, the BRAGGSS-PD sub-study was approved under Substantial Amendment 17a in November 2018. As part of the ethics application, a new study protocol was written by the author, and new patient information sheets and consent forms were written by the author and Sarah Ashton. As the study progressed, a further substantial amendment was submitted in order to incorporate combined patient information sheets and consent forms (written by the author and Sarah Ashton) for both BRAGGSS and BRAGGSS-PD, in order to rationalise paperwork; these were approved under substantial amendment 17b in November 2020. All the above documentation is provided in Appendix 1.

As part of the protocol, the first visit necessitated a patient home visit in most cases, as the majority of patients required education on use of their auto-injection device by a nurse employed by the medication provider (Healthcare at Home Limited) and blood samples were taken before and after the administration of a patient's first biologic dose. Hence, additional lone worker training and approvals were obtained by the author from the University of Manchester. A full risk assessment was carried out by the author and Sarah Ashton. Details of the lone worker risk assessment and policy for the BRAGGSS-PD sub-study are provided in Appendix 2.

3.1.2. Study setting and funding

The sponsor for this study was The University of Manchester. This study was funded by the National Institute for Health Research (NIHR) Manchester Biomedical Research Centre, both via a fellowship to the author and through consumable funding, as well as by Versus Arthritis as part of a core programme grant that covered BRAGGSS oversight.

BRAGGSS is a long-term multi-centre prospective observational study based in the UK. Its aim is to collect genetic, serological, clinical and psychological information from patients with RA commencing bDMARD and tsDMARD therapy in order to determine how these factors influence treatment response. The aim of the BRAGGSS-PD sub-study was to carry out a popPK study in patients with RA commencing either Amgevita (adalimumab biosimilar) or Benepali (etanercept biosimilar); patients underwent repeated serum drug level sampling over the first 12 weeks of treatment – this will be described in more detail in Section 3.1.5.

BRAGGSS-PD initially opened recruitment at one site in November 2018: Manchester Royal Infirmary, part of Manchester University NHS Foundation Trust. Apart from Visit One (in most patients' cases – this will be detailed in Section 3.1.6), all study visits took place at the NIHR Manchester Clinical Research Facility. However, recruitment was monitored throughout the study, and due to recruitment challenges, a second centre was opened at Bolton One (part of Bolton NHS Foundation Trust) in June 2019. All visits apart from Visit One were carried out at Bolton One.

Recruitment to BRAGGSS-PD was closed between March 2020 and September 2020, due to a combination of national lockdown due to the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic, and the redeployment of the author from PhD studies to work full-time on frontline NHS services to support the SARS-CoV-2 response. Manchester Royal Infirmary re-opened recruitment in September 2020 and Bolton One re-opened recruitment in October 2020, following confirmation of capacity and capability. Furthermore, because recruitment targets were further impaired due to the SARS-CoV-2 pandemic, a further recruitment centre was opened at Tameside General Hospital, part of Tameside and Glossop Integrated Care NHS Trust in October 2020. Further approvals from the University of Manchester's research governance department were obtained so that following resumption of BRAGGSS-PD after lockdown, all visits could be carried out as

patient home visits in order to avoid multiple hospital attendances for patients who were likely to be shielding.

Finally, in order to include participants from more diverse social and ethnic backgrounds, further approvals were gained from Health and Safety within the University of Manchester so that patients with inactive, previous infection with hepatitis B and C could be recruited. A new biological Control of Substances Hazardous to Health Regulations (BioCOSHH) form was written and submitted by the author, and the author attended a panel to discuss this new protocol and gained approval in April 2021. The new BioCOSHH form is included in Appendix 3.

Sample processing and storage was carried out at the Centre for Musculoskeletal Research (CfMR), the University of Manchester, UK.

3.1.3. BRAGGSS-PD study participants

Participants in the popPK study were recruited to the BRAGGSS-PD sub-study. Patients with RA starting either Amgevita or Benepali at the rheumatology departments of Manchester Royal Infirmary, Bolton One and Tameside General Hospital were informed about the study by local clinical staff and invited to participate. Following either a face-to-face or telephone initial visit by the author, patients were then invited to give consent to participate. Recruited patients provided written informed consent in compliance with Good Clinical Practice (GCP) and the Declaration of Helsinki. In-date GCP certification was held by the author throughout the course of this PhD (Appendix 4). All patients were prescribed either Amgevita 40mg SC every 2 weeks or Benepali 50mg SC every week throughout the duration of the study, in accordance with the licensed indications and dosages of these medications.

The inclusion and exclusion criteria were as follows:

Inclusion criteria

- bDMARD-naïve patients who were due to commence either adalimumab or etanercept biosimilars.
- Patients with a diagnosis of RA according to the ACR 1987 or ACR/EULAR 2010 criteria who were willing and able to participate in the study (including follow-up visits) after providing informed consent.

- Patients had to have a DAS28 greater than or equal to 5.1; this score had to be taken at the point of consent, or up to one calendar month prior to the date of consent.

Exclusion criteria

- Patients either unwilling or unable to donate a blood sample, provide informed consent or to participate in follow-up.
- Patients who received injectable steroids for 4 weeks prior to the last DAS28 measurement before starting the study. If a patient was on long-term steroids (4 weeks or longer) and then required a bDMARD because their DAS28 was ≥5.1, they could be recruited, as any improvement could be confirmed to be due to their bDMARD, and not their steroid.
- Patients who had not had a DAS28 within the last calendar month at the point of consent.
- Patients who were not able to attend for their clinical follow-up visits.
- Patients who were not willing to provide study blood samples either at the point of consent, or at another point prior to their treatment start date.

For popPK studies, it has previously been demonstrated that to achieve a power of at least 80% in the estimation of PK parameters (i.e. CL, volume, absorption rate), a sample size of 20-30 individuals is sufficient[191]. Therefore, a target sample size of at least 20 patients on any one drug (Amgevita or Benepali) was selected for this study; ethical approval was obtained for a total of 30 patients on each drug. Separate popPK models were developed for each agent. Recruitment was not restricted to any one agent for pragmatic reasons.

3.1.4. Clinical assessments

Baseline clinical and demographic data were collected via a case report form (CRF) using data from each patient's case notes both prior to bDMARD commencement. CRFs were completed by staff at each participating clinical site, including the author at Bolton One and Tameside General Hospital. Patient data was collected from a self-reported questionnaire that each patient was asked to complete. The following data were collected:

- Demographic data:
    - Age.
    - Sex.
    - Height.
    - Weight.

    - Smoking status (current/ex/never).
- Clinical data:
    - Year of diagnosis.
    - Joint replacement.
    - Components of ACR 1987 and ACR/EULAR 2010 criteria.
    - DAS28 and its components.
    - Current medications, including non-rheumatological medications.
    - Previous bDMARD/tsDMARD/csDMARD therapy.
    - Recent steroid use.
    - Tuberculosis screening information.
    - Herpes zoster immunity.
    - ACPA status.
    - Co-morbidities.
- Patient data:
    - Marital status.
    - Ethnicity.
    - Place of birth.
    - Employment status and occupational information.
    - Alcohol consumption.
    - Visual analogue scales (VAS) for pain, fatigue and overall health state.
    - HAQ and its components.
    - EQ-5D and its components.
    - The General Self-Efficacy Scale.
    - Hospital Anxiety and Depression Scale (HADS) and its components.
    - Illness Perception Questionnaire (IPQ) regarding RA.

After three months of bDMARD therapy, the following data were collected:

- Demographic:
    - Weight.
    - Patient vital status.
- Clinical:
    - Any changes to bDMARD/tsDMARD/csDMARD medications.
    - Adverse events to bDMARDs.
    - DAS28 and its components.
- Patient data: as for baseline collection.

3.1.5. Sampling schedules

Optimal sampling time intervals for Amgevita and Benepali were proposed by Dr Adam Darwich and Dr Kayode Ogungbenro, respectively, using previously defined popPK models from the literature in PopDes software, an application software that can be utilised for determining optimal sampling times or windows for PK and popPK studies[192]. All samples were collected by the author, and all doses given at study visits were witnessed by the author.

For Amgevita, a previous PK model derived for adalimumab by Ternant *et al* [180] was used to simulate multiple dosing interval models. The optimal sampling design for Amgevita was deemed to be at: baseline (pre-treatment), 1 hour post-first dose, then 2, 4, 6 and 12 weeks post-first dose. Doses of Amgevita were witnessed immediately after all samples had been obtained to ensure: 1) a true trough drug level; and 2) true patient adherence on initiation of drug. The sampling schedule is summarised in Figure 3.1.

For Benepali, previous PK models derived for etanercept by Lee *et al*[172], Yim *et al*[193] and Zhou *et al*[174] were utilised for simulation of multiple dosing interval models. The optimal sampling design for Benepali was deemed to be at: baseline, 1 hour post-first dose, 6 days post-first dose, then 2, 4, 6 and 12 weeks post-first dose. Again, administration of Benepali was witnessed immediately after all sampling, except after the Day 6 sample. The time for the second dose (Day 7) was contemporaneously communicated via text message immediately after the dose was administered. The sampling schedule is summarised in Figure 3.2.

3.1.6. Study protocol

Initial pre-treatment visit

The pre-treatment visit could either be carried out in person, when patients were attending drug education clinics at their parent rheumatology department, or via telephone, after suitable patients were identified by their parent rheumatology department. The patient information sheet (PIS) was discussed with each patient and if the visit was in person, patients would be invited to sign a hard-copy informed consent form (ICF). If the visit was conducted via telephone, patients would be invited to sign a consent form at Visit One.

Visit One

Prior to visit one, the author liaised with each patient to find out when their delivery of drug was due. Visit one was conducted using one of two methods, depending on whether a patient required auto-injectable device training from the medication provider, Healthcare at Home Limited.

When device training was not required, the visit initially consisted of:

- Signing the consent form (if not already completed).
- Baseline (pre-treatment) blood sample.
- Witnessed administration of drug.
- Further blood sampling one hour after first dose of drug.
- End of visit.

When device training was required, the visit had to be carried out as a patient home visit in order to coincide with attendance by a nurse provided by Healthcare at Home. The visit was conducted as described below:

- Signing of consent form (if not already completed).
- Baseline (pre-treatment) blood sample.
- Nurse attendance for device training.
- Witnessed administration of drug.
- Further blood sampling one hour after first dose of drug.
- End of visit.

Subsequent visits

Subsequent visits had a simplified form, as no one-hour blood sample was required:

- Trough blood sampling.
- Witnessed administration of drug.
- End of visit.

Figure 3.1. Amgevita sampling schedule overlaid on example PK profile of drug. DNA for genotyping and routine clinical blood sampling are routine samples for BRAGGSS. Drug levels and protein mapping are additional serum samples for the BRAGGSS-PD sub-study.

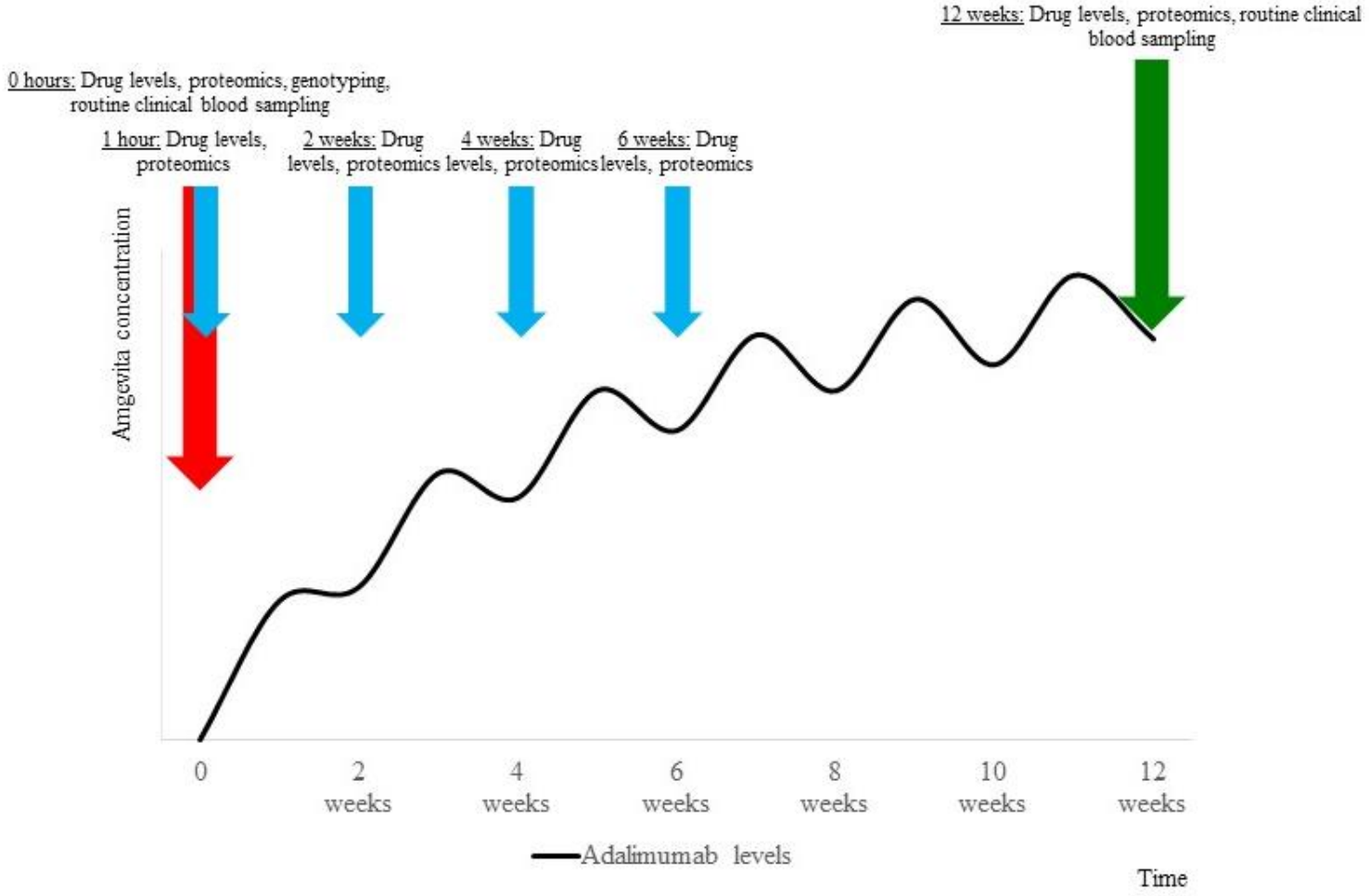
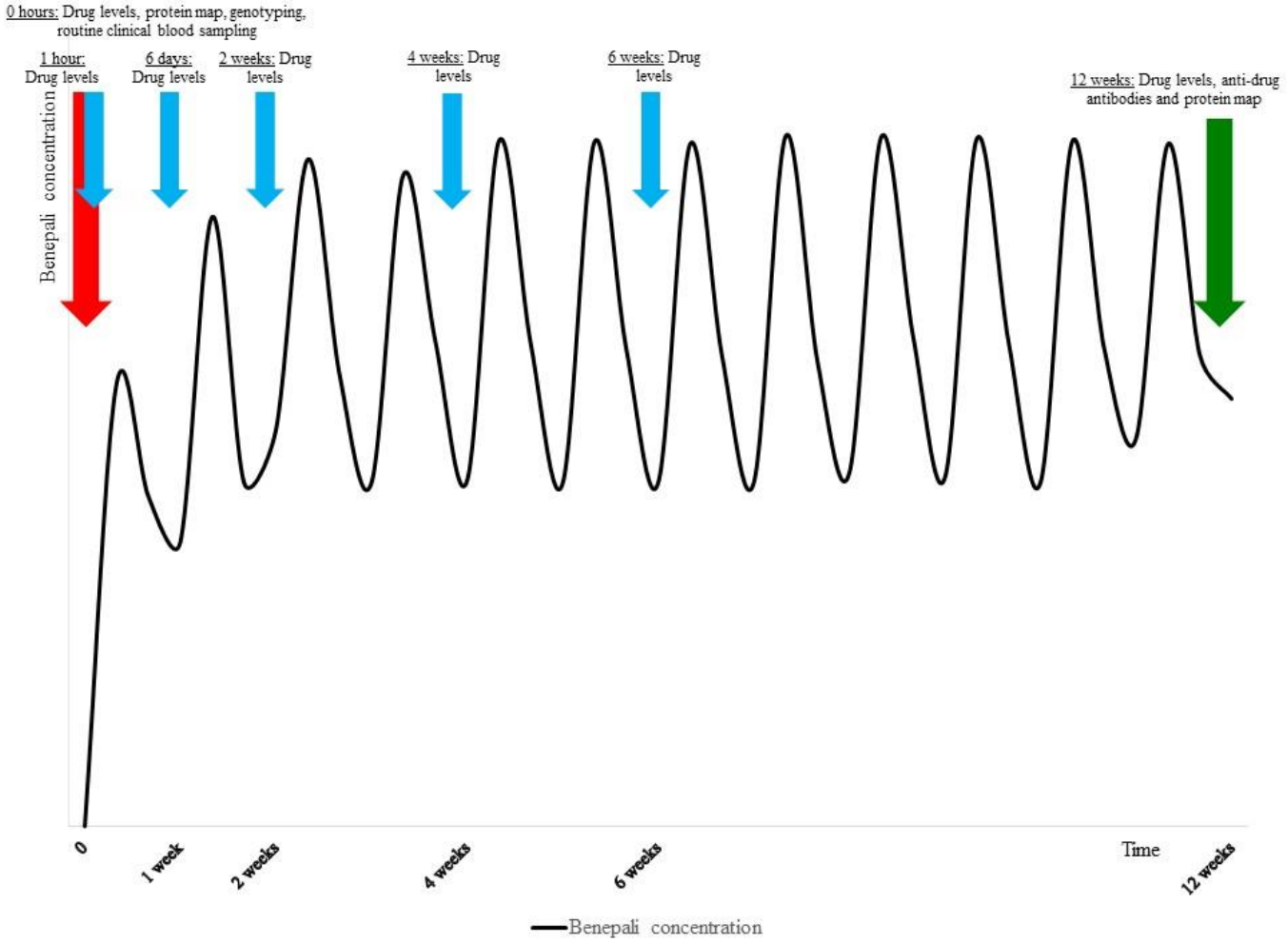


Figure 3.2. Benepali sampling schedule overlaid on example PK profile of drug. DNA for genotyping and routine clinical blood sampling are routine samples for BRAGGSS. Drug levels and protein mapping are additional serum samples for the BRAGGSS-PD sub-study.



3.1.7. Sample processing

Serum samples were collected from patients according to the sampling intervals detailed in Section 3.1.4. Additional samples of serum, RNA, plasma and PBMCs were collected at pre-treatment and 12 weeks, as per the pre-existing BRAGGSS protocol. All patient samples collected were processed and stored at the CfMR and were processed by CfMR laboratory staff. All sample blood tubes were spun at 1,720g for 10 minutes, then extracted into aliquots. Serum, plasma and RNA samples were stored in -80^{o}C freezers and PBMC samples were stored in -150^{o}C freezers.

3.1.8. Measurement of drug levels

Drug level measurements were carried out by CfMR laboratory staff using commercially-available ELISA-based test kits produced by Grifols International, SA (Barcelona, Spain). The Promonitor®-ADL-1DV kit was used to measure Amgevita levels, and the Promonitor®-ETN-1DV kit was used to measure Benepali levels. Standard laboratory equipment and a spectrophotometer (SpectraMax® Plus 384 Microplate Reader, Molecular Devices, LLC, San Jose, California, USA) were used during the experimental procedure. Samples were defrosted for two hours at room temperature, prior to thorough mixing before the experimental procedure.

Serum drug levels were measured using 96-microwell ELISA plates, which were pre-coated with anti-adalimumab and anti-etanercept human monocolonal antibody, according to which drug was being measured (Amgevita and Benepali, respectively). Patient samples were diluted to 1:50 concentration using a dilution buffer and were transferred to separate wells. Pre-diluted calibration samples and positive and negative controls were also included for purposes of quantification of results and quality control; these were also transferred to separate wells. Any drug present in the patient samples, calibration samples and controls became bound to the immobilised anti-drug antibodies during an incubation period of one hour at room temperature. Following incubation, any unbound material was removed by washing the wells with a 20X wash buffer containing phosphate-buffered saline and tween-20. Each well was then loaded with a second horseradish peroxidase-labelled anti-drug monoclonal antibody to form a sandwich complex. The plate was incubated for a further hour at room temperature to allow the labelled antibody to bind to the drug attached to the microwells. Unbound enzyme-labelled antibody was again washed away with wash buffer, and a substrate of pre-diluted stabilised tetramethylbenzidine was added to measure enzyme activity. After 15 minutes, a stop reagent of pre-diluted sulfuric acid solution was added to

halt the reaction. Colour intensity as a result of the enzymatic reaction was measured in triplicate using a spectrophotometer at wavelength 450nm. The generated optical density values were proportional to the drug concentration in each sample.

Softmax Pro 7 software (compatible with the SpectraMax® Plus 384 Microplate Reader, Molecular Devices, LLC, San Jose, California, USA) was used to interpolate the optical density values and determine drug level concentrations. Interpolated values were multiplied by the dilution factor (x50) to obtain drug levels in patient samples.

3.1.9. PopPK analysis

All data cleaning and formatting and statistical analysis in this section was carried out by the author.

3.1.9.1. Software

PK data were analysed using a population approach with Monolix v.2019R2 software (Lixoft, Antony, France). This is a non-linear mixed-effects modelling software package based on the stochastic approximation expectation maximisation (SAEM) algorithm which optimises maximum likelihood without any approximation. SAEM convergence has been proven to be robust and reliable[194]. Monolix works by optimising the maximum likelihood in order to produce optimal population parameter values; final estimates maximise the likelihood of the data, given the model. Maximisation of likelihood is equivalent to minimisation of minus two times the logarithm of the likelihood (-2LL) of the data, given the model. Both positive and negative values of -2LL are possible following model development, and no special value is assigned to direction. The default simulated annealing option for SAEM was used in order to maintain a larger parameter space for a prolonged period of time (in comparison to without simulated annealing); this enables the escape of local maximum values and improves convergence towards a global maximum. A maximum of 500 iterations was set in order to ensure best possible convergence. For medication doses that were self-administered by patients outside of study visits, the nominal dose timings were used for the purpose of modelling.

3.1.9.2. Structural models

For each drug studied, one, two- or three-compartment mammillary models assuming first-order absorption and elimination were tested. Estimated PK parameters were given as apparent values, due to extravascular administration via the SC route. PK parameters were

parameterised as clearance (CL) and volume of distribution (V_D). Structural models were compared using the Akaike information criterion (AIC), which is defined as:

$$AIC = -2LL + 2p$$

where p is the total number of model parameters to be estimated. For each drug, the model with the lowest AIC value with the most parsimonious combination of estimates, covariates and correlations was selected over competing models.

3.1.9.3. Between-subject variability and unexplained residual error models

BSV in PK parameters was described using an exponential model, defined as:

$$\theta_i = \theta_{TV} \times \exp(\eta_i)$$

where θ_i is the estimated individual parameter, θ_{TV} is the typical individual value of the parameter and η_i is the random effect for the *i*th patient i.e. the *i*th patient's deviation from the typical value, θ_{TV}. Values of η_i were assumed to be normally distributed, with a mean of zero and a variance of ω^2. For parameters where BSV could not be estimated, this was removed from the analysis and therefore, only typical individual values were estimated. Correlations between parameters were also tested during model development. Additive, proportional or combined additive and proportional models were tested for residual unexplained variability (RUV) during model development. This represents the error difference between the model prediction at each time point between an individual and the observed data.

3.1.9.4. Covariate model development

Due to a limited number of patients, three covariates were tested in the analysis: age, body weight (continuous covariates) and biological sex (binary covariate). Covariate models were compared using both -2LL and AIC. Models with the lowest significant -2LL value (assessed using a likelihood ratio χ^2 test, LRT) and the lowest AIC, with the simplest combination of covariates and between-variable correlations were selected. The effect of each covariate on each PK parameter was tested using an LRT with $\alpha = 0.05$. The most significant covariates were kept in the final model if shown to usefully improve model fit whilst maintaining the simplest possible structural model, given the low number of study subjects. Stepwise

forwards/backwards selection was not required, due to the low number of pre-selected covariates.

3.1.9.5. Model goodness of fit and evaluation

The goodness of fit for each model was visually assessed using plots of:

- Population-predicted (PRED) and individual-predicted (IPRED) measurements versus observed measurements.
- IPRED and observed concentrations (DV) versus time.
- Residuals, represented in plots of:
    - Population weighted residual distributions (PWRES).
    - Individual weighted residual distributions (IWRES).
    - Normalised prediction distribution errors (NPDE).

NPDE is expected to have a normal distribution, therefore, distribution was tested using the Shapiro Wilk test at a level of $\alpha = 0.05$; NPDE was deemed to be non-normally distributed if $p < 0.05$.

3.1.9.6. Simulation of altered dosing intervals

Using model parameters estimated from the final popPK model for each drug, simulations of altered dosing intervals were then carried out. Simulations were carried out in R v.4.0.5[195]. The `mvrnorm` function from the MASS[196] package was used in order to simulate random distributions of each of the model parameters, and the ggplot2[197] package was used to visualise simulations. All other analyses were performed using the base R package. For each drug, a total of 10,000 patients were simulated to receive doses over the first 12 weeks of treatment, as per the BRAGGSS-PD study protocol. Simulations for the median, 5th percentile and 95th percentiles of the population were plotted to show population trend and variability in the population.

Simulations of altered dosing intervals (with the same dose of pre-filled syringe as usually prescribed) were also carried out using the same methods. For Amgevita, dosing intervals of 7 days, 14 days (usual dosing interval) and 21 days were simulated. For Benepali, dosing intervals of 5 days, 7 days (usual dosing interval) and 10 days were simulated.

3.1.9.7. Simulation of drug response

For Amgevita patients, further PD simulation was able to be performed using parameters defined by Ternant *et al* from trial data of RA patients commencing adalimumab (the originator medication of Amgevita)[180]. Changes in CRP and DAS28 as the outcome measures linked to changes in Amgevita concentrations were simulated. CRP was described with an indirect response model with inhibition of CRP input by Amgevita plasma concentration. A direct E_{max} inhibitory model was used to describe the relationship between DAS28 as an outcome measure and Amgevita serum concentrations. These relationships are summarised in Figure 3.3.

For the Amgevita model, input parameters for K_{in} and K_{out} (rate constants for synthesis and degradation, respectively) for the CRP model were taken from Ternant *et al*[180], as were the values for population baseline DAS28 (DAS_0) and the Amgevita concentration leading to a 50% decrease of baseline DAS28 (IC_{50}). Therefore, K_{in} was set at 22 mg/L/day, K_{out} was set at 0.875/day, DAS_0 was set at 5.5 and IC_{50} was set at 11.0 mg/L. The remaining parameters were based on the final popPK model for Amgevita developed in Monolix using BRAGGSS-PD subject data, as described in previous sections. CRP and DAS28 values were simulated using dosing intervals of 7, 14 and 21 days for Amgevita.

For Benepali patients, further PD simulation was carried out using parameters defined by Hsu and Huang from a meta-analysis of RA patients commencing etanercept (the originator medication of Benepali)[176]. Only changes in DAS28 as the outcome measure linked to Benepali concentrations were simulated; a model for CRP was not available in the reference paper. A simple inhibitory E_{max} model was used to describe the relationship between DAS28 as an outcome measure and Benepali serum concentrations. This relationship is summarised in Figure 3.4.

For the Benepali model, input parameters for E_0 (population baseline DAS28), E_{max}, and the cumulative AUC that produced half of E_{max} (AUC_{50}) for the DAS28 model were taken from Hsu and Huang[176]. Therefore, E_0 was set at 6.22, E_{max} was set at 2.89 and AUC_{50} was set at 2,440 μg*hr/ml. The remaining parameters were based on the final popPK model for Benepali developed in Monolix using BRAGGSS-PD subject data, as described in previous sections. DAS28 values were simulated using dosing intervals of 5, 7 and 10 days for Benepali.

Simulations were carried out in R v.4.0.2. The RxODE[198] package was used to implement the ordinary differential equations (ODEs) for these models using various dosing schedules. This package was used for convenience of simulation in the Benepali models, as ODEs were not required, but it was used to implement ODEs in the Amgevita model. Cumulative AUC (Figure 3.4) was calculated at each time point in the Benepali model using the `cumtrapz` function of the pracma[199] package. Visualisations were carried out using the ggplot2 package.

Figure 3.3. PK and PD models describing serum Amgevita concentrations in RA patients, adapted from Ternant et al[180].

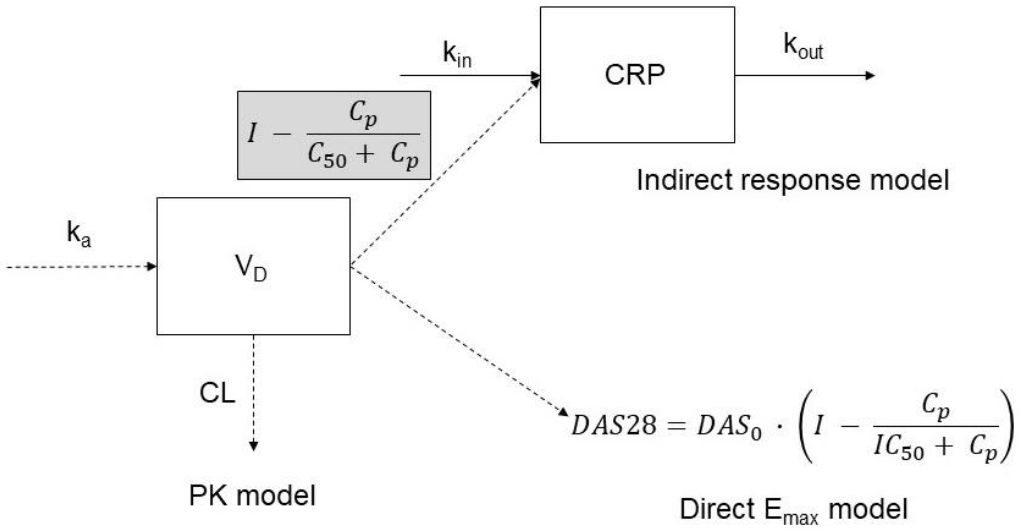
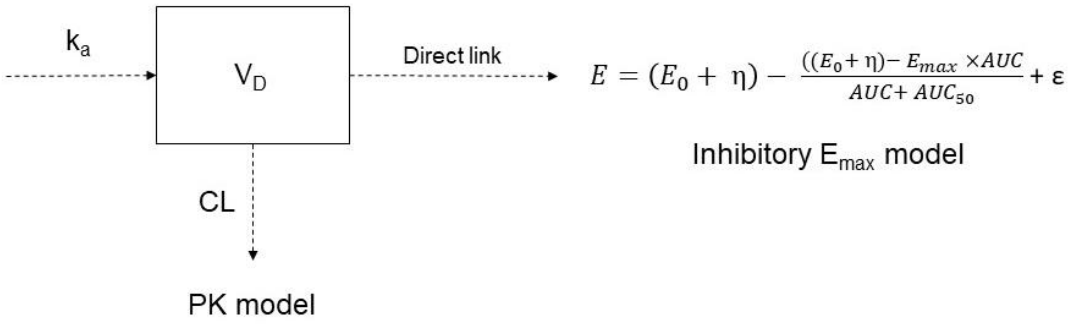


Figure 3.4. PK and PD models describing serum Benepali concentrations in RA patients.



3.2. Discovery proteomics: proteomic predictors of treatment response

3.2.1. Study participants

One cohort of participants consisted of all patients recruited to the BRAGGSS-PD sub-study – please see Section 3.1.2 for more details. Serum protein mapping was carried out at all sampling time points in the BRAGGSS-PD sub-study. In a second cohort, further samples

were selected from pre-existing participants to the prospective arm of BRAGGSS who were commencing etanercept or an etanercept biosimilar from across all 60 participating sites. Serum samples were obtained at baseline (pre-treatment) and after three months of treatment for all participants, and also at six months for 82 participants. In addition, HCs were recruited from the National Repository Study (REC reference 99/8/084), a study consisting of healthy volunteers to provide samples for comparison cohorts and protocol, technique and method development. Samples for HCs were only available at one time point. Inclusion and exclusion criteria were as detailed in Section 3.1.2. Clinical assessments were as detailed in Section 3.1.3. Samples were processed as detailed in Section 3.1.5.

Data cleaning was carried out by the author. For any missing clinical variables (e.g. patients not seen at follow-up for a particular time point, patient did not return questionnaire by post), values were imputed at each time point using a random forest algorithm, implemented in the R package MissForest[200]. Please see Section 3.2.5 for a description of methods used to compare the accuracy of imputation techniques. A random forest algorithm was chosen for imputation as it had the best accuracy of imputation of proteomic data, and this method was also chosen for imputation of clinical variables for consistency.

3.2.2. Sample preparation

Sample preparation and SWATH-MS for discovery proteomics was carried out at the Stoller Biomarker Discovery Centre (SBDC), the University of Manchester, UK. Two datasets of MS data were generated by SBDC:

1. A longitudinal dataset of protein expression over the first 12 weeks of treatment with either Amgevita (adalimumab biosimilar) or Benepali (etanercept biosimilar) in patients recruited to the BRAGGSS-PD sub-study; this was processed in two batches.
2. A dataset of extant samples from patients treated with etanercept/etanercept biosimilars from the wider BRAGGSS cohort; this was processed in three batches. The third batch was processed at the same time as some of the BRAGGSS-PD samples.

Samples were transferred from CfMR to SBDC on dry ice. Following thorough thawing, serum samples were depleted of abundant proteins (e.g. albumin, IgA, IgG, IgM). Within each batch, all samples were plated in a random order and control samples were used at the beginning, during and at the end of MS runs in order to detect run-order effects. The BRAGGSS-PD and third batch of etanercept cohort samples were depleted of the top 14

most abundant proteins using commercially-available HighSelect™ Top14 Abundant Protein Depletion Mini Spin Columns (Thermo Fisher Scientific, Waltham, MA, USA), according to the manufacturer's methods. The remaining BRAGGSS etanercept samples were depleted of the top 12 most abundant proteins using commercially-available Top-12 kits (Pierce, Thermo, Loughborough, UK), according to the manufacturer's methods. Different immunodepletion kits were used as the SBDC protocol had been updated before the later batches of samples were processed.

Protein amount was then assayed in the resultant solution using a Bradford reagent (Bio-Rad, Watford, UK). Solution containing 40µg of protein was subsequently processed further. The BRAGGSS-PD and third batch of etanercept cohort samples were reduced, alkylated and digested by using S-trap columns prior to lyophilisation. For the remaining etanercept cohort samples, samples were reduced using 60 mM tris (2-carboxyethyl) phosphine at 60^{o}C for 60 minutes, and then were alkylated using 10 mM iodoacetamide for 30 minutes in the dark. Protein digestion was carried out overnight using trypsin (Promega, Southampton, UK) at 37^{o}C in a 10:1 ratio of protein-to-enzyme.

For all samples, once digested, peptides were cleaned using a SepPak (Waters, Wilmslow, UK) 96-well plate solid-phase extraction system. Following digestion, gel electrophoresis was then carried out in order to assess the efficiency of immunodepletion and digestion, and to identify any discrepancies or handling errors during prior sample processing. Each gel was run using a Mini Gel Tank containing Bolt 4 – 12% Bis-Tris gel and SDS Running Buffer (all Thermo Fisher Scientific, Waltham, MA, USA). Tanks were supplied with a PowerEase® 90W Power Supply (also Thermo Fisher). Gels were assessed for adequacy of digestion: digested proteins should be small enough to run to the end of the gel.

3.2.3. Bespoke RA protein library generation

A library of proteins associated with RA as determined from previous studies was generated following a detailed literature search and review by the author. The Ovid MEDLINE database was searched from 1946 until October week 5 2018 using the following search terms:

- "Rheumatoid arthritis" as a keyword; "Arthritis, Rheumatoid" as a subject heading (all subheadings included) – search **1**.
- "Proteomics" as a keyword; "Proteomics" as a subject heading (all subheadings included) – search **2**.

- "Proteins" as a keyword (all subheadings included) – search **3**.
- Combine searches **2** OR **3** – search **4**.
- Combine searches **1** AND **4** – search **5**.

Studies were included if they were carried out in human participants with RA, with at least 10 participants in each comparison group. The full text of each manuscript had to be available via the University of Manchester Library. In addition, only proteins that were specifically identified were included; spectral peaks alone were excluded. Review articles, studies utilising only cell lines, clinical trials and studies of DNA and RNA were also excluded. A library of proteins from the remaining included studies was compiled, and identifier (ID) numbers were allocated following a search of the Universal Protein Resource (UniProt)[201] database.

In addition to proteins identified from the above literature search, proteins involved in the TNF pathway according to the Kyoto Encyclopedia of Genes and Genomes (KEGG)[202] were included.

3.2.4. SWATH-MS analysis

Samples were analysed by SWATH-MS with a micro-flow LC-MS system, comprising an Eksigent nanoLC 400 autosampler and an Eksigent nanoLC 425 pump, coupled to a SCIEX 6600 TripleTOF mass spectrometer (all equipment listed: SCIEX, Framingham, MA, USA). The LC method consisted of a 120-minute gradient between a buffer A of 98% water, 2% (volume per volume, *v/v*) acetonitrile and 0.1% (weight per volume, *w/v*) formic acid, and a buffer B of 80% acetonitrile, 20% water and 0.1% formic acid. Samples were injected in duplicate. Spectra were acquired with the mass spectrometer in SWATH mode, using the 100 variable window method, with MS2 windows ranging from 399.5 to 1,249.5 *m/z*, with optimised collision energy equations. The MS1 mass range was from 100 to 1,500 *m/z*, with an accumulation time of 0.05 seconds and a cycle time of 2.6 seconds.

Spectral data files were then converted using wiffconverter (SCIEX, Framingham, MA, USA) to mzML format, an open extensible markup language (XML) format that has been collaboratively developed specifically for use in MS data output files[203]. A protein library search was then carried out using OpenSWATH v.2.0.0[204] against both a publically available twin plasma library[205] (version published 5th January 2015) and a bespoke library of proteins associated with RA (as detailed in Section 3.2.3). OpenSWATH results files were then

processed using PyProphet[206], an algorithm developed for targeted proteomics, particularly large-scale data generated from OpenSWATH or DIANA[206] (an alternative set of bioinformatics tools for analysis of SWATH data). Following processing with PyProphet, files were then aligned using the feature alignment script from the MS Proteomics Tools[207] online repository of bioinformatics tools developed to aid analysis of proteomic MS data. A target false discovery rate (FDR) was set at 0.01 at the peptide-spectrum match (PSM) level.

Further data pre-processing was carried out using R v.3.4.1, using Bioconductor v.3.5[208] packages MSstats[209] and SWATH2Stats[210] for downstream processing. Coefficients of variation (CV) were calculated between technical replicates; any samples with a median CV of ≥ 20% were re-run. Data were filtered by m-score (also known as q-value; this constitutes the minimum FDR acceptable in samples). m-scores were generated using the `filter_mscore_fdr` function in the SWATH2stats package; the overall protein FDR target was set at 0.02 and the upper overall peptide FDR limit was set to 0.05. Data were then converted from a feature alignment input to MSstats input using the `convert4MSstats` function in SWATH2stats. MSstats was then used to normalise and summarise protein intensity data, using the `dataprocess` function with default arguments. Protein intensity readings were then log2-transformed and centred and scaled, so that all coefficients were scaled within each batch that was processed. All sample processing and bioinformatics processing was carried out in-house by SBDC up until this point, but all further analysis in subsequent sections was carried out by the author.

3.2.5. Quality control of proteomics data

Further quality control (QC) of proteomics data was carried out in R v.4.0.2, primarily in the dataset generated from extant samples drawn from the wider BRAGGSS collection. Firstly, data readouts from both plasma and bespoke RA libraries were merged for each study participant and duplicate proteins removed, using the tidyverse[211] package. Proteins with near-zero variance across all participants were removed using the caret[212] package. Density plots were generated in order to visually compare mean expression of proteins versus levels of protein missingness i.e. <25% missing, 25-50% missing, 50-75% missing, >75% missing. Heatmaps were generated using the ComplexHeatmap[213] package to visually inspect for any patterns of protein missingness e.g. by time point or treatment response.

Once it had been ascertained that missing proteins were likely missing at random (for example, due to not being captured in a particular SWATH window at a given time point),

missing proteins were imputed. Prior to imputation, accuracy of various imputation methods was carried out on a subset of complete data with no missing values. The `simIm` function of the imputeR[214] package was used to randomly spike 30% missing values into the data subset, before several different methods of imputation were compared for accuracy:

- Lasso regression, a form of linear regression where data values are shrunk towards a central point, such as the mean. This method favours more sparse models with fewer parameters. The `lassoR` function was used in the imputeR package.
- Partial least squares regression, another form of linear regression that is similar to principal components analysis (PCA), where predicted and observed variables are projected into a new theoretical space. The `plsR` function was used in the imputeR package.
- K-nearest neighbours, a machine learning algorithm that assumes similarity of data points in close proximity. The `knn.impute` function was used in the bnstruct[215] package.
- Multiple imputation by chained equations (MICE), a multiple imputation method that assumes that data is missing at random, with prediction of new data based on a regression model. This was carried out using the mice[216] package.
- Random forest, an ensemble machine learning algorithm that utilises multiple decision trees. This was carried out using the MissForest[200] package.

The root mean square error (RMSE) was used to determine the most accurate method of imputation; the lower the value, the better the accuracy. Using this metric, random forest was selected as the imputation method of choice. New datasets were created from the original subset of complete data, with values of 25%, 50% and 75% missing data randomly spiked in. RMSE of random forest imputation across different levels of protein missingness was then compared between these new datasets.

The above random forest algorithm was then used to impute missing values in the wider dataset. This was carried out separately in each batch, at each time point. It was hypothesised that levels of any given protein would change over time, so imputation was carried out separately at each time point in order to increase imputation accuracy; imputed values would not then be affected by increased or decreased expression of the same protein at different time points. Following imputation, further density plots were created for data pre- and post-imputation to visually inspect whether imputation altered protein expression densities. Hierarchical clustering was used to identify and remove any outlying samples. Then, using

the caret package, PCA was carried out to determine whether there was any batch effect in protein expression. Finally, the `ComBat` function in the sva[217] package was used in order to correct for batch effect in the entire imputed dataset.

In the dataset generated from the BRAGGSS-PD sub-study, QC was carried out with imputation using a random forest algorithm as discussed above. Hierarchical clustering was performed to identify and remove any outlying samples, then PCA was carried out to identify batch effects. The `ComBat` function was again used to correct for batch effect in the entire imputed dataset.

3.2.6. Statistical analysis

All analyses were carried out in R v.4.0.2.

3.2.6.1. Differentially expressed proteins between cases and controls

Analysis was carried out in the etanercept sub-cohort drawn from the wider BRAGGSS collection. One of three batches processed by SBDC contained HCs, and this batch was used to carry out analysis of differential expression of proteins between cases and controls. A Welch's t-test was carried out between RA patients (cases) at baseline (pre-treatment) and HCs using the `col_t_welch` function in the MatrixTests[218] package. Proteins with a significance of $p < 0.05$ were retained, meaning that these proteins were significantly increased or decreased in cases compared to controls. These proteins were selected for the next stage of analysis detailed in later sections, in order to reduce the high dimensionality of the dataset by focusing only on proteins that were differentially expressed between patients with active RA and HC, who represented a healthy physiological state.

3.2.6.2. Longitudinal analysis of protein expression in the first 12 weeks of treatment with Amgevita or Benepali

Analyses were carried out in the dataset drawn from the BRAGGSS-PD sub-study. Samples from patients on each drug (Amgevita and Benepali) were both pooled (for therapeutic drug level analysis only) and then analysed separately. In addition to the MS data described above, clinical parameters were also included in the analyses, namely, drug concentrations (the outcome measure or dependent variable), age at inclusion, biological sex, weight (as this could affect CL and hence, drug concentration) and concurrent csDMARD therapy (as this is known to reduce the chance of developing ADAbs). Proteins measured using SWATH-MS were filtered to only significant proteins from the case-control analysis carried out in

Section 3.2.6.1, in order to limit analysis to only proteins differentially expressed in RA patients compared with HC. Linear mixed effects models were developed to investigate any potential associations between measured protein levels and RA disease outcomes. Fixed effects were defined as age, biological sex, weight and concurrent csDMARD therapy, and independent random effects were defined as each patient's identifier and each sampling time point.

Time can be considered as either a fixed or a random effect in a linear mixed effects model. The overall time period of this study is a continuous variable, but proteins have only been measured at pre-determined time points. These sampling points are representative of protein levels over time as part of a pragmatic study design, but to have protein levels measured at every single possible time throughout the study would be unfeasible. Random effects data can be considered to be a sample of all possibilities, just as sampling time points are just a sample of all possible times in this study. Because there are multiple sampling points for each patient included in the study, sampling time point could be considered to be a grouping variable (random effects are usually considered to be grouping factors which are trying to be controlled), with protein expression values compared between patients at each time point. Furthermore, an advantage of assigning sampling time point as a random effect is that it can then be determined how much variation in protein expression can be attributable to this effect.

Drug concentration levels were chosen as the outcome, as opposed to DAS28, because values for this variable were available at all time points, to correspond with protein MS data; DAS28 was only measured at baseline and 12 weeks. In addition, drug concentration levels may represent a more objective clinical outcome measure than DAS28, as values do not rely on subjective, patient-reported outcome measures[52]. Protein expression was regressed against:

1. Therapeutic drug concentration levels as a dichotomous binary variable, defined as:
    a. Between 5 – 8 mg/L for Amgevita, as per Pouw *et al*[219]; this range is based on trough drug levels and all samples in the Pouw study were obtained immediately before administration of the next adalimumab dose, as in the BRAGGSS-PD study.
    b. Between 2.1 – 4.7 mg/L for Benepali, as per Jamnitski *et al*[220]; it was not explicitly stated whether this range is based on trough or random drug levels in the Methods section of the Jamnitsky study.

2. Actual drug concentration levels as a continuous linear variable.

An FDR set at a threshold of 5% was used in order to control for type 1 error. Analyses were carried out using the lme4[221] package. Significance values for each model component were obtained using the lmerTest[222] package.

3.2.6.3. Analysis of protein expression and association with RA disease outcomes following treatment with etanercept

Analysis was carried out in the wider BRAGGSS etanercept sub-cohort in order to determine whether protein expression as determined using SWATH-MS was predictive of RA clinical response outcomes. Analysis was carried out in R v.4.0.2. The R base package was used to carry out linear regression between expression of each significant protein selected from Section 3.2.6.1 and the following continuous RA disease outcomes:

- Primary outcome measures:
    - DAS28, calculated using high-sensitivity CRP as measured at the CfMR; where this was missing, CRP or ESR from each patient's recruiting centre as documented on the CRF was used, and where no value was available at all, CRP was imputed statistically as per Section 3.2.1.
    - Change in DAS28 from baseline (ΔDAS28) – only analysed at time points after treatment i.e. three and six months.
- Secondary outcome measures:
    - TJC.
    - SJC.
    - Patient-reported VAS of global health.
    - High-sensitivity CRP, as measured at CfMR using ELISA, and not with SWATH-MS.

Logistic regression was carried out between expression of each significant protein from Section 3.2.6.1 and the following categorical RA disease outcomes:

- EULAR response.
- DAS28 improvement of > -1.2, defined as the minimally clinically important difference (MCID).

Both univariate analysis and multivariable analysis (with adjustment for age, biological sex and RA disease duration) were carried out for each protein in both linear and logistic

regression analyses. Adjustment for FDR due to multiple testing was carried out using the Benjamini-Hochberg procedure[223]. Significant proteins (following multiple testing adjustment with $p < 0.05$) were then added into a multivariable model with one another and the above confounding covariates. Proteins at baseline were compared with outcomes at 3 and 6 months. Proteins at three months were compared with outcomes at three and six months. Proteins at six months were compared with outcomes at six months. Linear and logistic regression were used, as opposed to linear mixed effects modelling again, because random effects were not shown to significantly contribute to the variance of models from Section 3.2.6.2, indicating that mixed effects models are unlikely to provide additional benefit over conventional linear and logistic regression.

3.2.6.4. Differential expression of proteins over time following treatment with etanercept

Analysis was carried out in the wider BRAGGSS etanercept sub-cohort using R v.4.0.2. This analysis aimed to determine whether the overall protein expression profile of the study population was altered between time points of interest. While patients were stratified by treatment response, this analysis was agnostic to RA clinical response outcomes and focused on protein expression levels between specific time points in the study population.

Significant proteins from Section 3.2.6.1 were selected for analysis. The limma[224] package was used to assess for differential expression of each protein between available time points i.e. between baseline and three and six months, and between three months and six months. Age at baseline, biological sex, RA disease duration and pre-treatment DAS28 were included as confounders in analysis of all dependent variables. The `contrasts.fit` function was used to compute estimated coefficients and standard errors, and the `eBayes` function was used to compute moderated t- and F-statistics and log-odds of differential expression using empirical Bayes moderation of the standard errors towards a common value.

3.2.6.5. Machine learning methods to determine proteomic predictors of treatment response

Analyses were carried out in the dataset drawn from the wider BRAGGSS etanercept sub-cohort. Proteins that were significantly increased or decreased in RA patients compared to HCs were retained from the analysis described in Section 3.2.6.1. A number of machine learning algorithms were used to determine whether proteins at baseline were predictive of treatment response after three months of treatment. The binary classifiers of failure to

achieve MCID in DAS28 and poor EULAR response were used to define treatment non-response. Due to class imbalance in these classification outcomes, a synthetic minority oversampling technique (SMOTE) algorithm was used to create synthetic data for the minority class. Data were partitioned into completely separate training and validation sets using an 80%/20% split, so SMOTE was carried out separately in each training and validation set in order to avoid data leakage influencing model development. SMOTE was carried out using the package smotefamily[225].

Using the mlr[226] package, the following machine learning algorithms were used to develop classification models predictive of treatment response/non-response to etanercept/etanercept biosimilars:

- Penalised regression: this an umbrella term for a family of regression methods (i.e. ridge, lasso, elastic net) that estimate the regression coefficients of a model through the minimisation of the residual sum of squares. The tuning parameters consisted of:
    - α, the elastic net mixing parameter, tuned between zero (ridge regression) and one (lasso regression).
    - s, equivalent to the regularisation parameter λ, tuned between zero and one.

  Penalised regression models can demonstrate improved prediction on new data by shrinking coefficient size and retaining predictors with coefficients greater than zero[227].
- K-nearest neighbours: this method was previously outlined in Section 3.2.5. The tuning parameters consisted of:
    - k, the number of nearest neighbours considered in algorithmic decisions, tuned between two and ten.
    - Distance, the Minkowski distance, which is a generalisation of both the Euclidean and Manhattan distances, tuned between one and three.
    - Kernel, which allows mapping to a high-dimensional feature space, tuned over rectangular, Gaussian, rank and optimal values.

  K-nearest neighbours is an advantageous algorithm, with a quick calculation time, a simple algorithm and good predictive accuracy.
- Random forest: this method was previously outlined in Section 3.2.5. The tuning parameters consisted of:
    - mtry, the number of variables to potentially split at each decision tree node, tuned between three and five.
    - num.trees, the maximum number of decision trees, which was set at 500.

Random forest algorithms have good predictive accuracy due to in-built reduction in over-fitting when building decision trees. However, it is computationally and time-intensive during model training.

- Support vector machine: this method is a non-probabilistic binary linear classifier that functions by mapping training data to points in space, which maximises the distance between two categories. The tuning parameters consisted of:
    - Cost, a function that controls for training errors and margins, tuned between 0.1 and ten.
    - γ, a parameter required for compatibility with different kernels, tuned between 0.1 and ten.
    - Degree, a parameter required for a polynomial kernel, tuned between one and four.
    - Kernel, tuned over polynomial, radial and sigmoid values.

  Support vector machine algorithms are effective in high-dimensional spaces, particularly when dimensions outnumber samples. However, these algorithms struggle when there is not a clear margin of separation between classes.

These different methods all utilise a variety of statistical methods and were chosen to represent algorithms based on regression, clustering and decision trees, as well as an algorithm based on data segregation by hyperplane formation. By using this ensemble of different methods, a broad overview of model training and prediction on the same data could be determined.

During model training, algorithms were tuned using nested cross validation. Inner loops were tuned using 10 folds of cross-validation. Outer loops were tuned over five repeats of 10-fold cross-validation. Parameters were tuned over a random tuning grid, with a maximum number of iterations of 20. Models developed using each algorithm were compared in a benchmarking experiment, and the best model from each benchmarking experiment was taken forward and re-trained on the validation data. Model performance was then tested using area under the ROC curve, accuracy and mean misclassification error (MMCE). Model calibration was assessed via visual inspection of plots generated using the classifierplots[228] and RBPcurve[229] packages, as well as by carrying out the Hosmer-Lemeshow goodness-of-fit test using the `hoslem.test` function of the ResourceSelection[230] package.

3.2.6.6. Pathway analysis

All significant proteins from Sections 3.2.6.2, 3.2.6.3 and 3.2.6.4 were combined into a list and searched in the STRING database v. 11.5[231 232]. The STRING database is an online, interactive repository that aims to collect, appraise and amalgamate all publicly-available information on protein-protein interactions. UniProt IDs of significant proteins as described above were input into the browser-based interface with STRING using the "multiple proteins" search function and filtered to only results from *Homo sapiens*. The resultant search returned information on protein interactions and networks, including both known and predicted interactions, and interactions derived from other online sources.

3.3. Genetics of protein expression in patients with RA

Genotyping was carried out on the same extant BRAGGSS prospective arm etanercept samples included in the proteomic analysis, as detailed in Section 3.2.1. Analysis was not carried out in the longitudinal samples recruited from the BRAGGSS-PD sub-study.

3.3.1. DNA extraction and genotyping

Both DNA extraction and genotyping procedures were carried out at the CfMR by the laboratory technician team. DNA extraction was carried out via standard phenol-chloroform extraction on whole blood, as previously described[233]. Genotyping was carried out using the Illumina Infinium HumanCoreExome 12 BeadChip kit (Illumina, San Diego, California, USA). 200 ng of DNA was used, according to the manufacturer's guidance. Genotype calling was carried out using GenomeStudio software (Illumina, San Diego, California, USA).

3.3.2. SNP and sample QC and imputation

Genetic QC and imputation was carried out by Dr Chuan Fu Yap, research associate at CfMR. Analysis was conducted using PLINK v.1.9[234 235]. Samples with unassigned chromosomes (i.e. "CHR=0"), Y chromosomes and mitochondrial SNPs were initially excluded. Then, samples and SNPs were pre-filtered at a 98% call rate. A threshold of 2% was set for SNP missing rate, and SNPs above this threshold were excluded. A threshold for low allele frequency was set at 1%, and alleles with a minor allele frequency below this threshold were excluded. SNPs that did not fulfil conditions for Hardy-Weinberg equilibrium, set at a p-value of 1E-04, were then excluded. Duplicate SNPs, based on chromosome and base position, were then removed. Autosomal heterozygosity was calculated in order to exclude in-bred, outlier genotypes. Finally, genotype data for the X

chromosome was checked for discrepancies in sex data with corresponding pedigree data. Sex was initially determined using X chromosome heterozygosity rates, then compared to a gender pedigree data file. Mismatched samples were compared with phenotype data, and if a discrepancy persisted, samples were excluded.

The resultant dataset was then aligned to a reference genome, namely the Haplotype Reference Consortium[236]; this step is necessary prior to imputation. Samples were checked for identity by descent to determine sample relatedness, in order to remove potentially identical or duplicate genotype files, as well as related individuals. Because population stratification drives spurious association in any final analysis, the genetic ancestry of each individual was determined using the HapMap3 reference panel[237] by merging the genotype data with the reference panel. Areas in high linkage disequilibrium (LD) were excluded, then PCA was carried out in R v.3.6.1 in order to identify and remove outliers. The results of the PCA were used to ensure that only individuals of European ancestry were retained for analysis. Finally, phasing and imputation were carried out on the Michigan Imputation Server[238]: phasing was performed using Eagle v2.4[239] and the imputation engine was Minimac4[240]. The HRC[241] (Version r1.1 2016) reference panel was used; this panel consists of 64,940 haplotypes of predominantly European ancestry. During post-imputation QC, duplicate SNPs were removed, as were variants with low-quality scores ($R^2 < 0.5$, indicative of poor imputation).

3.3.3. pQTL analysis to determine whether genetic markers are correlated with protein levels

All analyses were carried out by the author using R v.4.0.2. Imputed PLINK files from Section 3.3.2 were input into R using the `read_plink` function of the genio[242] package. The ensembldb[243] package was used to annotate UniProt IDs to genes as documented in the Ensembl[244] database. Duplicate annotations were dropped. Once proteins had been annotated to genes, pQTL analysis was carried out with the Matrix eQTL[245] package using a linear model. The following variables were included as covariates: age at baseline, RA disease duration prior to starting etanercept, biological sex, concurrent csDMARD therapy, BMI and seropositivity to either RF or ACPA. A significance level of $p < 1E\text{-}05$ was set for *cis* pQTLs, and of $p < 5E\text{-}20$ for *trans* pQTLs. The low p-value for *trans* pQTLs was because the study was not designed to be powered to detect *trans* effects, so this significance level was set to limit the number of associations detected. p-values from the pQTL analysis were adjusted for FDR using the Benjamini-Hochberg procedure. pQTLs were sought at baseline and after

three months of treatment with etanercept. dbSNP was used to identify which chromosome each SNP was on[246]. The most associated SNPs for each protein identified were then input into the GTEx Portal[247] in order to identify any corresponding tissue-specific eQTL to pQTL. SNPs of proteins of interest were input into the STRING database[231] to determine whether there were any known interactions between these proteins.

3.3.4. Genetic risk score derivation

Data pre-processing was carried out using R v.4.0.2 by the author. Significant *cis* pQTLs (as derived in Section 3.3.3) at level of $p < 1E-05$ were selected for analysis to derive a genetic risk score for treatment non-response. pQTL results output files from Section 3.3.3 were merged with SNP information present in the .bim file generated in 3.3.2 and used as the base data. The target data consisted of 1,563 patients with RA recruited to BRAGGSS (but independent from patients included in the pQTL analysis in Section 3.3.3) receiving a combination of TNFi, including adalimumab, etanercept and infliximab; a detailed description of this cohort and genotyping methods have previously been described[248]. Polygenic risk scores (PRS) were calculated using the PRSice[249] package. The distance for clumping was set to 250kb, the R^2 threshold for clumping was set to 0.1 and the p-value threshold was set to 1. A logistic regression was then carried out, using the generated PRS as a predictor of EULAR non-response after three or six months of treatment as the target trait, adjusting for the covariates of biological sex, baseline DAS28 and concurrent csDMARD therapy; age at baseline was not available in this dataset. A significance level of $p < 0.05$ was set to determine model fit.

3.4. Chapter summary

This chapter outlined methodology for the following planned analyses:

- popPK studies of patients initiating Amgevita or Benepali, with subsequent modelling and simulation of altered dosing intervals.
- Acquisition of proteomics data using SWATH-MS and subsequent data QC.
- Analysis of proteomics data to determine predictors of treatment response and differential expression profiles.
- pQTL analysis using protein expression data obtained at SWATH-MS combined with genotype data from the same patients.
- Genetic risk score calculation using the results of the pQTL analysis.

CHAPTER FOUR: popPK MODELLING AND SIMULATION OF AMGEVITA AND BENEPALI

Summary of chapter contents:

4.1. Results

4.1.1. Development of a popPK model for patients initiating Amgevita

Ten patients with RA who commenced Amgevita were recruited to the BRAGGSS-PD sub-study. Nine patients were female and one patient was male. All patients were Caucasian. The median age was 50.5 years (interquartile range, IQR, 46 – 61) and the median pre-treatment DAS28 was 5.71 (IQR 5.20 – 6.09). Detailed patient characteristics are outlined in Table 4.1.

Table 4.1. Patient characteristics at baseline, prior to treatment with Amgevita.

Characteristic	**Statistic**
Female sex, n (%)	9 (90.00)
Age (years), median [IQR]	50.5 [46 – 61]
Body weight (kg), median [IQR]	85.5 [66 – 111]
Concurrent csDMARD, n (%)	8 (80.00)
DAS28, median [IQR]	5.71 [5.20 – 6.09]

ABBREVIATIONS: Conventional synthetic disease-modifying anti-rheumatic drug (DMARD), disease activity score of 28 joint counts (DAS28), interquartile range (IQR).

A total of 58 serum samples of Amgevita drug concentrations were available for analysis. As outlined in Section 3.1.9.2, a one-compartment PK model was found to be sufficient to describe the popPK profiles of patients initiating Amgevita recruited from the BRAGGSS-PD sub-study. A combined additive and proportional model was used to describe the residual errors in the data. No covariate was found to be significant in the model, as covariates tested only demonstrated a modest improvement in AIC and -2LL, whilst complicating the model built from sparse sampling. Furthermore, due to a large residual error from estimation of ka, this value was fixed to 0.01167 $hour^{-1}$ as per Ternant *et al*[180] and the random effect (BSV) was not estimated on this parameter.

Plots were generated by Monolix of predicted versus observed measurements for Amgevita serum concentrations, demonstrating that PK parameters were able to describe the data (Figure 4.1). PK parameters estimated are presented in Table 4.2.

Table 4.2. PopPK parameter estimates for Amgevita.

Parameter (units)	**Definition**	**Estimate**	**Relative standard error (RSE, %)**
V_D (L)	Apparent volume of distribution	9.19	12.7
CL (L/hr)	Apparent clearance	0.00283	23.3
ka (/hr)	Rate constant for absorption	0.1167	**Fixed**
ω_{VD} (%)	Coefficient of variation (CV) of between-subject variability (BSV) on V_D	15.60	141.0
ω_{CL} (%)	CV of BSV on CL	68.90	24.8
σ_{prop} (%)	Standard deviation (SD) of proportional residual error	26.00	15.7
σ_{add} (mg/L)	Standard deviation of additive residual error	10.80	16.0

The relative standard error (RSE, %) was calculated as: RSE = (estimate / standard error) x 100.

ABBREVIATIONS: Additive error (add), between-subject variability (BSV), clearance (CL), coefficient of variation (CV), proportional error (prop), proportional rate constant for absorption (ka), standard deviation (SD), volume of distribution (V_D).

All diagnostic plots were obtained from the final popPK model for Amgevita. Population weighted residuals (PWRES) and individual weighted residuals (IWRES), as well as normalised prediction distribution error (NPDE) plots demonstrated no gross model misspecification (Figure 4.2). Residuals were shown not to be normally distributed using the Shapiro Wilk test, with p-values of 4.08E-05, 1.26E-05 and 2.93E-07 for IWRES, PWRES and NPDE, respectively. However, on this basis alone, it is not sufficient to reject a model, and estimates from this model had values similar to those obtained by Ternant *et al*[180] and relative standard errors (RSE) were mostly satisfactory; the RSE for ω(%) for V_D was high, but the estimate for the actual value was in-keeping with what was expected from prior knowledge. Furthermore, a visual predictive check (VPC) revealed adequate model fit (Figure 4.3), with only a very small area of outlying predictions. Therefore, parameter estimates from this popPK model were taken forward and utilised in simulations of

alternative dosing intervals for Amgevita, as well as simulations of how Amgevita treatment might influence CRP and DAS28 values in a typical patient during the first 12 weeks of treatment.

4.1.2. Simulation of alternative dosing intervals for Amgevita

Initially, a simulation of 10,000 individuals was carried out using PK parameters obtained from the popPK model derived in Section 4.1.1. Typical individual profiles, the median and 5^{th} and 95^{th} percentiles were simulated using the usual dosing interval of Amgevita 40mg every 14 days. Simulated values were overlaid with measured values from the study subjects included in the popPK model, and this demonstrated visually that simulated values agreed well with measured values (Figure 4.4). Figure 4.4 differs from Figure 4.3 in that Figure 4.3 is based on actual dosing history and sampling times of BRAGGSS-PD patients used for the modelling. However, Figure 4.4 is based on nominal dosing history (Amgevita 40mg every 14 days) alongside the sampling times in the observed data. Therefore, Figure 4.3 is a true reflection of the data when compared with Figure 4.4.

Once it had been established that simulations agreed with actual measured values, alternative dosing regimens were additionally simulated for typical individuals. These alternative regimens were Amgevita 40mg administered every 7 or every 21 days, alongside the usual dosing regimen of every 14 days (Figure 4.5). Both the usual dose rates of 40mg every 14 days and the increased dose rate of 40mg every 7 days achieved steady-state drug concentrations within the therapeutic window of adalimumab, defined by Pouw *et al* as between 5 – 8 mg/L[219]. However, the reduced dose rate of 40mg every 21 days from initiation of treatment did not achieve steady-state concentrations within this therapeutic window. Time to reaching steady-state drug concentrations had negligible difference between the administration of Amgevita 40mg every 7 or 14 days.

Figure 4.1. Observed values of Amgevita concentrations versus population model-predicted values (PRED) and individual predicted values (IPRED).

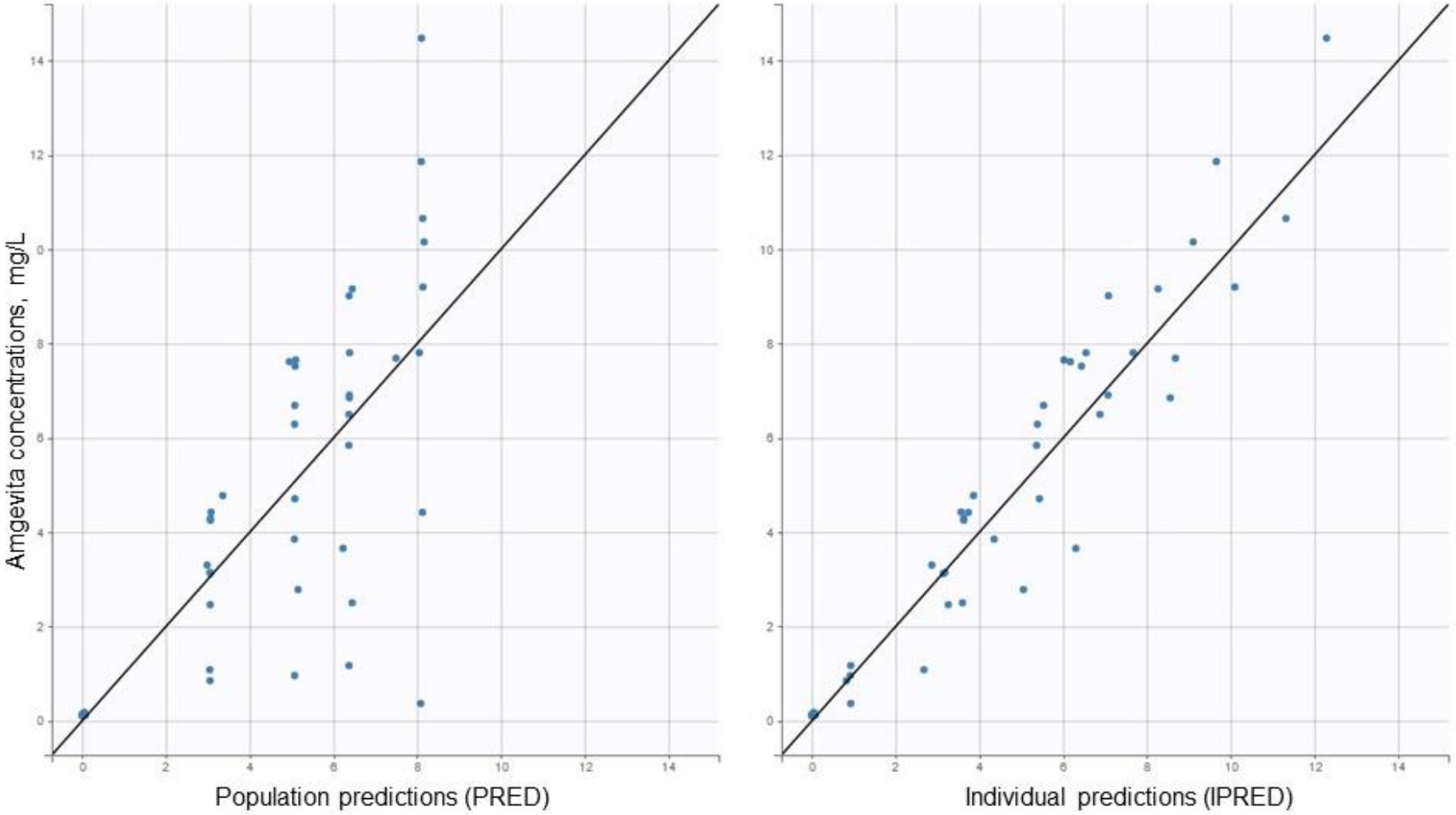


Figure 4.2. Distribution of population (PWRES) and individual weighted residuals (IWRES) versus individual predictions and normalised prediction distribution error (NPDE) for Amgevita concentrations.

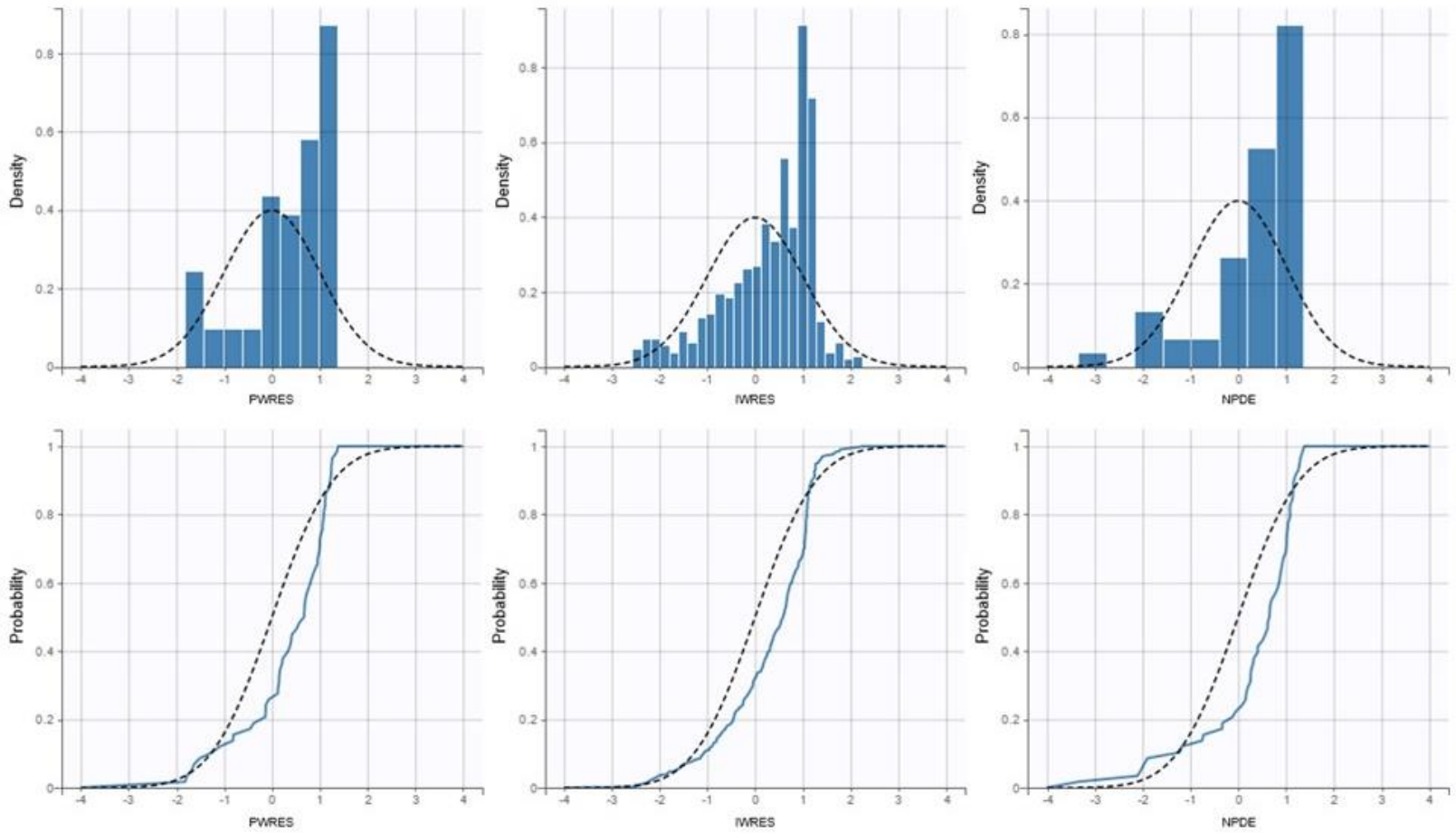
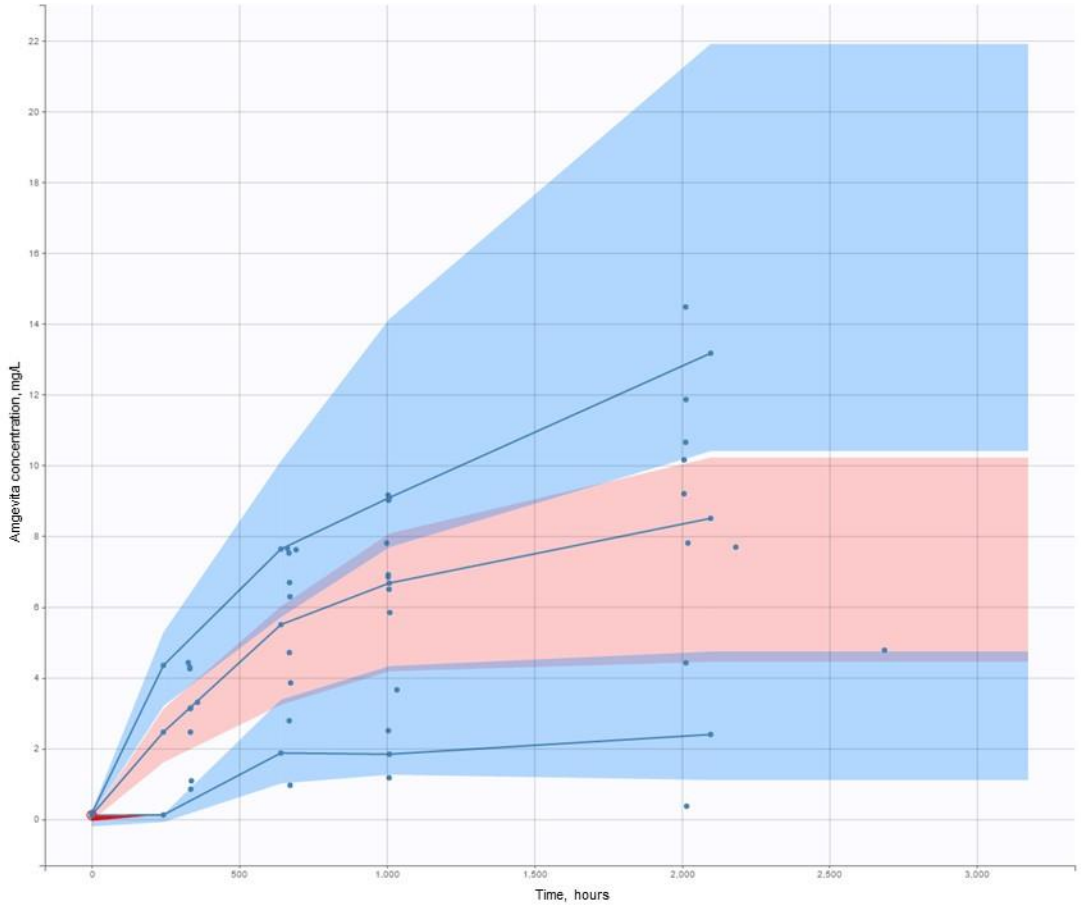
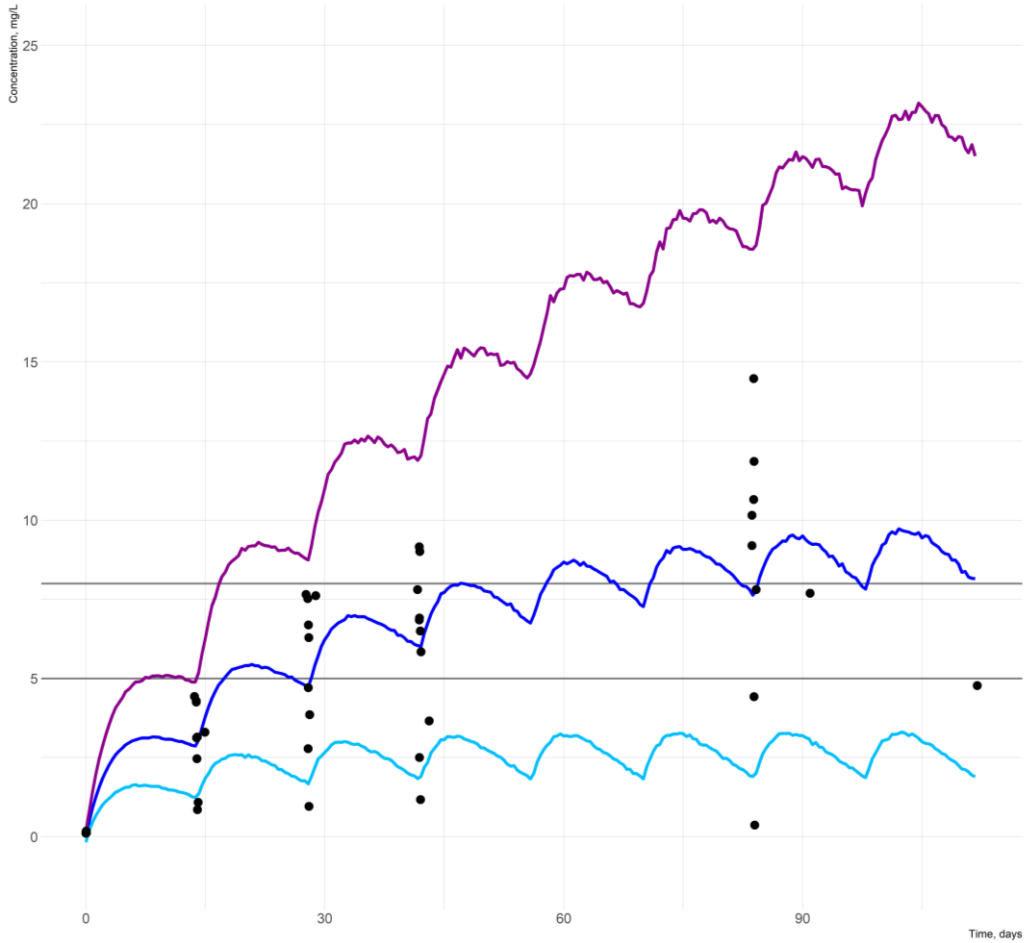


Figure 4.3. Visual predictive check (VPC) of popPK model fit for Amgevita.



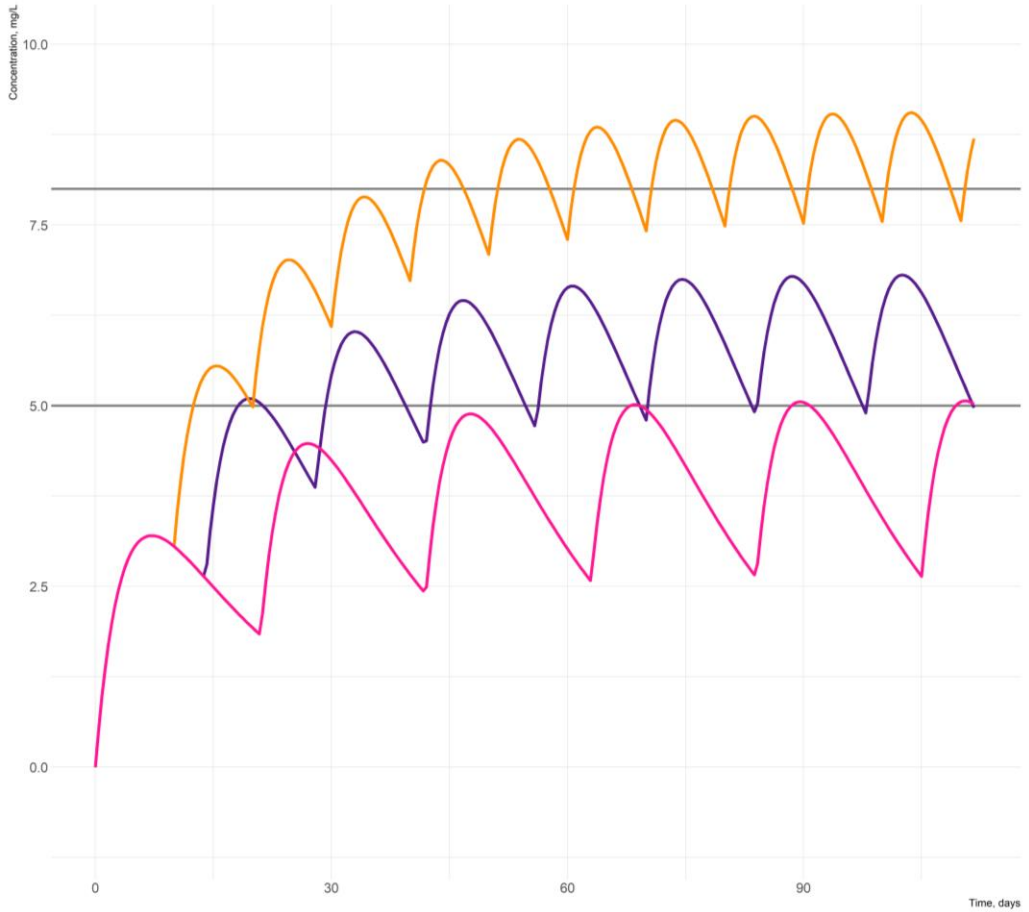
INTERPRETATION OF VISUAL PREDICTIVE CHECK: The light red zone represents a simulation-based 95% confidence interval (CI) around the median Amgevita concentration for the population, which is denoted by the middle solid blue line. The 10^{th} and 90^{th} percentiles are represented by the lower and upper solid blue lines, respectively, and their 95% CI are represented by the surrounding light blue areas. Outliers are highlighted with the bright red area. Solid blue circles are the actual Amgevita concentration values of the sample population.

Figure 4.4. Simulation of Amgevita PK profile for 10,000 individuals using the final popPK model developed for this drug.



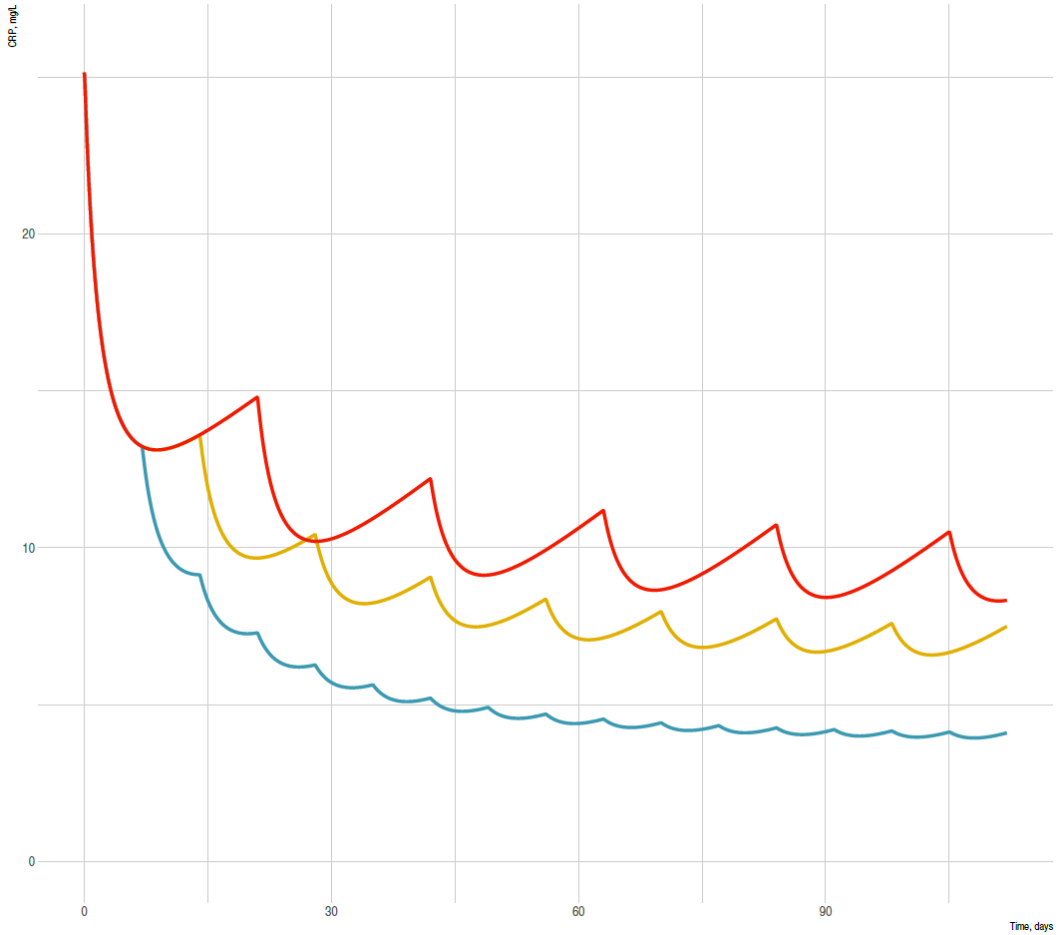
LEGEND: Horizontal grey lines represent the Amgevita therapeutic window of 5 – 8 mg/L proposed by Pouw *et al*[219]. The dark blue line represents the population median, the light blue line represents the 5th percentile and the purple line represents the 95th percentile of the population. The black dots represent actual population drug concentration values.

Figure 4.5. Simulated usual and alternative dosing intervals of Amgevita for a typical patient with different dose rates.



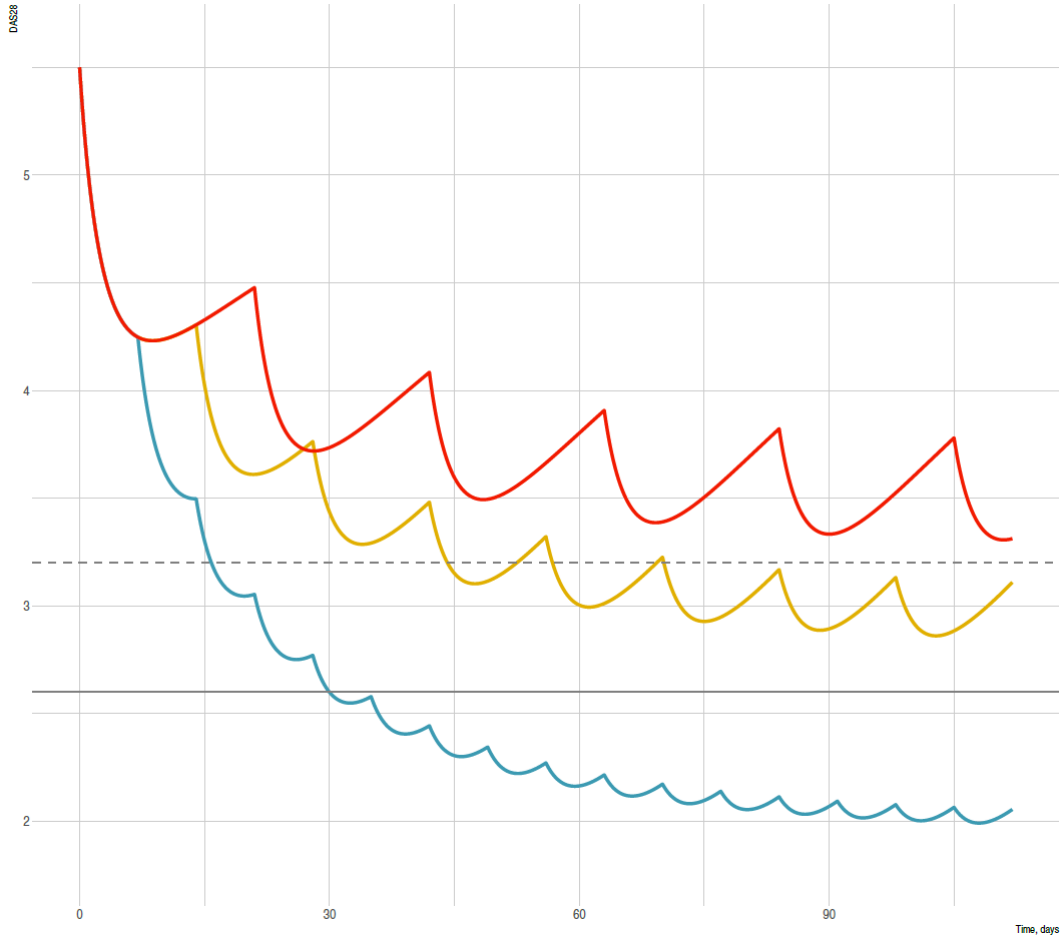
LEGEND: Horizontal grey lines represent the Amgevita therapeutic window of 5 – 8 mg/L proposed by Pouw *et al*[219]. The purple line represents the usual Amgevita dosing interval of 40mg every 14 days, the orange line represents a dosing interval of 40mg every 7 days and the pink line represents 40mg every 21 days.

Figure 4.6. Simulated PD profiles of alternative dosing regimens of Amgevita with CRP as the outcome measure.



LEGEND: The yellow line represents the usual dosing regimen of Amgevita 40mg every 14 days, the blue line represents dosing every 7 days and the red line represents dosing every 21 days.

Figure 4.7. Simulated PD profiles of alternative dosing regimens of Amgevita with DAS28 as the outcome measure.



LEGEND: The yellow line represents the usual dosing regimen of Amgevita 40mg every 14 days, the blue line represents dosing every 7 days and the red line represents dosing every 21 days. The horizontal dashed grey line represents a DAS28 of 3.2 (low disease activity) and the horizontal solid grey line represents a DAS28 of 2.6 (disease remission).

4.1.3. Simulation of PD models for patients with RA initiating Amgevita

An indirect response model of CRP as the outcome measure of response to treatment with Amgevita was successfully simulated for the usual dosing regimen of 40mg every 14 days, as well as for 40mg every 7 and 21 days; simulations are presented in Figure 4.6. Only a regimen of 40mg every 7 days was expected to reduce CRP levels to a level below 5mg/L, although improvement was dramatic in the first two weeks of treatment with all three regimens. A CRP level of 5mg/L is the lower limit of detection for most clinical CRP measurements and, as such, a measurement below this level represents absent systemic inflammation.

A direct E_{max} inhibitory model of DAS28 as the outcome measure of response to treatment with Amgevita was successfully simulated for the usual dosing regimen of 40mg every 14 days, as well as for 40mg every 7 and 21 days; simulations are presented in Figure 4.7. A regimen of 40mg every 7 days successfully simulated reaching steady-state DAS28 levels that would be classified as remission. A regimen of 40mg every 14 days (the usual dosing regimen) reached steady-state DAS28 levels that would be classified as low disease activity, but not remission. A regimen of 40mg every 21 days reached steady-state DAS28 levels that would be classified as a moderate EULAR response, but still with moderate disease activity, indicating ongoing systemic inflammation inadequately controlled by immunosuppressive medication.

4.1.4. Development of a popPK model for patients initiating Benepali

Six patients with RA who commenced Benepali were recruited to the BRAGGSS-PD sub-study. Four patients were female and the remaining two patients were male. One patient was West African and the remaining patients were Caucasian. The median age was 57.5 years (IQR 56 – 59) and the median pre-treatment DAS28 was 5.33 (IQR 4.96 – 5.58). Detailed patient characteristics are outlined in Table 4.3.

Table 4.3. Patient characteristics at baseline, prior to treatment with Benepali.

Characteristic	**Statistic**
Female sex, n (%)	4 (66.67)
Age (years), median [IQR]	57.5 [56 – 59]
Body weight (kg), median [IQR]	70.5 [69 – 84]
Concurrent csDMARD, n (%)	4 (100.00) [2 missing]
DAS28, median [IQR]	5.33 [4.96 – 5.58]

ABBREVIATIONS: Conventional synthetic disease-modifying anti-rheumatic drug (DMARD), disease activity score of 28 joint counts (DAS28), interquartile range (IQR).

A total of 40 samples of Benepali drug concentrations were available for analysis. As outlined in Section 3.1.9.2, a one-compartment PK model was found to be sufficient to describe the popPK profiles of patients initiating Benepali recruited from the BRAGGSS-PD sub-study. A combined additive and proportional model was used to describe the residual error in the data during modelling. No covariates were included in the model, as these only demonstrated a modest improvement in AIC and -2LL, whilst complicating the model developed from only sparse sampling. Furthermore, due to a large residual error from estimation of ka, this value was fixed to 0.0396 $hour^{-1}$ as per Korth-Bradley *et al*[250] and the random effect (BSV) was not estimated on this parameter. Additionally, the model had a large %RSE for estimated V_D, so again, the random effect (BSV) was removed from this parameter. Finally, the additive error SD was fixed at 0.0001 in order to ensure model stability.

Plots were generated by Monolix of predicted versus observed measurements for Benepali serum concentrations, demonstrating that PK parameters were able to describe the data (Figure 4.8). PK parameters were estimated and are detailed in Table 4.4.

Figure 4.8. Observed values of Benepali concentrations versus population model-predicted values (PRED) and individual predicted values (IPRED).

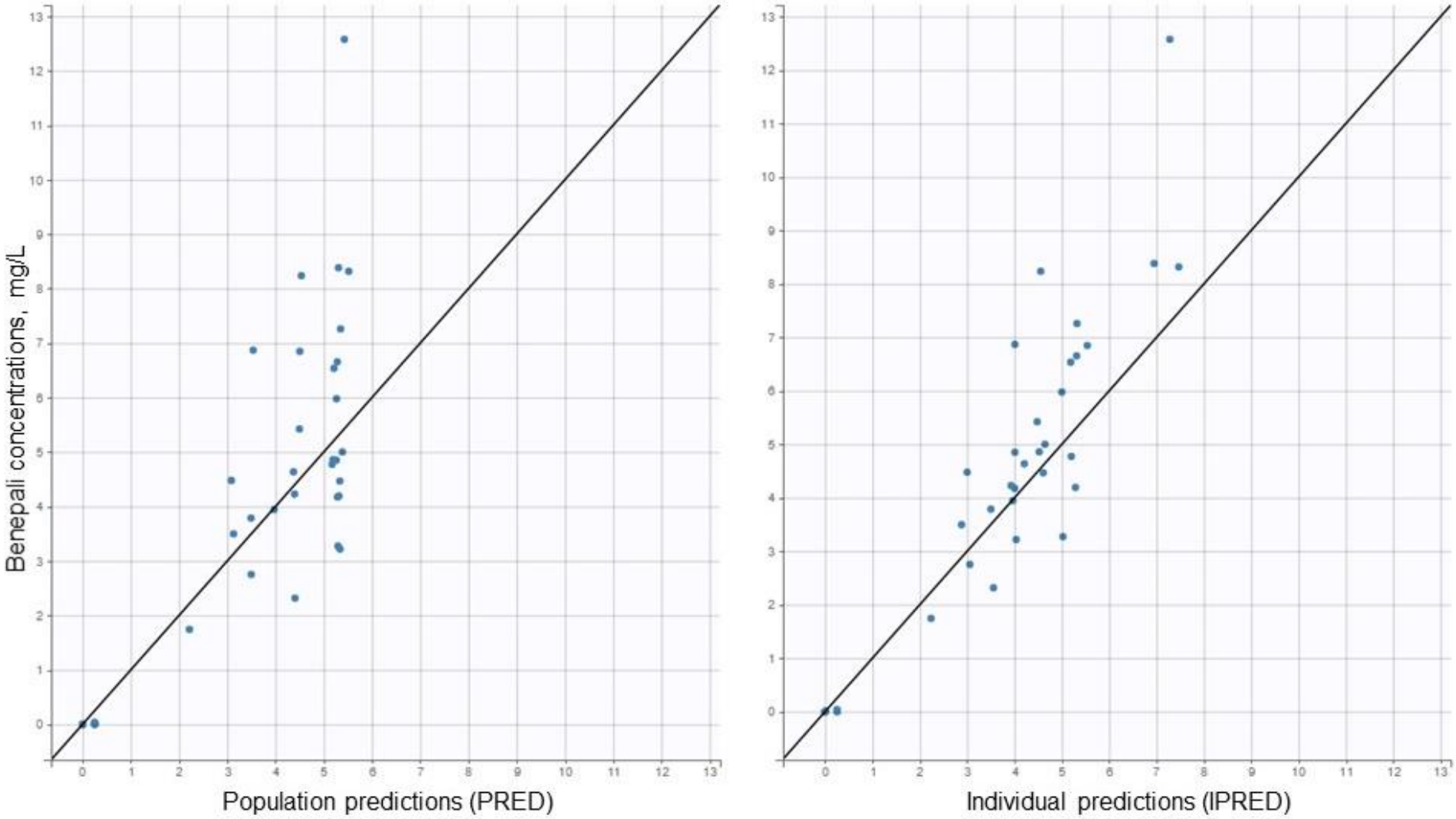


Table 4.4. PK parameter estimates for Benepali popPK model.

Parameter (units)	**Definition**	**Estimate**	**Relative standard error (RSE, %)**
V_D (L)	Apparent volume of distribution	7.76	18.2
CL (L/hr)	Apparent clearance	0.0404	10.7
ka (/hr)	Rate constant for absorption	0.0396	**Fixed**
ω_{CL} (%)	Coefficient of variation (CV) of between-subject variability (BSV) on CL	0.173	61.0
σ_{prop} (%)	CV of proportional residual error	0.46	13.6

The relative standard error (RSE, %) was calculated as: RSE = (estimate / standard error) x 100.

ABBREVIATIONS: Between-subject variability (BSV), clearance (CL), coefficient of variation (CV), proportional error (prop), rate constant for absorption (ka), volume of distribution (V_D).

All diagnostic plots were obtained from the final popPK model for Benepali. PWRES and IWRES, as well as NPDE plots, demonstrated no gross model misspecification (Figure 4.9). Residuals for IWRES were shown not to be normally distributed using the Shapiro Wilk test, with a p-values of 9.77E-03. However, PWRES and NPDE residuals were normally distributed, with non-significant p-values following Shapiro Wilk testing, where the null hypothesis assumes a normal distribution. A VPC revealed adequate model fit (Figure 4.10), with two very small areas of outlying predictions. Therefore, parameter estimates from this popPK model were taken forward and utilised in simulations of alternative dosing intervals for Benepali.

Figure 4.9. Distribution of population (PWRES) and individual weighted residuals (IWRES) versus individual predictions and normalised prediction distribution error (NPDE) for Amgevita concentrations.

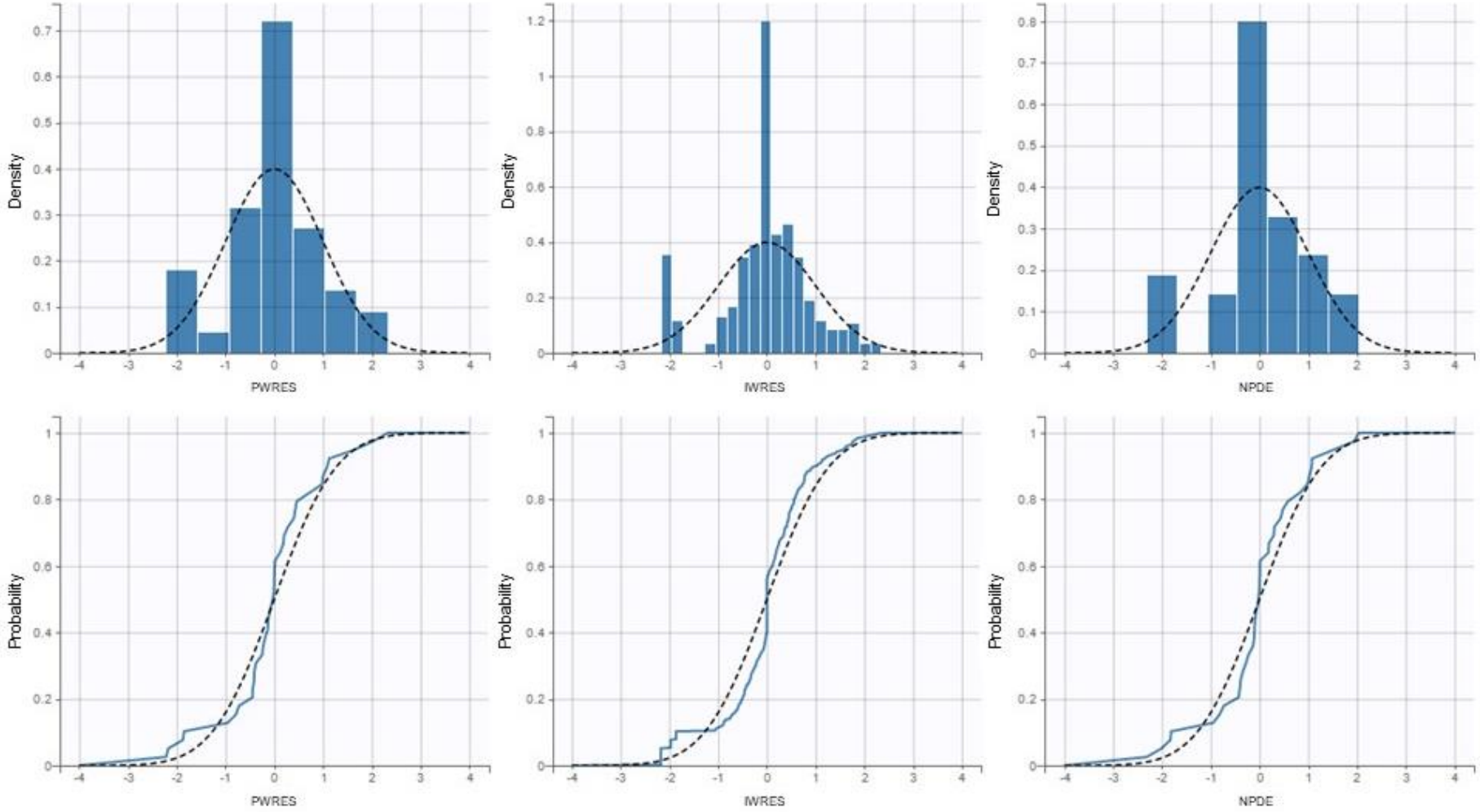
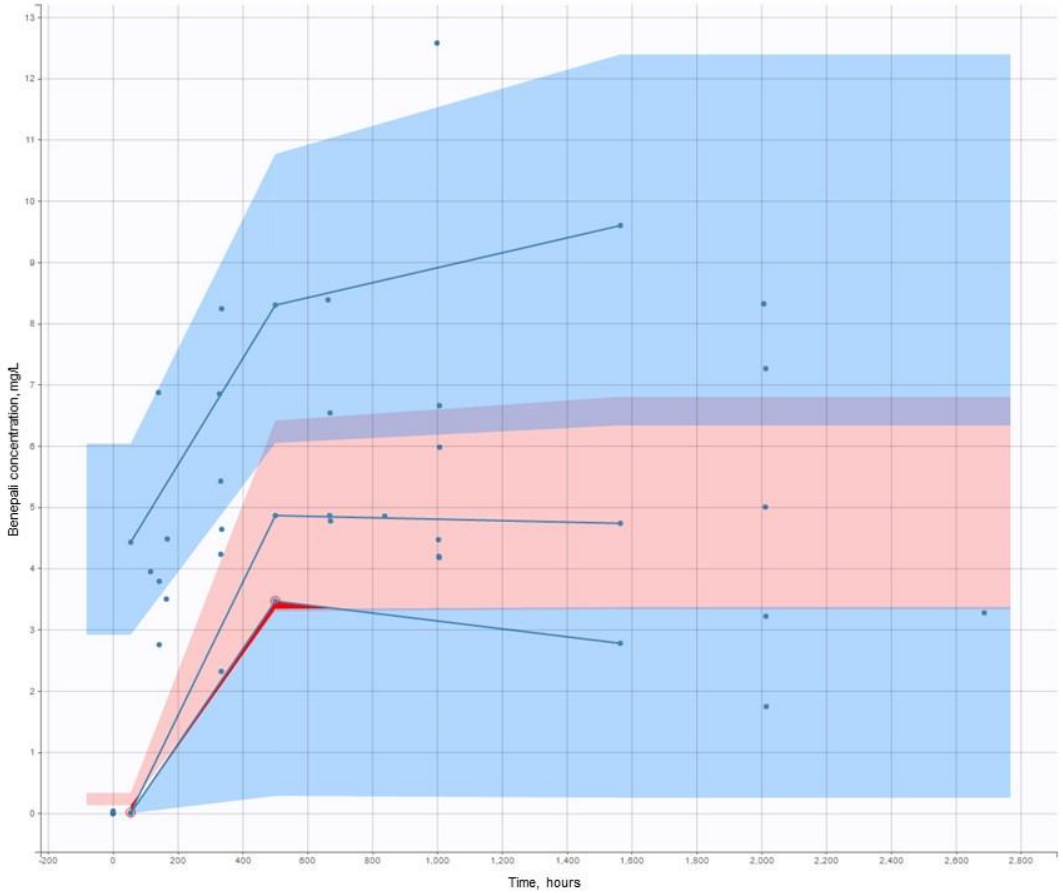


Figure 4.10. VPC of popPK model fit for Benepali.



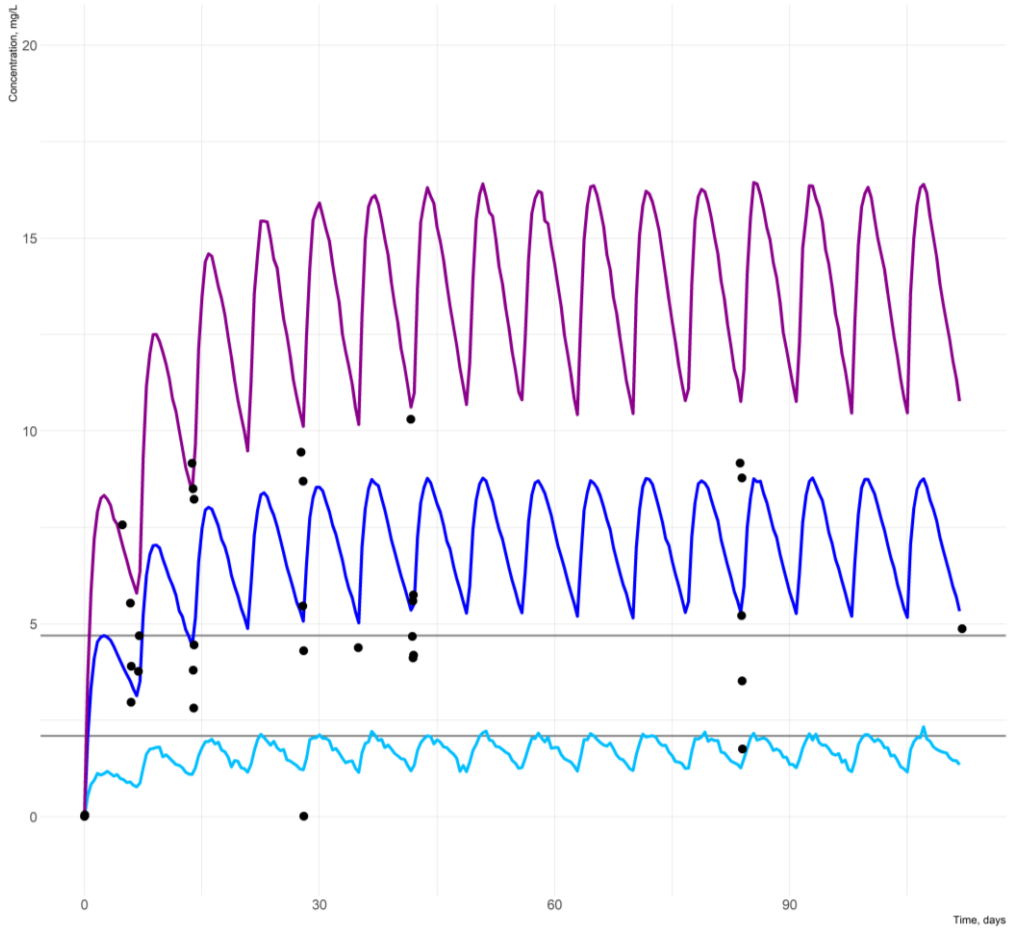
INTERPRETATION OF VISUAL PREDICTIVE CHECK: The light red zone represents a simulation-based 95% CI around the median Benepali concentration for the population, which is denoted by the middle solid blue line. The 10th and 90th percentiles are represented by the lower and upper solid blue lines, respectively, and their 95% CI are represented by the surrounding light blue areas. Outliers are highlighted with the bright red areas. Solid blue circles are the actual Benepali concentration values of the sample population.

4.1.5. Simulation of alternative dosing intervals for Benepali

Initially, a simulation of 10,000 individuals was carried out using PK parameters obtained from the final popPK model derived in Section 4.1.4. Typical profiles for the median, 5^{th} and 95^{th} percentiles were simulated using the usual dosing interval of Benepali 50mg every 7 days. Simulated values were overlaid with measured values from the study subjects included in the popPK model, and this demonstrated that simulated values correlated well with actual measured values (Figure 4.11). Figures 4.10 and 4.11 are different in that simulations in Figure 4.10 were based on practical dosing records that were obtained from BRAGGSS-PD patients, whereas a nominal dosing record of Benepali 50mg every 7 days was used for all patients in Figure 4.11

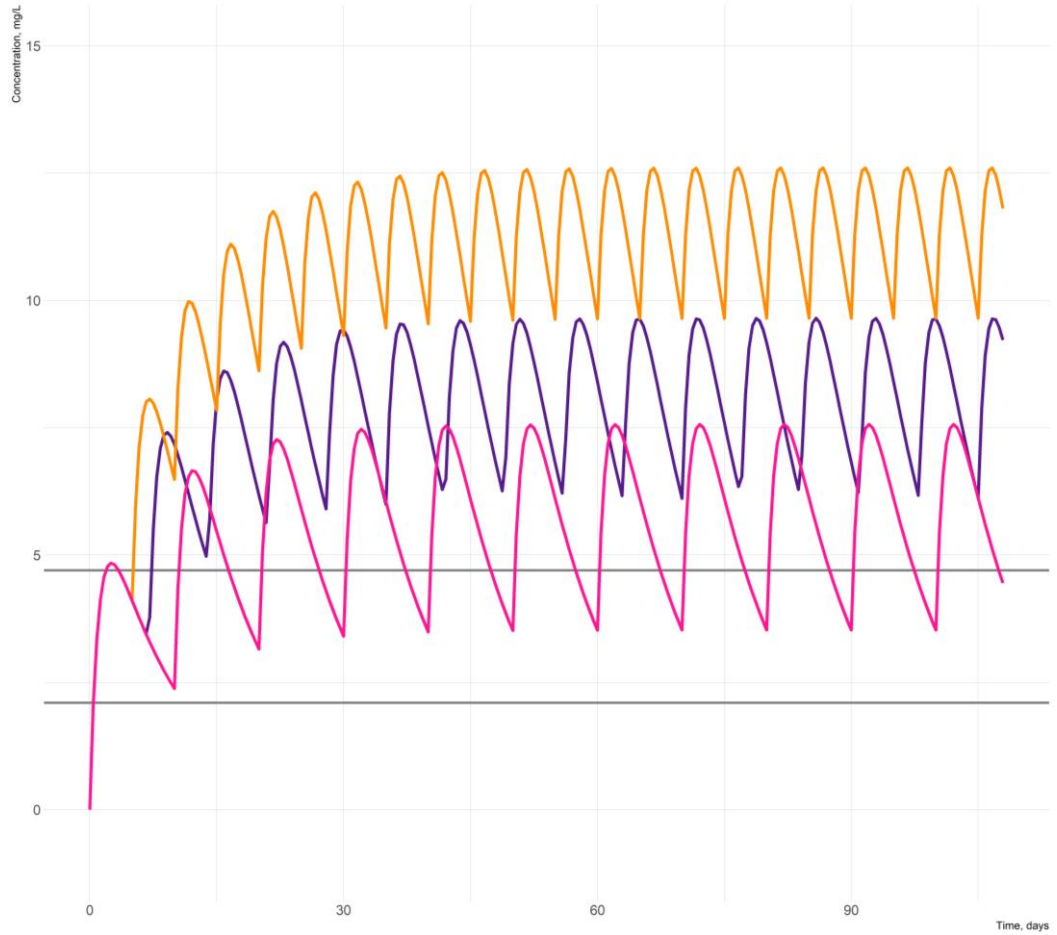
Once it had been established that simulations agreed with actual measured values, alternative dosing regimens were additionally simulated for typical individuals. These alternative regimens were Benepali 50mg administered every 5 or every 10 days, alongside the usual dosing regimen of every 7 days (Figure 4.12). All simulated doses achieved steady-state drug concentrations well above the therapeutic window of etanercept, proposed by Jamnitski *et al* as between 2.1 – 4.7 mg/L[220]. As expected, time to reaching steady-state drug concentrations had negligible difference between the three simulated dosing regimens of Benepali.

Figure 4.11. Simulation of Benepali PK profile for 10,000 individuals using the final popPK model developed for this drug.



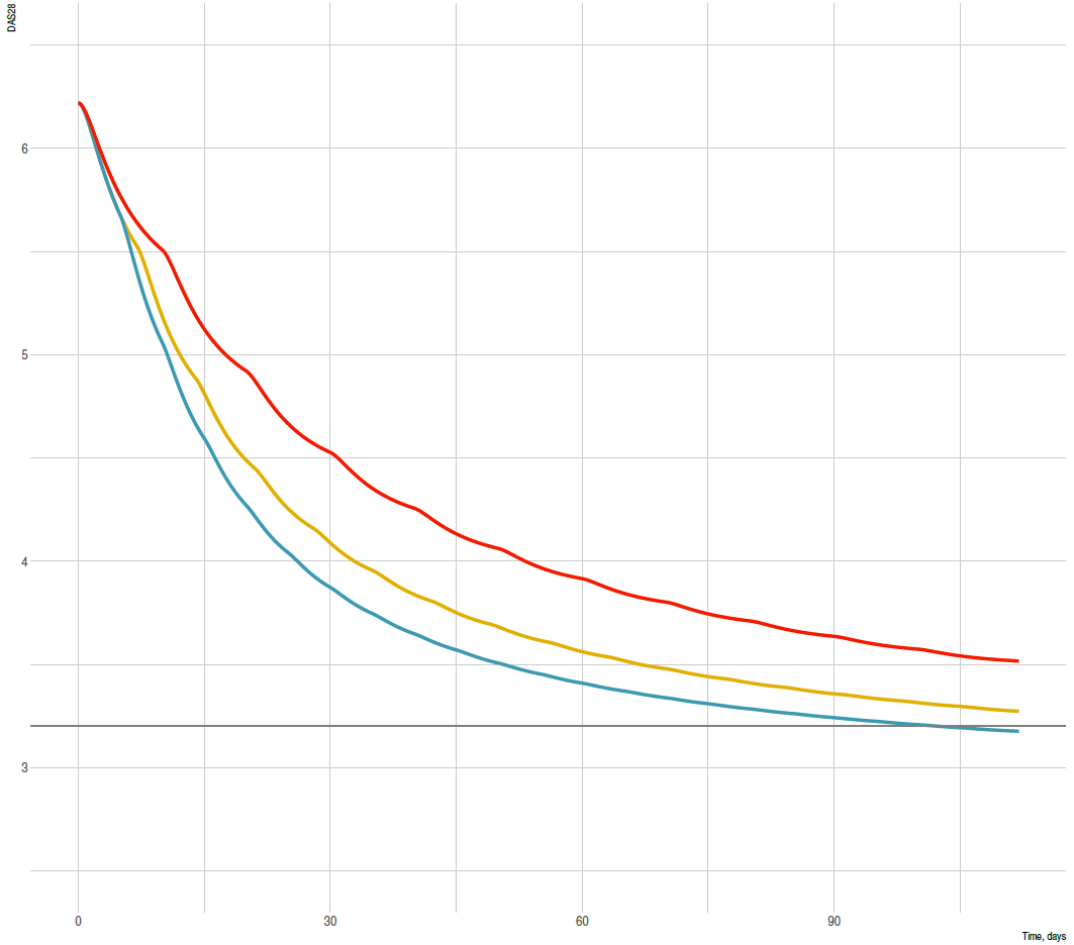
LEGEND: Horizontal grey lines represent the Benepali therapeutic window of 2.1 – 4.7 mg/L proposed by Jamnitski *et al*[220]. The dark blue line represents the population median, the light blue line represents the 5th percentile and the purple line represents the 95th percentile of the population. The black dots represent actual population drug concentration values.

Figure 4.12. Simulated usual and alternate dosing intervals of Benepali for a typical patient with different dose rates.



LEGEND: Horizontal grey lines represent the Benepali therapeutic window of 2.1 – 4.7 mg/L proposed by Jamnitski *et al*[220]. The purple line represents the usual Benepali dosing interval of 50mg every 7 days, the orange line represents a dosing interval of 50mg every 5 days and the pink line represents 50mg every 10 days.

Figure 4.13. Simulated PD profiles of alternative dosing regiments of Benepali with DAS28 as the outcome measure.



LEGEND: The yellow line represents the usual dosing regimen of Benepali every 7 days, the blue line represents dosing every 5 days and the red line represents dosing every 10 days. The horizontal grey line represents a DAS28 of 3.2 (low disease activity).

4.1.6. Simulation of PD models for patients with RA initiating Benepali

A simple inhibitory E_{max} model of DAS28 as the outcome measure of response to treatment with Benepali was successfully simulated for the usual dosing regimen of 50mg every 7 days, as well as for 50mg every 5 and 10 days; simulations are presented in Figure 4.13. A regimen of 50mg every 5 days successfully simulated reaching DAS28 levels below 3.2 (low disease activity) at 12 weeks. Regimens of 50mg every 7 days (licensed dose rate) and every 10 days did not reach DAS28 $\leq$ 3.2 at 12 weeks, although these dose rates showed a trend towards eventual DAS28 $\leq$ 3.2. Improvement in DAS28 in all three dosing rates was most dramatic in the first 30 days of treatment with Benepali.

4.2. Discussion

4.2.1. Development of a popPK model for patients initiating Amgevita and subsequent simulation

Using 58 drug concentration samples from 10 patients with RA, collected over a 12-week period for each patient, a popPK model was developed to describe the study population and estimate population PK parameters. These parameter estimates were similar to those from previous studies, and this, alongside satisfactory visual checks, meant that model fit was determined to be satisfactory. Parameter estimates were then used to simulate drug concentrations based on altered dosing intervals, as well as PD responses to altered dosing intervals.

A one-compartment popPK model with first-order absorption and elimination for Amgevita in the first 12 weeks of treatment was developed using the study population. No covariates were included in the final model, as these did not achieve dramatic reductions in AIC or -2LL, whilst reducing model stability due to the sparse sampling that the model was built on. The value of ka was fixed as per previous findings[180] and no BSV was estimated for this parameter, due to high residual error in previous iterations of the model. However, diagnostic plots were satisfactory and estimated parameters were in-keeping with prior knowledge, so these values were used to successfully simulate models in further scenarios:

1. An Amgevita PK model using the usual dosing interval (40mg every 14 days) as proof-of-concept and validity of PK parameter estimates obtained from the popPK model, simulated in 10,000 subjects.
2. An Amgevita PK model illustrating altered dosing intervals compared to usual dosing in a typical individual.

3. An Amgevita PD indirect response model to describe the relationship between CRP as an outcome measure and Amgevita serum concentrations.
4. An Amgevita PD direct E_{max} inhibitory model to describe the relationship between DAS28 as an outcome and Amgevita serum concentrations.

There is only one other popPK study of subcutaneously administered adalimumab in RA patients, carried out by Ternant *et al*[180], and this study was the basis for corroboration of estimated PK parameters, as well as utilisation of values for simulation of PD models. The Ternant study design was different from the present study, in that it was a post-hoc analysis of a single-centre observational study based in France carried out in RA patients over 52 weeks, whereas this study was a prospective multi-centre study based in the UK carried out over 12 weeks. The Ternant study did not mention whether patients included were adherent with medication for the full 52 weeks of study. 77% of patients in the Ternant study were female, compared to 90% in this study. Furthermore, sparse samples were taken at baseline and weeks 6, 12, 24 and 52, compared to baseline, 1 hour and weeks 2, 4, 6 and 12 in the current study. However, despite these differences in design, findings in the current study still broadly agreed with Ternant *et al*.

The typical individual V_D was estimated at 9.19 L, which is higher than trial data for Humira, the proprietary originator form of adalimumab, where in a number of studies using both IV and SC administration over a range of doses the V_D was estimated at between 2.98 – 7.5 L[251]. CL was estimated at 0.0121 L/hr, which falls within the range determined from Humira trial data of 0.00676 – 0.0322 L/hr[251]. Given that CL was not reduced below expected values, the estimate of V_D may be above expected due to a number of reasons. Nine out of ten patients in this study were female, and this may have skewed V_D estimates compared to other studies with a more even distribution between biological sexes. Sex will affect physical and physiological properties, such as body composition, so drug distribution in this patient cohort may be altered compared to in other studies, for example, a greater proportion of drug in whole blood, and not just confined to plasma. Another possibility is that the population estimate is not as accurate as it could have been if a greater number of patients had been recruited to the study. However, this limitation has been somewhat mitigated by the use of a smarter study design, based on ideal sample timings simulated from previous data.

PK parameter estimates derived from modelling were successfully used to simulate the popPK profile of 10,000 subjects. The population median Amgevita concentration PK

profile reached steady-state at the upper end/just above the therapeutic window of drug concentrations between 5 – 8 mg/L, as previously defined by Pouw *et al*[219]. Figure 4.4 demonstrates the median and 5th and 95th percentiles for the simulated population, and the black dots that have been overlaid onto this profile represent actual Amgevita concentrations from the BRAGGSS-PD patients used to derive the originating popPK model. These actual values agree well with simulated values, so this simulation is proof-of-concept that PK parameter estimates from the originator popPK model can be used for simulation of alternative dosing intervals, as well as PD model simulation. Of note, the 95th percentile in the simulation is estimated as being much higher than actual values, and this could be because nominal dosing records instead of actual dosing records were used for this simulation.

Simulations for a typical individual were carried out with alternative dosing intervals to the usual regimen of Amgevita 40mg SC every 14 days, namely, every 7 or 21 days, and this is shown in Figure 4.5. This simulation predicted that the usual dosing regimen of Amgevita, based on the final popPK profile developed, reached steady-state concentrations just at the lower end of the therapeutic window. This is the optimum range for this medication, as it will maximise chances of a patient achieving therapeutic efficacy, whilst also minimising the risk of drug toxicity and unwanted adverse effects. The dosing simulation of 40mg every 7 days reached the therapeutic window more quickly, but took the same amount of time to reach steady-state concentrations. Furthermore, the steady-state range was at the upper end of and above the therapeutic window, increasing the risk of drug toxicity. The dosing simulation of 40mg every 21 days achieved steady-state concentrations at the same time as the other two dosage regimen simulations, but the steady-state range was below the therapeutic window and would be unlikely to achieve therapeutic efficacy *in vivo*.

Simulations for a typical individual were performed to predict PD profiles based on the established relationship between Amgevita concentration and the disease activity outcome measures of CRP and DAS28. CRP and DAS28 scores were not available for patients recruited from BRAGGSS-PD at all sampling time points, as repeated assessment of these measures had not originally been designed into the study protocol, so instead, values for k_{in}, k_{out}, DAS_0 and IC_{50} were taken from the Ternant *et al* study[180].

The PD model for CRP was simulated as an indirect response model for a typical individual with alternative dosing intervals alongside the usual regimen of Amgevita 40mg SC every

14 days, namely, every 7 or 21 days, and this is shown in Figure 4.6. All three simulated dosing regimens predicted a dramatic reduction in CRP in the first week of treatment, but reduction plateaued at levels that would indicate ongoing systemic inflammation for both dosing at 14 days (usual dose) and 21 days. Only the simulated dosing interval of every 7 days led to CRP levels being reduced below a clinically meaningful level. The accuracy of this prediction would likely have been improved if CRP was measured at all sampling time points in the BRAGGSS-PD patients, as values from the Ternant *et al* study may not have been transferable to this patient population. However, results indicate that a dosing regimen of Amgevita 40mg every 21 days likely would be unable to control systemic inflammation to a level that would correlate with decreased RA disease activity.

The PD model for DAS28 was simulated as a direct E_{max} inhibitory model for a typical individual with alternative dosing intervals alongside the usual regimen of Amgevita 40mg SC every 14 days, again, every 7 or 21 days; this is shown in Figure 4.7. All three dosing regimens predicted a dramatic reduction in DAS28 in the first week of treatment, but then started to diverge. Usual dosage at 40mg every 14 days led to a range of DAS28 around the cut-off for low disease activity at ≤3.2, but without reaching complete remission at ≤2.6. A dosing interval of every 7 days led to the most dramatic reduction in DAS28 steady-state range to around 2, which is below the threshold for clinical disease remission. However, a dosing interval of 21 days led to a DAS28 within the moderate disease activity range. Whilst this would be enough to ensure continuation of treatment, according to NICE guidance[8 146], it implies ongoing systemic inflammation and risk of irreversible joint damage due to uncontrolled RA disease activity. Therefore, starting with a reduced dosing interval has been shown to be unlikely to be beneficial beyond the first few weeks of treatment, as demonstrated from simulation of CRP and DAS28 levels using this dosing interval. There might be scope in future to reduce to this dosing interval once a significant reduction has been achieved with a more frequent dosing interval e.g. every 7 or 14 days. There could be an argument for initiating treatment at 40mg every 7 days in patients with very high disease activity until steady-state drug concentrations and low PD outcomes are reached, and then stepping down to every 21 days as maintenance therapy after the first three months of treatment, given what has been demonstrated from simulations. However, simulations were not carried out for dosing intervals of varying lengths for a typical patient, and there is scope to do this work in the future.

This study's strengths are that it is the first prospective popPK study of patients with RA commencing Amgevita; previous studies have been carried out on patients starting Humira (the proprietary adalimumab originator compound). The majority of patients in the UK commence a biosimilar drug and it is, therefore, important to explore the PK profile of these drugs and how this may influence disease outcome. This study was designed specifically for this purpose, and was not carried out as part of post-hoc analysis. Furthermore, patients were real-world patients managed in NHS rheumatology outpatient clinics, and were not clinical trial patients. This means that drug concentrations obtained are likely to be more reflective of day-to-day clinical practice. All samples were collected by the author and were immediately placed on ice following the blood draw; samples were also delivered to the central processing laboratory by the author within 24 hours of the blood draw, ensuring the highest possible quality of serum sample. All blood draws and witnessed dose administrations were documented to the nearest minute by the author, maximising the accuracy of the popPK model. Finally, the witnessing of doses at each study visit ensured true trough drug concentrations, as on each occasion, phlebotomy was carried out immediately before patient self-administration of each dose.

This study does have a number of weaknesses, however, with recruitment throughout the course of the PhD below target. The initial target was to recruit approximately 20 – 30 patients starting Amgevita, which was felt to be a realistic target prior to the commencement of this PhD. However, a number of factors have contributed to low recruitment. Firstly, recruitment coincided with the approval of tsDMARD agents in England by NICE[147 148]. While these agents were placed at a similar or increased price point in comparison to bDMARD agents at the same stage in the therapeutic escalation pathway, many patients may have been commenced on these drugs preferentially if they found the prospect of regular injectable medication unacceptable, thus reducing the potential number of recruits.

Initially, patients were only recruited from one centre (Manchester Royal Infirmary), and some progress was made towards improving recruitment by opening a second centre at Bolton One. However, this had limited impact, because the UK went into national lockdown due to the spread of the worldwide 2019 coronavirus disease (COVID-19) pandemic, starting in March 2020. All recruitment to clinical studies at the University of Manchester was suspended and the author was re-deployed to front-line NHS services for five months. After a return to full-time research, recruitment to the BRAGGSS-PD study only re-opened in October 2020 due to limited site capacity, although a third centre was opened at Tameside

General Hospital. There was also a reduced number of patients compared to pre-pandemic starting bDMARDs, as many appointments were via telephone, and accurate DAS28 calculation in order to determine the threshold for treatment escalation was unable to be carried out.

Undoubtedly, an increased number of patients would have led to a more accurate popPK model, particularly if there had been less disparity in the split between male and female patients (nine females versus one male). Improved accuracy of modelling would also have been achieved if there had been some mechanism to accurately record the precise time and date of administration of medication doses given between study visits, instead of using nominal administration times during the modelling process. This would be an area of development for the future, perhaps by sending reminder text messages to patients just prior to the scheduled dosage time and asking them to reply immediately after self-administration of medication.

Future work could include further recruitment of patients using the same study protocol in order to develop a more robust popPK model. An increased number of patients would also mean that a more stable model could be generated if covariates were included, potentially leading to a more accurate population model. Another area of development would be to include drug response outcome measure collection at each follow-up visit e.g. CRP, DAS28, so that a *de novo* PK-PD model could be generated. In future, patients could then be recruited to a personalised dosing trial with sparse serum drug concentration sampling. The specific PK profile of a patient could be determined based on prior modelling and the patient's drug concentration in relation to the last time and date of administration of Amgevita, and advice could be given regarding further dosing intervals. This would then tailor the dose of drug to each individual patient, to ensure a steady-state drug concentration range within the therapeutic window, but also minimising the risk of toxicity and adverse events.

In conclusion, a popPK model of patients with RA commencing Amgevita, an adalimumab biosimilar, was successfully derived from ten real-world patients. Parameter estimates from this model were used to simulate alternative dosing intervals, and how these would behave in simulated PD models of treatment response to Amgevita. This work forms the basis for future work with larger study numbers and potentially a prospective trial of personalised dosing for patients with RA starting Amgevita. Finally, because this study was carried out over a period of only 12 weeks, compared to over a period of 52 weeks in the Ternant study,

future implications of this could be that sampling need only be carried out in the first few months of treatment with Amgevita in order to determine an individual patient's PK profile and whether they are likely to achieve steady-state concentrations within the therapeutic window on their current dosing regimen.

4.2.2. Development of a popPK model for patients initiating Benepali and subsequent simulation

Using 40 drug concentration samples from six patients with RA, collected over a 12-week period for each patient, a popPK model was developed to describe the study population and estimate population PK parameters. These parameter estimates were similar to those from previous studies, and this, alongside satisfactory visual checks, meant that model fit was determined to be satisfactory. Parameter estimates were then used to simulate drug concentrations based on altered dosing intervals, as well as PD responses to altered dosing intervals.

A one-compartment popPK model with first-order absorption and elimination for Benepali was developed using the study population. This is in-keeping with previous findings[176]. No covariates were included in the final model, as these did not achieve dramatic reductions in AIC or -2LL, whilst reducing model stability due to the sparse sampling that the model was built on. The value of ka was fixed as per previous findings[250] and no BSV was estimated for this parameter, due to high %RSE in previous iterations of the model. The random effect (BSV) parameter was also not estimated for V_D for the same reason. However, diagnostic plots were satisfactory and estimated parameters were in-keeping with prior knowledge, so these values were used to successfully simulate models in further scenarios:

1. A Benepali PK model using the usual dosing interval (50mg every 7 days) as proof of concept and validity of PK parameter estimates obtained from the popPK model, simulated in 10,000 subjects.
2. A Benepali PK model illustrating altered dosing intervals compared to usual dosing in a typical individual.
3. A Benepali PD inhibitory E_{max} model to describe the relationship between DAS28 as an outcome and Benepali serum concentrations

This is the first popPK study of patients with RA starting the etanercept biosimilar Benepali, and furthermore, it is the first using the usual dosage regimen of 50mg SC every 7 days; other studies have used mixed dosing regimens with a mixture of IV and SC administration.

Another study also used a mixed population of HC and patients with RA, AS and JIA, pooling patients with different diagnoses for analysis[174]. A previous PK-PD study was carried out by Hsu and Huang[176], but that study used Enbrel, the proprietary etanercept originator compound, and not one of the currently more commonly prescribed biosimilar versions. In addition, the analysis by Hsu and Huang was a meta-analysis of previously published data, as opposed to the prospective approach used in the current study.

Interestingly, findings of the current study agree broadly with those of Korth-Bradley *et al*[250], who carried out popPK analysis in a cohort of HC with no RA pathology who received only a single dose of etanercept 25mg SC. However, mean CL (0.0446 L/hr) was higher and mean V_D (7.76 L) was lower in the current study compared to Korth-Bradley *et al*, which is likely reflective of altered PK in study subjects with ongoing inflammation and active RA pathology compared to HC. Findings cannot be directly compared with the findings of Zhou *et al* due to their use of a two-compartment model, but their estimate of CL of 0.072 L/hr was slightly higher compared with the estimate in the current study. The CL estimate of the current study is also lower than the estimate of Shennak *et al* for Enbrel (0.11 L/hr)[175], but that study was only carried out in biologically male patients with RA, which could explain the increased CL compared to the current study cohort, where four out of six patients were female. However, the V_D of Enbrel in the Shennak *et al* study was estimated at 15.04L, which is almost double the estimate of the current study. There is no certain cause for these discrepancies, but estimates from the current popPK model may have been different with improved recruitment and a more equal balance between male and female patients.

PK parameter estimates derived from modelling were successfully used to simulate the popPK profile of 10,000 subjects. The population median Benepali concentration PK profile reached steady-state at above the therapeutic window of drug concentrations between 2.1 – 4.7 mg/L, as previously defined by Jamnitski *et al*[220]. Figure 4.11 shows the median and 5th and 95th percentiles for the simulated population, and the black dots that have been overlaid onto this profile represent actual Benepali concentrations from the BRAGGSS-PD patients used to derive the originating popPK model. These actual values agree well with simulated values, so this simulation is proof-of-concept that PK parameter estimates from the originator popPK model can be used for simulation of alternative dosing intervals.

Simulations for a typical individual were carried out with alternative dosing intervals to the usual regimen of Benepali 50mg SC every 7 days, namely, every 5 or 10 days, and this is

shown in Figure 4.12. This simulation predicted that the usual dosing regimen of Benepali, based on the popPK analysis of the originator compound, reached steady-state concentrations above the upper limit of the therapeutic window, which could suggest over-dosing of patients and increased risk of toxicity and dose-related adverse events. The dosing simulation of 50mg every 5 days took the same amount of time to reach steady-state concentrations, which were also above the therapeutic window, and at an even higher level than usual dosing of every 7 days. The simulation of 50mg every 10 days achieved steady-state concentrations marginally faster than the other two dosing interval simulations, and steady-state concentrations were at the upper end/above the therapeutic window. This dosing interval could potentially be brought forward into a prospective trial to compare efficacy versus the usual dosing interval of every 7 days, as time to steady-state is almost identical, and steady-state concentrations are still within the therapeutic range. Reduced dosage of Benepali, particularly from initiation of the drug, is likely to represent significant cost-savings, while also reducing the risk of dose-related adverse events. A future research recommendation is to evaluate this approach as part of a clinical trial, with inclusion of health economic evaluation.

The PD model for DAS28 was simulated as a simple inhibitory E_{max} model for a typical individual with alternative dosing intervals alongside the usual regimen of Benepali 50mg SC every 7 days, namely, every 5 or 10 days; this is shown in Figure 4.13. All three dosing regimens demonstrated steep improvement in DAS28 in the first 2 – 4 weeks of treatment, but then improvement began to plateau. Improvement was similar for Benepali given every 5 and 7 days over the first week of treatment, but then began to diverge. Usual dosage of 50mg every 7 days lead to a DAS28 just above the cut-off for low disease activity of 3.2 after 12 weeks of treatment, and a reduced dosing rate of every 10 days did not reach this target. However, an increased dosing rate of every 5 days reached a DAS28 of below 3.2 at approximately 14 weeks, with this simulation indicating that this increased dosing rate may be likely to control active RA more rapidly. However, starting Benepali with a reduced dosing interval of every 10 days is unlikely to be beneficial beyond the first few weeks of treatment. There could potentially be scope in future to reduce this dosing interval once a significant reduction had been achieved with a more frequent dosing interval e.g. every 5 or 7 days. Again, there could be an argument for initiating treatment at the licensed dosing rate of 50mg every 5 days in patients with very high disease activity until steady-state drug concentrations and low PD outcomes are reached, and then stepping down to every 10 days as maintenance therapy after the first three months of treatment, given what has been

demonstrated from simulations. Future work could involve carrying out simulations for dosing intervals of varying lengths for the same typical patient.

This study's strengths are that it is the first prospective popPK study of patients with RA commencing Benepali; previous studies have been carried out on patients starting Enbrel (the proprietary etanercept originator compound) or alternate biosimilar compounds other than Benepali. As with the Amgevita study, participants were real-world patients managed in NHS rheumatology outpatient clinics, as opposed to a controlled cohort of clinical trial patients. Again, drug concentrations in this study are likely to be more reflective of day-to-day clinical practice. Furthermore, sample collection, processing and documentation of blood draws and dose administration had the same robust procedure as documented for Amgevita in Section 4.2.1.

As with the Amgevita popPK study, recruitment was well below the target of 20 – 30 patients on Benepali. As well as the reasons outlined for Amgevita recruitment problems, another explanation for poor recruitment could be due to the relative costs of Amgevita and Benepali. When this PhD commenced, Benepali was the preferred first-line bDMARD agent in the Greater Manchester area, where this study was conducted. However, soon after recruitment opened, Amgevita became the first-line agent of choice due to availability at a lower price point, and Benepali became reserved for patients who were more at risk of serious infections. As with the Amgevita study, nominal drug administration times were used for medication doses given outside of study visits that were not witnessed by the author; more accurate reporting of administration times by patients, as well as increased participant numbers, may have led to improved model accuracy, as discussed in Section 4.2.1.

Future work could include additional recruitment of participants in order to improve model accuracy, improved recording of medication administration outside of study visits, and a potential additional PD study with recording of treatment outcome measures at each study visit. The finding of steady-state Benepali concentrations within the therapeutic window in the simulation of dosing every 10 days could provide the basis for a multi-dose comparative trial of efficacy from initiation of therapy. As with Amgevita, detailed and accurate modelling could also form the basis for a future personalised dosing trial, with drug concentration sampling leading to determination of a patient's Benepali PK profile, which could then be used to give advice on current dosage.

In conclusion, a popPK model of patients with RA commencing Benepali, an etanercept biosimilar, was successfully derived from six real-world patients. Parameter estimates from this model were used to simulate alternative dosing intervals, and a reduced dose of 50mg every 10 days was found to still achieve steady-state drug concentrations within the therapeutic window of etanercept. This work forms the basis for future work with larger study numbers and inclusion of PD data, and potentially a prospective trial of personalised dosing for patients with RA starting Benepali.

4.3. Chapter summary

PopPK models for the first 12 weeks of treatment with Amgevita or Benepali in a cohort of prospectively recruited patients with RA from the BRAGGSS-PD sub-study were successfully developed. PK parameter estimates from these models were then used to simulate alternative dosing regimens to determine whether these would lead to steady-state concentrations within each drug's previously defined therapeutic window. PD models were then simulated for both drugs studied. Findings could form the basis of future personalised dosing studies for these two agents in RA patients who are being initiated on these therapies.

CHAPTER FIVE: PROTEOMIC PREDICTORS OF TREATMENT RESPONSE TO ETANERCEPT IN PATIENTS WITH RHEUMATOID ARTHRITIS - RESULTS

Summary of chapter contents:

5.1. Study participants

The patients recruited to the BRAGGSS-PD sub-study are detailed in Sections 4.1.1 and 4.1.4. Their demographic details are summarised in Table 5.1.

A total of 180 patients were included in the etanercept sub-cohort from the wider BRAGGSS dataset. 134 patients were female and 46 patients were male. All patients were of Caucasian ethnicity. The median age was 57.40 years (IQR 50.02 – 65.09) and the median pre-treatment DAS28 was 5.85 (IQR 5.24 – 6.37). Detailed baseline patient characteristics are outlined in Table 5.2. 26 values for CRP were imputed at baseline. At three months follow-up, the following values were imputed: three values for TJC, three values for SJC, six values for VAS of global health and 27 values for CRP. At six months follow-up, the following values were imputed: seven values for TJC, seven values for SJC, eight values for VAS of global health and 26 values for CRP.

Table 5.1. Baseline characteristics of patients recruited to the BRAGGSS-PD sub-study.

Characteristic	**Statistic**
Female sex, n (%)	13 (81.25%)
Age (years), median [IQR]	56.00 [49.50, 60.00]
Disease duration prior to starting bDMARD (years), median [IQR]	4 [2, 6.5]
Body weight (kg), median [IQR]	76.50 [67.50, 104.50]
Concurrent csDMARD, n (%)	13 (86.67) [1 missing]
DAS28, median [IQR]	5.42 [5.03, 6.02]
Drug	
Amgevita, n (%)	10 (62.50)
Benepali, n (%)	6 (37.50)
Patients achieving therapeutic drug levels	
All time points, n/total samples (%)	45/91 (49.45)
Baseline, n (%)	0/16 (0.00)
1 hour, n (%)	0/13 (0.00) [3 missing]
1 week (etanercept only), n (%)	6/6 (100.00)
2 weeks, n (%)	6/15 (40.00) [1 missing]
4 weeks, n (%)	10/13 (76.92) [3 missing]
6 weeks, n (%)	12/15 (80.00) [1 missing]
12 weeks, n (%)	11/15 (73.33) [1 missing]

ABBREVIATIONS: Conventional synthetic disease-modifying anti-rheumatic drug (csDMARD), biological disease-modifying anti-rheumatic drug (bDMARD), disease activity score of 28 joint counts (DAS28), interquartile range (IQR).

Table 5.2. Baseline characteristics of patients in the etanercept cohort from the wider BRAGGSS dataset.

Characteristic	**Statistic**	**Missing, n (%)**
Female sex, n (%)	134 (74.44)	0 (0.00)
Age (years), median [IQR]	56.90 [49.96, 64.93]	0 (0.00)
Disease duration prior to starting bDMARD (years), median [IQR]	6 [2, 14]	0 (0.00)
Body mass index (BMI, kg/m^2), median [IQR]	27.56 [23.86, 32.54]	0 (0.00)
Concurrent csDMARD, n (%)	147 (81.67)	0 (0.00)
DAS28, median [IQR]	5.85 [5.25, 6.39]	26 (14.44)
Ever seropositive (RF and/or ACPA), n (%)	120 (66.67)	0 (0.00)
HAQ (maximum score: 3), median [IQR]	1.39 [1.00, 1.75]	30 (16.67)
HADS Anxiety Score (maximum score: 21; defined as "Anxiety" if score ≥ 11), median [IQR]	7.27 [5.00, 9.00]	29 (16.11)
HADS Depression Score (maximum score: 21; defined as "Depression" if score ≥ 11), median [IQR]	6.74 [5.00, 9.00]	23 (12.78)

ABBREVIATIONS: Anti-citrullinated peptide antibody (ACPA), biological disease-modifying anti-rheumatic drug (bDMARD), body mass index (BMI), conventional synthetic disease-modifying anti-rheumatic drug (DMARD), disease activity score of 28 joint counts (DAS28), Health Assessment Questionnaire (HAQ), Hospital Anxiety and Depression Score (HADS), interquartile range (IQR), rheumatoid factor (RF).

Finally, 14 HCs were recruited for case-control analysis. 12 of the 14 patients were female (85.71%). The median age was 78 years (IQR 73 – 82).

5.2. Protein library generation

A total of 476 unique proteins were identified from the literature search previously outlined. A complete list of individual proteins is included in Appendix Five. Summaries of proteins included according to whether a study was based on RA pathogenesis, diagnosis, prognosis and treatment response/monitoring are in Tables 5-8, respectively.

Table 5.3. Summary of protein studies in RA pathogenesis.

Study	**Sample**	**Groups of study**	**Proteomic strategy**	**Proteins identified in RA**
Degré 1983[252]	Synovial fluid	10 RA	Folin phenol reagent method	IFN-γ
Biemond *et al* 1984[253]		17 RA, 6 OA, 7 traumatic effusions	Chromatography, immunoassay	CP
Gysen *et al* 1985[254]		15 RA, 18 OA	ELISA	A1AT, MMP1
Malyak *et al* 1993[255]		16 RA, 18 non-RA inflammatory arthritis (IA)	ELISA	IL-1RA
Okano *et al* 1996[256]		38 RA, 45 OA, 11 HC	RIA	PTHrP
Schäffler *et al* 2003[257]		24 RA, 29 OA	ELISA	ADIPOQ, RETN
Kim *et al* 2006[258]		25 RA	2-DE and MALDI-TOF MS	FN1, GRB7
Tabushi *et al* 2008[259]		10 RA	MALDI-TOF MS	FGA, FN1, VIM
Katano *et al* 2009[260]		16 RA, 13 OA	MALDI-TOF-MS	11 proteins upregulated following stimulation of neutrophils with GM-CSF
Baillet *et al* 2010[261]		30 RA, 18 IA controls	MALDI-TOF MS	S100A8, S100A9, S100A12
Mateos *et al* 2012[262]		20 RA, 20 OA	Nano-LC and MALDI-TOF/TOF	17 differentially expressed proteins
Noh *et al* 2014[263]		11 RA, 15 non-RA	2-DE and MALDI-TOF	32 differentially expressed proteins
Yang *et al* 2015[264]		25 RA, 10 HC	GC-TOF MS	8 differentially expressed enzymes
Meng *et al* 2016[265]		34 RA, 24 non-RA effusions	LC-MS/MS	4 histone proteins acting as ACPA autoantigens
Firestein *et al* 1992[266]	Synovial tissue	12 RA, 12 OA	*In-situ* hybridisation	IL-1RA
Yamasaki *et al* 2001[267]		10 RA, 10 OA	Electrophoretic mobility shift	NF-κB
De Rycke *et al* 2005[268]		19 RA	Immunostaining	Intracellular citrullinated proteins
Chang *et al* 2009[269]		10 RA, 10 OA, 6 AS	2-DE and MALDI-TOF MS	10 differentially expressed proteins
Wang *et al* 2012[270]		50 RA, 10 HC	2D nano-ESI LC-MS/MS	100 differentially expressed proteins in RA
Yan *et al* 2012[271]		10 RA, 10 OA, 10 SpA	2-DE MALDI-TOF/TOF MS	VDBP
Chang *et al* 2013[272]		10 RA, 10 OA	2-DE and MALDI-TOF/TOF MS	VDBP
Doran *et al* 1995[273]	Synovial mononuclear cells	17 RA	Western blotting	HAPLN1
Swedlund *et al* 1974[274]	Serum	24 RA, 40 HC	Radial immune-diffusion and biuret method	A1AT
Baskol *et al* 2006[275]		57 RA, 25 HC	Enzymatic spectrophotometry	Increased MPO levels in RA

Study	Sample	Groups of study	Proteomic strategy	Proteins identified in RA
Grazio *et al* 2013[276]	Plasma	20 RA, 20 PsA, 20 OA	SDS-PAGE and LC-MS	13 differentially expressed proteins
Yang *et al* 2018[277]		12 RA-MCI, 12 RA non-MCI, 12 HC	2D-LC-MS/MS	SHH, TTR
Schulz *et al* 2007[278]	PBMCs	32 RA, 33 HC	2-DE and MALDI-TOF MS	9 differentially expressed proteins
Lu *et al* 2010[279]	Monocytes and macrophages	13 RA, 10 OA	ELISA	Citrullinated GRP78
Darrah *et al* 2017[280]	CD4+ T cells	11 RA, 5 SSc, 8 PsA, 2 HC	Flow cytometry	PADI4
Olszewski *et al* 2001[281]	Lymphatic fluid	20 RA, 20 HC	ELISA	Increased levels of pro-inflammatory cytokines in lymph versus blood
Doroshevskaya *et al* 2014[282]	Bone marrow	30 RA, 20 OA	Immunohistochemistry and immunocytochemistry	Reduced MDM2

ABBREVIATIONS: 78 kDa glucose-regulated protein (GRP78), adiponectin (ADIPOQ), alpha-1 antitrypsin (A1AT), ankylosing spondylitis (AS), anti-citrullinated protein antibody (ACPA), caeruloplasmin (CP), E3 ubiquitin-protein ligase Mdm2 (MDM2), electrospray ionisation (ESI), enzyme-linked immunosorbent assay (ELISA), fibrinogen alpha chain (FGA), fibronectin (FN1), gas chromatography (GC), granulocyte-macrophage colony-stimulating factor (GM-CSF), growth factor receptor-bound protein 7 (GRB7), healthy controls (HC), hyaluronan and proteoglycan link protein 1 (HAPLN1), IL1 receptor antagonist (IL1-RA), interferon (IFN)-γ, inflammatory arthritis (IA), liquid chromatography (LC), mass spectrometry (MS), matrix-assisted laser desorption/ionisation (MALDI), matrix metalloproteinase 1 (MMP1), mild cognitive impairment (MCI), myeloperoxidase (MPO), nuclear factor-κB (NFκB), osteoarthritis (OA), parathyroid hormone-related peptide (PTHrP), protein-arginine deiminase type-4 (PADI4), protein S100-A8 (S100A8), protein S100-A9 (S100A9), protein S100-A12 (S100A12), psoriatic arthritis (PsA), radioimmunoassay (RIA), resistin (RETN), rheumatoid arthritis (RA), sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), sonic hedgehog protein (SHH), spondyloarthritis (SpA), systemic sclerosis (SSc), time-of-flight (TOF), transthyretin (TTR), two-dimensional gel electrophoresis (2-DE), vimentin (VIM), vitamin D binding protein (VDBP).

Table 5.4. Summary of protein studies in RA diagnosis.

Study	**Sample**	**Groups of study**	**Proteomic strategy**	**Proteins identified in RA**
Wagatsuma *et al* 1996[283]	Serum	71 RA, 60 HC	Ethanol precipitation, PAGE, Western blotting	EZR, RDX, MSN
Chandra *et al* 2011[284]		120 RA, 27 AS, 28 PsA, 25 HC	Multiplex and single automated assays	6 differentially expressed proteins
Urbaniak *et al* 2017[285]		50 RA	Elution following chromatography with mixtures E1 and E2 with MALDI-TOF	E1: panel of 5 proteins; E2: panel of 6 proteins
Seok *et al* 2017[286]		Discovery: 36 RF+ RA, 18 RF- RA; validation: 40 RF+ RA, 40 RF- RA, 40 HC	Discovery: LC-MS/MS; validation: ELISA	SAA4
Kim *et al* 2018[287]		Discovery: 18 RF+ RA, 18 RF- RA; validation: 20 RF+ RA, 22 RF- RA, 23 RA (RF status not defined)	Discovery: LC-MS/MS; validation: ELISA.	HRG, LBP
Giusti *et al* 2010[288]	Saliva	20 RA, 20 HC	2-DE and MALDI-TOF/TOF MS	8 differentially expressed proteins
Siebert *et al* 2017[289]	Urine	50 RA, 50 PsA, 50 OA, 50 IBD, 50 HC	CE-MS	7 differentially expressed proteins in RA
Yang *et al* 2015[290]	FLS	44 RA, 15 OA, 15 AS, 15 HC	ELISA, Western blotting	CEMIP

ABBREVIATIONS: Ankylosing spondylitis (AS), capillary electrophoresis (CE), cell migration-inducing and hyaluronan-binding protein (CEMIP), enzyme-linked immunosorbent assay (ELISA), ezrin (EZR), healthy control (HC), histidine-rich glycoprotein (HRG), inflammatory bowel disease (IBD), lipopolysaccharide-binding protein (LBP), mass spectrometry (MS), matrix-assisted laser desorption/ionisation (MALDI), moeisin (MSN), osteoarthritis (OA), polyacrylamide gel electrophoresis (PAGE), psoriatic arthritis (PsA), radixin (RDX), rheumatoid arthritis (RA), rheumatoid factor (RF), serum amyloid A4 (SAA4), time-of-flight (TOF).

Table 5.5. Summary of protein studies in RA prognosis.

Study	**Sample**	**Groups of study**	**Proteomic strategy**	**Proteins identified in RA**
Forslind *et al* 2004[32]	Serum	379 RA	ELISA	ACPA
Hueber *et al* 2007[291]		56 RA, 21 PsA/AS, 19 HC	Microarrays and commercially available cytokine assays	Distinct autoantibody epitopes associated with increased pro-inflammatory cytokine levels
Shi *et al* 2011[37]		571 RA, 305 HC	ELISA	Anti-CarP antibodies
Cheng *et al* 2014[292]		30 RA, 30 HC	SDS-PAGE and LC-MS/MS	26 differentially expressed proteins
Hueber *et al* 2005[75]	Synovial fluid	76 RA, 27 other IA/OA, 11 HC	Microarrays	Distinct autoantibody epitopes associated with prognosis
Kang *et al* 2014[293]	Urine	Discovery: 20 RA, 19 OA; validation: 30 RA, 30 OA; soluble CD14 (sCD14) analysis: 274 RA, 120 OA, 60 SLE	SDS-PAGE and LC-MS/MS	134 differentially expressed proteins
Park *et al* 2016[294]		264 RA, 187 HC	ELISA	sCD14, ORM1, ORM2, GSN

ABBREVIATIONS: Ankylosing spondylitis (AS), anti-carbamylated protein (anti-CarP), anti-citrullinated protein antibody (ACPA), cluster of differentiation (CD), enzyme-linked immunosorbent assay (ELISA), gelsolin (GSN), healthy control (HC), inflammatory arthritis (IA), liquid chromatography (LC), mass spectrometry (MS), orosomucoid (ORM), osteoarthritis (OA), psoriatic arthritis (PsA), rheumatoid arthritis (RA), sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), soluble CD14 (sCD14), systemic lupus erythematosus (SLE).

Table 5.6. Summary of protein studies in treatment response to biologic agents in RA.

Study	Biologic agent	Sample	Groups of study	Proteomic strategy	Proteins identified in RA
Sekigawa *et al* 2008[126]	Infliximab	Serum/plasma	10 RA	2D LC-MS/MS	22 differentially expressed proteins before and after treatment
Trocmé *et al* 2009[128]		Plasma	60 RA	SDS-PAGE and SELDI-TOF MS	8 differentially expressed proteins between good and non-responders
Serada *et al* 2010[127]		Serum	33 RA, 9 Behçet's, 22 Crohn's, 50 HC	iTRAQ with nanoLC-MS/MS	71 differentially expressed proteins in RA
Fabre *et al* 2008[129]	Etanercept		33 RA	Protein biochip array	MCP-1 and EGF associated with treatment response
Hueber *et al* 2009[130]			93 RA	RA antigen microarray, FLEX®, ELISA	24-biomarker signature associated with treatment response
Obry *et al* 2015[131]			Discovery: 22 RA; validation: 16 RA	SDS-PAGE with nanoLC-MS	12 biomarkers associated with treatment response
Fabre *et al* 2009[132]	Rituximab		46 RA	Protein biochip array	Cytokine profile associated with treatment response at 3 months
Murota *et al* 2016[134]	Mixture		28 RA, 30 Sjögren's, 30 HC	SOMAscan® assay	33 differentially expressed proteins
Cuppen *et al* 2017[137]			Discovery: 65 RA; validation: 185 RA	xMAP®	12 differentially expressed proteins

ABBREVIATIONS: Epidermal growth factor (EGF), healthy control (HC), isobaric tag for relative and absolute quantitation (iTRAQ), liquid chromatography (LC), mass spectrometry (MS), monocyte chemoattractant protein (MCP), rheumatoid arthritis (RA), sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), surface-enhanced laser desorption-ionisation (SELDI), time-of-flight (TOF), two-dimensional (2D).

5.3. Proteins acquired via SWATH-MS

A first batch of 82 patients with RA from the etanercept sub-cohort was processed at SBDC in April 2018. These patients had serum samples available for processing collected pre-treatment and after three and six months of treatment. Although this was before the commencement of this PhD, *in silico* data extraction was carried out using the bespoke RA protein library after the library had been curated as part of this PhD. A total of 775 proteins were detected using a generic plasma library[205] and 397 proteins were detected using the bespoke RA library. Following removal of duplicated proteins with the plasma library, 233 unique proteins remained in the data extraction using the RA library. Therefore, a total of

1,008 unique proteins were identified from this first batch. Patients included in this batch had proteomics data available at pre-treatment and at three and six months after treatment.

A second batch of 74 RA patients from the etanercept cohort and 14 HCs was processed at SBDC in March 2019. These patients had serum samples available for processing collected pre-treatment and after 3 months of treatment. A total of 871 proteins were detected using the generic plasma library and 430 proteins were detected using the bespoke RA library. Some discrepancy in the number of proteins detected between batches is expected, as SWATH-MS methods were constantly being developed, optimised and improved throughout the collaboration with the SBDC. Following removal of duplicated proteins with the plasma library, 240 unique proteins remained in the data extraction using the RA library. Therefore, a total of 1,111 unique proteins were identified from this second batch. Patients included in this batch had proteomics data available at pre-treatment and three months after treatment. HCs only had cross-sectional data at a single time point.

A third batch of 22 RA patients from the etanercept cohort and nine patients from the BRAGGSS-PD cohort (receiving either Amgevita or Benepali) was processed at SBDC in February 2021. The patients from the etanercept sub-cohort had serum samples available at pre-treatment and after three months of treatment. The BRAGGSS-PD patients had serum samples available as per Figures 3.1 and 3.2. A total of 668 proteins were detected using the generic plasma library and 408 proteins were detected using the bespoke RA library. Following removal of duplicated proteins with the plasma library, 271 unique proteins remained in the data extraction using the RA library. Therefore, a total of 939 unique proteins were identified from this third batch. Patients in this batch from the etanercept cohort had proteomics data available at pre-treatment and after three months of treatment. Patients from the BRAGGSS-PD cohort had data available at pre-treatment and the following time points after treatment: 1 hour, 6/7 days (patients on Benepali only, n = 4), 2 weeks, 4 weeks, 6 weeks and 12 weeks.

A final batch of seven patients from the BRAGGSS-PD cohort was processed in September 2021. These patients had serum samples available as per Figures 3.1 and 3.2. A total of 621 proteins were detected using the generic plasma library and 412 proteins were detected using the bespoke RA library. Following removal of duplicated proteins with the plasma library, 286 unique proteins remained in the data extraction using the RA library. Therefore, a total of 907 unique proteins were identified from this fourth batch. Patients in this batch had data

available at pre-treatment and the following time points after treatment: 1 hour, 6/7 days (patients on Benepali only, n = 2), 2 weeks, 4 weeks, 6 weeks and 12 weeks.

5.4. QC of proteomics data

QC was initially carried out on patients from the etanercept cohort and HC. A total of 392 samples were initially quality controlled. Two duplicate samples were removed from the second batch to be processed at SBDC, leaving 390 remaining samples. 261 proteins with near-zero variance were removed from the first batch, 208 from the second batch and 40 from the third batch.

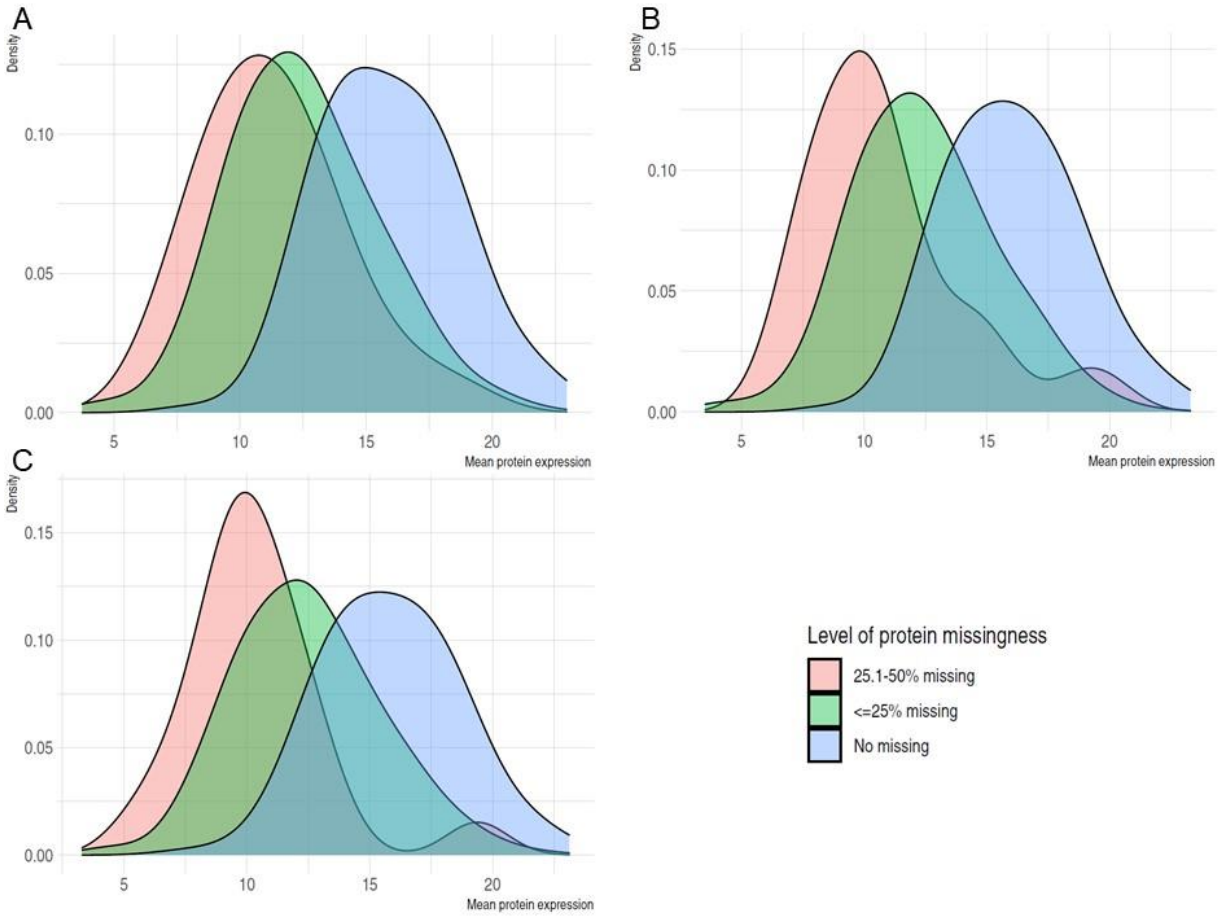
5.4.1. Assessment of mean protein expression compared with protein missingness

Density plots were generated to assess mean protein expression in each batch at each time point and at all time points combined according to levels of missingness, defined as:

- <25% missing values for each protein.
- 25 – 50% missing.
- 50 – 75% missing.
- >75% missing.

Density plots are shown in Figures 5.1 – 5.5.

Figure 5.1. Mean protein expression multi-density plots for the first batch of SWATH-MS data to be processed at SBDC, stratified by time point.



LEGEND: A) Proteins measured at baseline (pre-treatment). B) Proteins measured after three months of treatment. C) Proteins measured after six months of treatment.

Figure 5.2. Mean protein expression multi-density plot for the first batch of SWATH-MS data to be processed at SBDC, all time points combined.

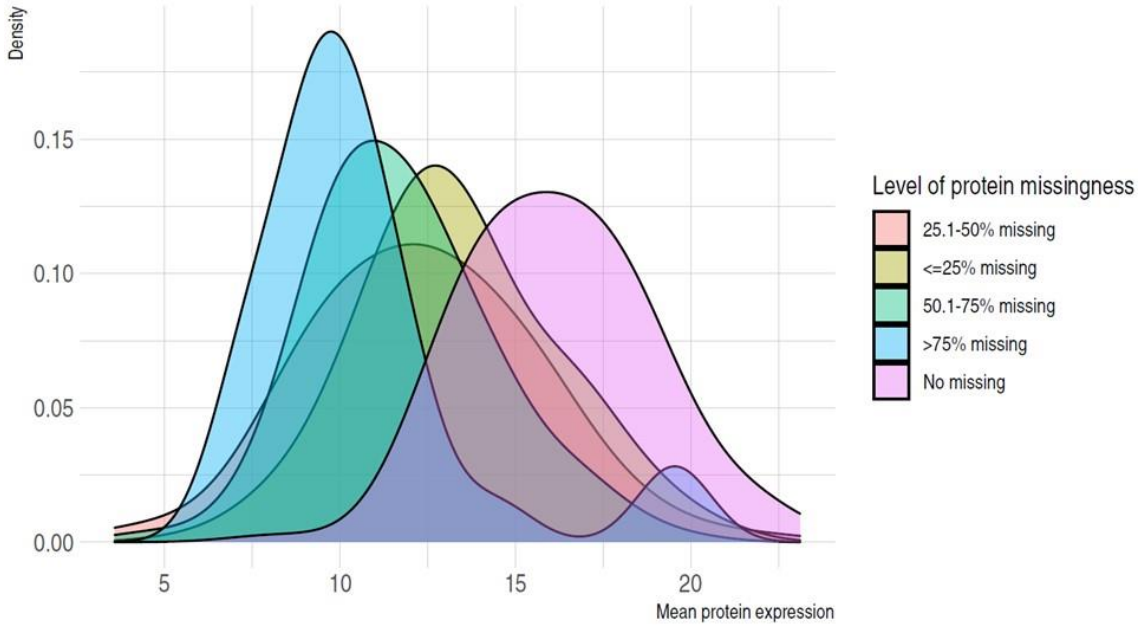
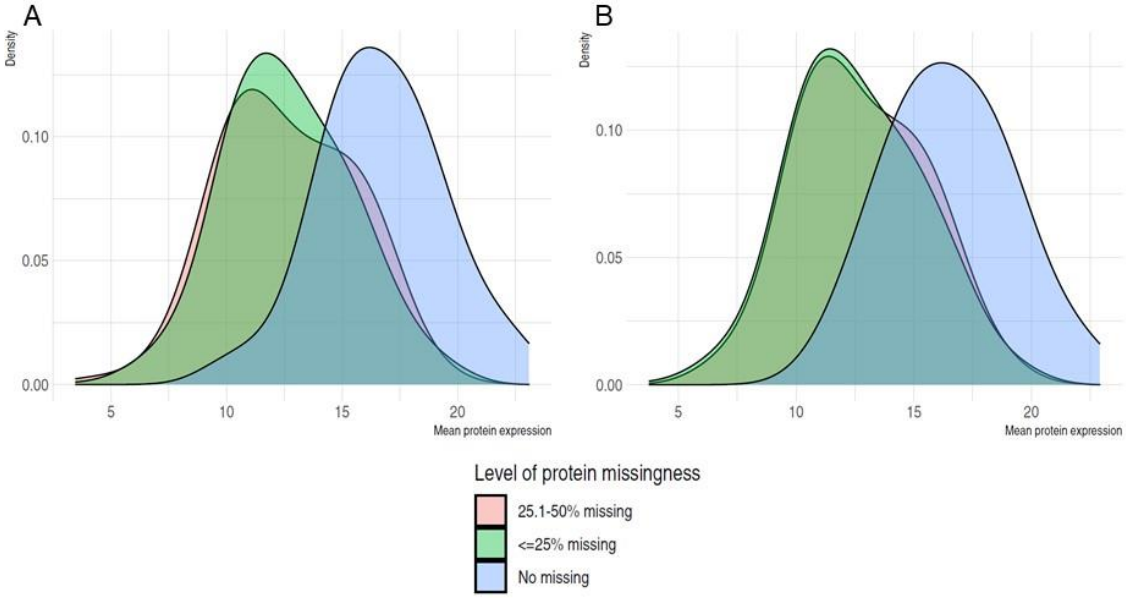
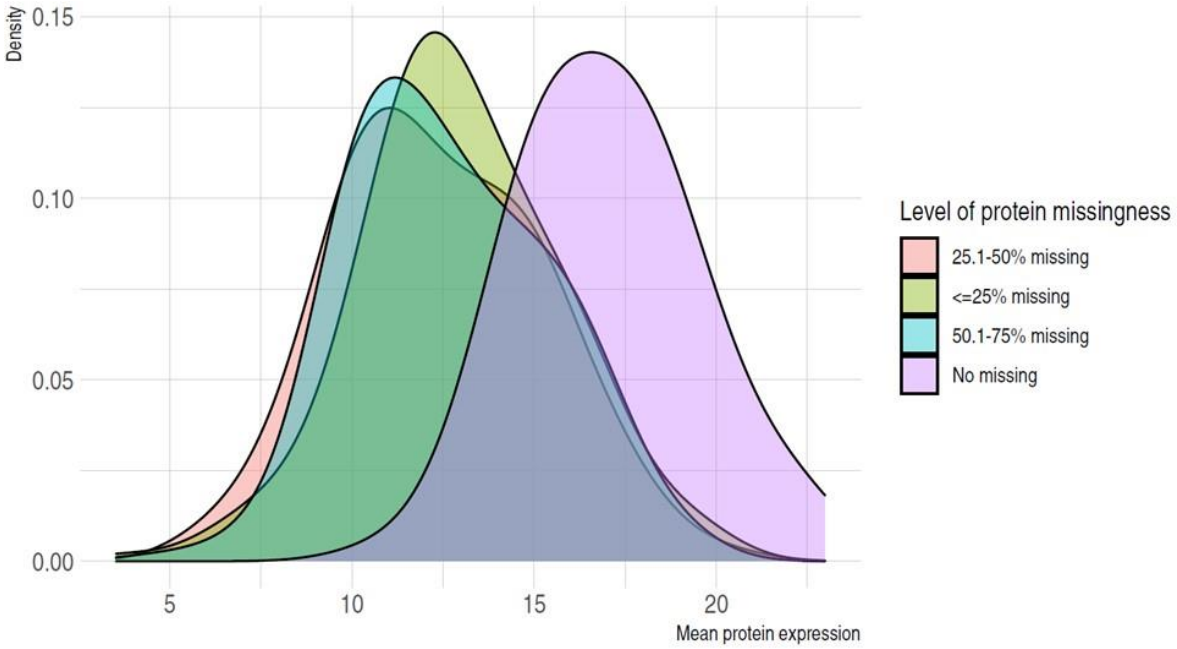


Figure 5.3. Mean protein expression multi-density plots for the second batch of SWATH-MS data to be processed at SBDC, stratified by time point.



LEGEND: A) Proteins measured at baseline (pre-treatment). B) Proteins measured after three months of treatment.

Figure 5.4. Mean protein expression multi-density plot for the second batch of SWATH-MS data to be processed at SBDC, all time points combined.

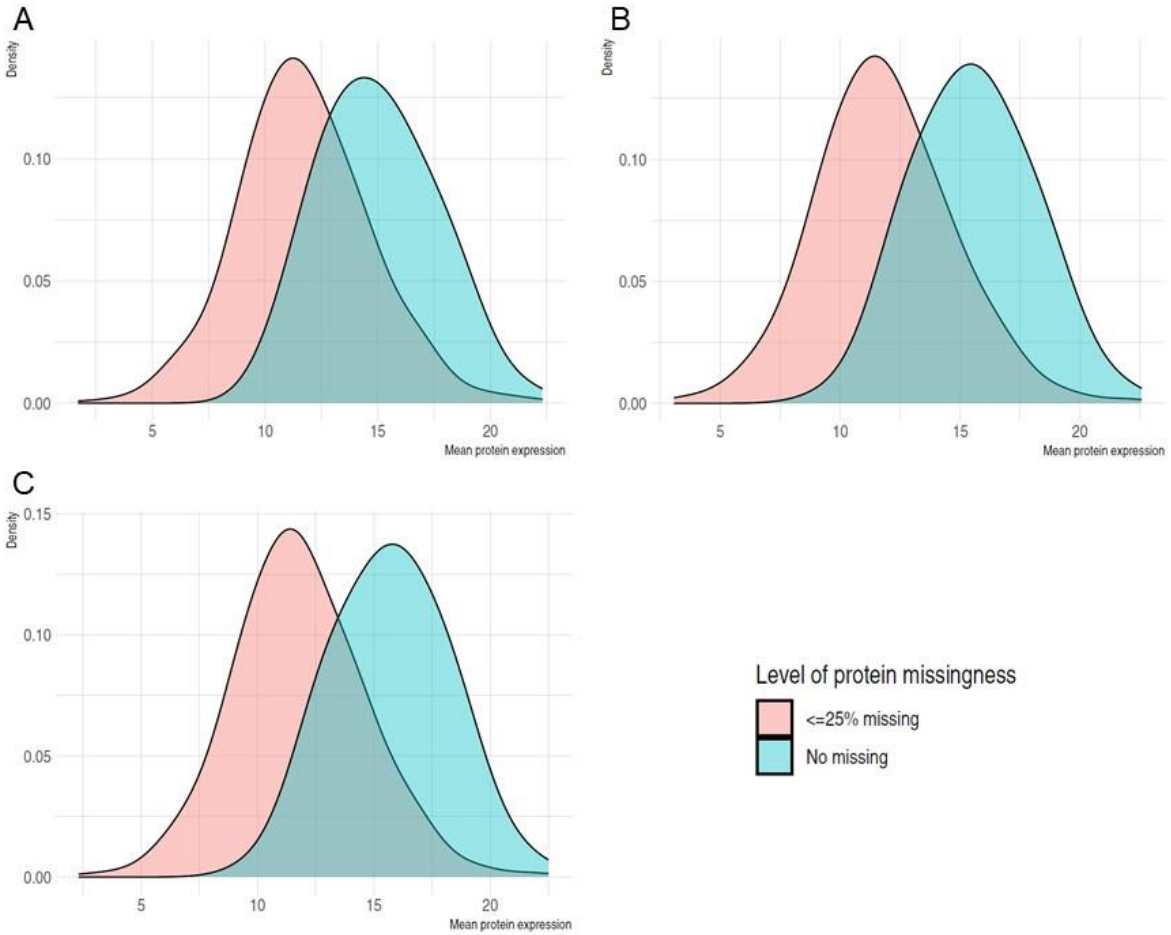


All multi-density plots demonstrate increased mean protein expression with lower levels of missingness. This would suggest that missing proteins may simply be caused by a lower read-rate at MS. Therefore, a decision was made to impute missing proteins, as it was likely that they were missing due to limitations in MS, and not due to physiological reasons.

5.4.2. Assessment of patterns of protein missingness

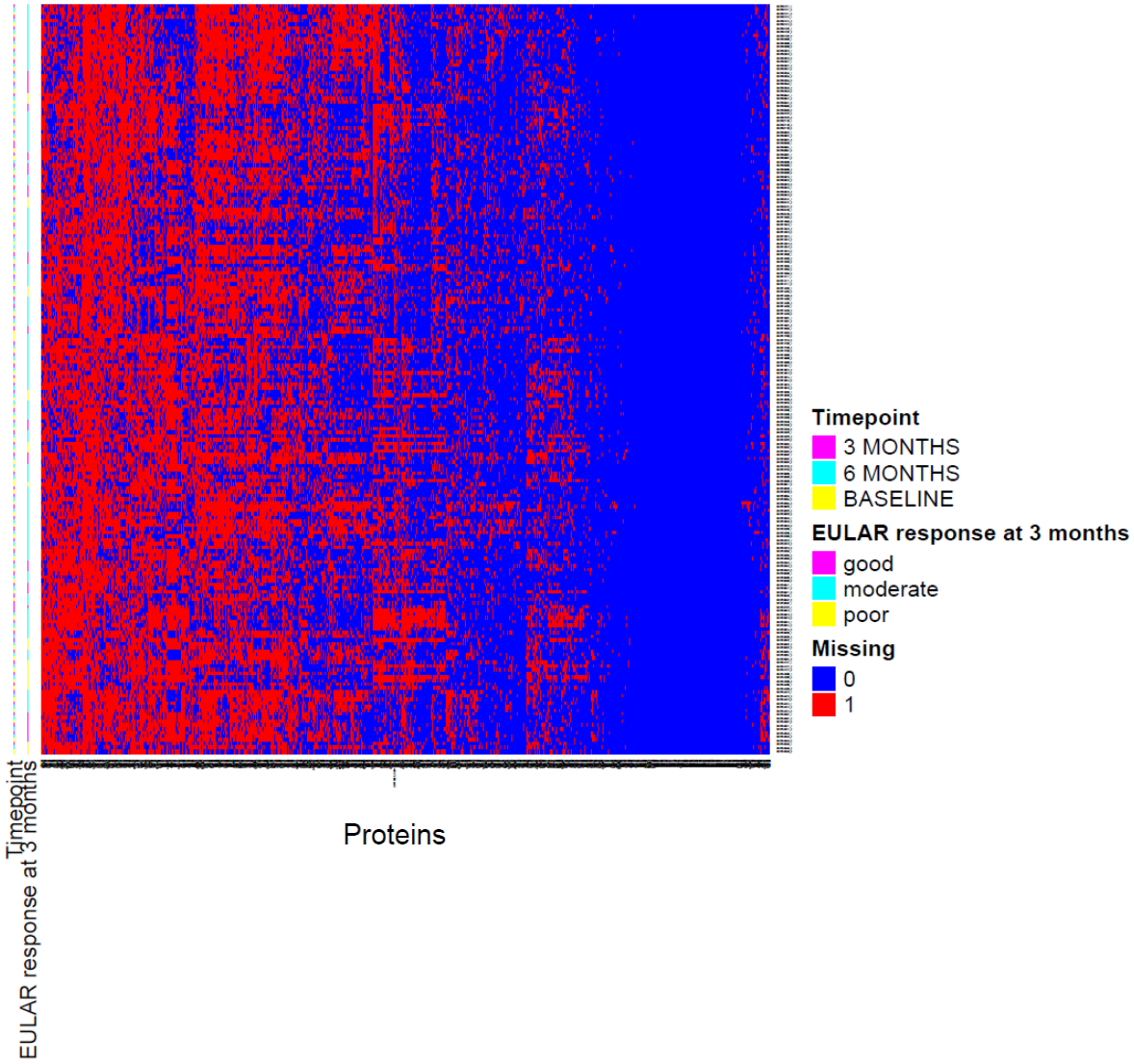
Heatmaps were generated for each batch to visualise any potential patterns of protein missingness, such as time point or treatment response (i.e. EULAR response); these are shown in Figures 5.6 – 5.8. Proteins demonstrated no clear patterns of missingness, neither according to time point of measurement nor EULAR response.

Figure 5.5. Mean protein expression multi-density plots for the etanercept sub-cohort patients from the third batch of SWATH-MS data to be processed at SBDC.



LEGEND: A) Proteins measured at baseline (pre-treatment). B) Proteins measured after three months of treatment. C) Proteins measured at all time points, combined.

Figure 5.6. Heatmap to assess protein missingness, first batch of etanercept sub-cohort samples. Samples are listed on the right-hand vertical axis.



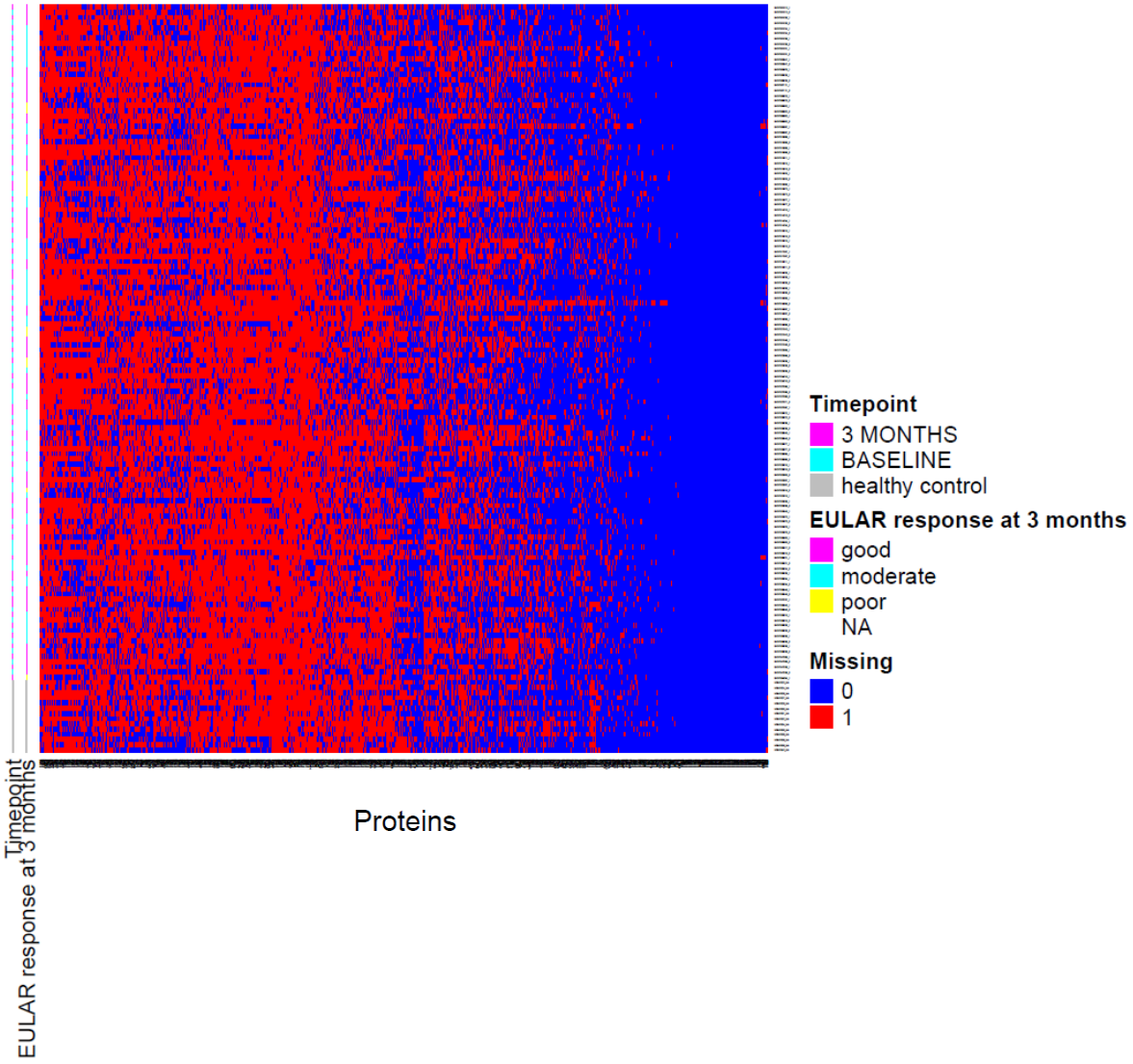
A high-quality PDF of this figure is available to download at:
https://github.com/oobergirl/thesis/blob/main/missingness_heatmap_batch_1.pdf

Table 5.7. Comparison of different imputation methods in simulated dataset.

Imputation method	**RMSE**
Lasso	0.3259
Partial least squares	0.3193
k-nearest neighbours	0.4871
MICE	0.4587
Random forest	0.2702

ABBREVIATIONS: Multiple imputation by chained equations (MICE), root mean square error (RMSE).

Figure 5.7. Heatmap to assess protein missingness, second batch of etanercept sub-cohort samples. Samples are listed on the right-hand vertical axis.



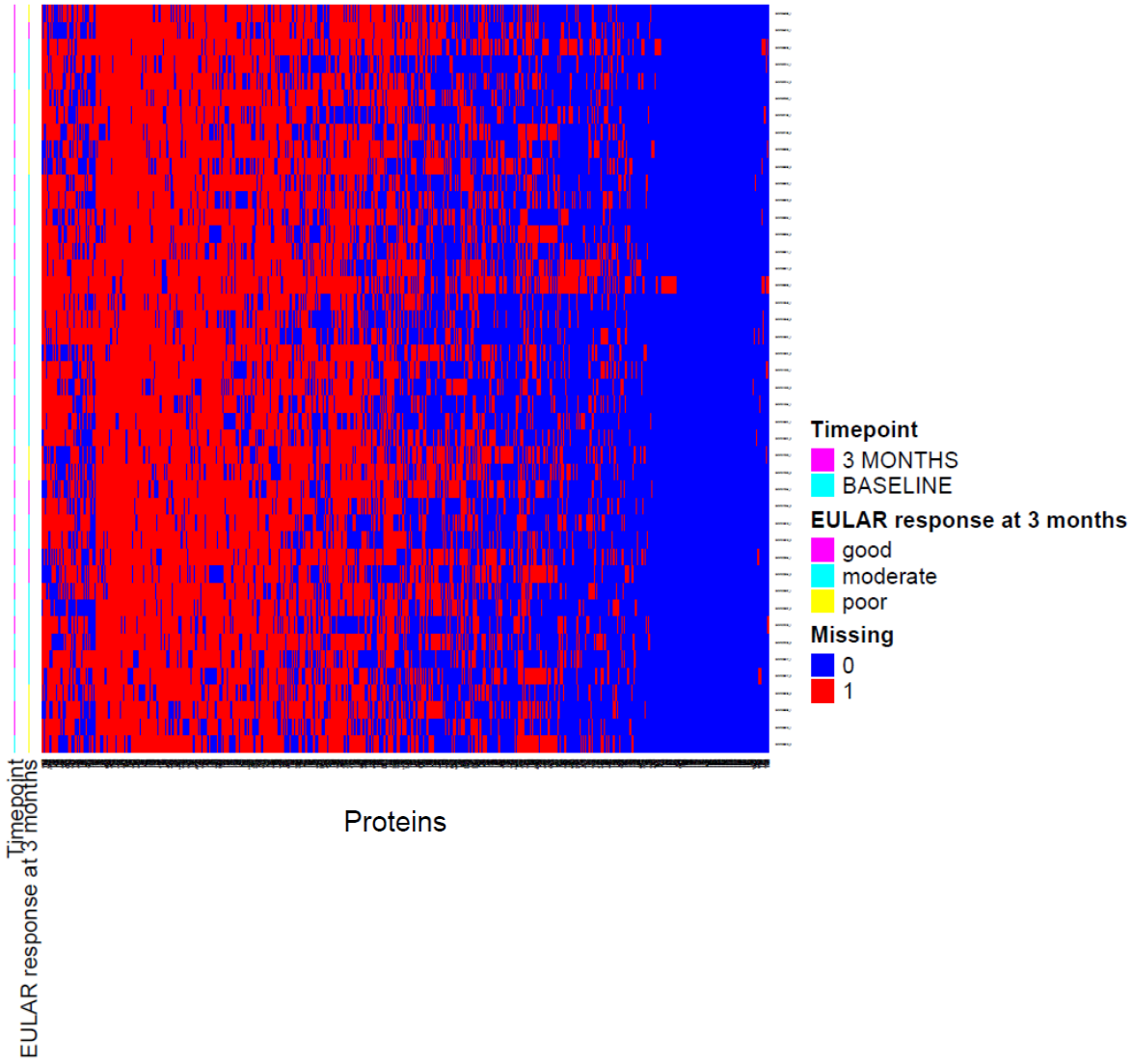
A high-quality PDF of this image is available to download at:
https://github.com/oobergirl/thesis/blob/main/missingness_heatmap_batch_2.pdf

Table 5.8. Comparison of random forest imputation across simulated datasets with different percentage missing values.

Total % missing values in simulated dataset	**RMSE**
25	0.2669
50	0.2738
75	0.2899

ABBREVIATIONS: Root mean square error (RMSE).

Figure 5.8. Heatmap to assess protein missingness, third batch of etanercept sub-cohort samples. Samples are listed on the right-hand vertical axis.



A high-quality PDF of this image is available to download at:
https://github.com/oobergirl/thesis/blob/main/missingness_heatmap_batch_3.pdf

5.4.3. Comparison of imputation methods on simulated dataset

80 patients from the first batch of the etanercept cohort to be processed for SWATH-MS were used to create a simulation dataset to compare different methods of imputation. 108 proteins with no missing values were retained, then 30% missing values were spiked into this dataset at random. Following imputation, imputed datasets were then compared to the original complete dataset of 108 proteins in 80 patients to determine the accuracy of imputation. The results of different imputation methods are shown in Table 5.7. The parameter of RMSE was used to compare methods, with a lower value representing improved accuracy. With the lowest RMSE of 0.2702, random forest was deemed to be the most accurate method of imputation for missing protein values.

New datasets were then simulated from the original complete dataset of 80 patients and 108 proteins, with values of 25%, 50% and 75% missing data randomly spiked in. Accuracy of random forest imputation was then compared across these new simulated datasets using RMSE, and this is detailed in Table 5.8. RMSE was comparable across all levels of percentage missing values, although this parameter increased as percentage missing values increased, which was not unexpected. However, because RMSE values were similarly low across different levels of missing values, a decision was made to impute all missing values, instead of excluding proteins with missing values above a certain threshold e.g. >75%.

5.4.4. Imputation of missing protein values in all patients with SWATH-MS data

Imputation was carried out in each batch of both the BRAGGSS-PD and etanercept cohorts using a random forest algorithm. Imputation was carried out at each time point for each batch of the etanercept cohort. Imputation was carried out on all samples from the BRAGGSS-PD cohort (all patients at all time points), despite being processed in two batches; this pragmatic approach was required due to low sample numbers. Following imputation, density plots were created with post-imputation data overlaid on pre-imputation data to visually inspect the effect of imputation on protein expression densities. Density plots demonstrate good agreement between pre- and post-imputation protein expression densities, and are shown in Figures 5.9 – 5.12.

Figure 5.9. Overlaid density plot comparing original with imputed data in first batch of etanercept sub-cohort.

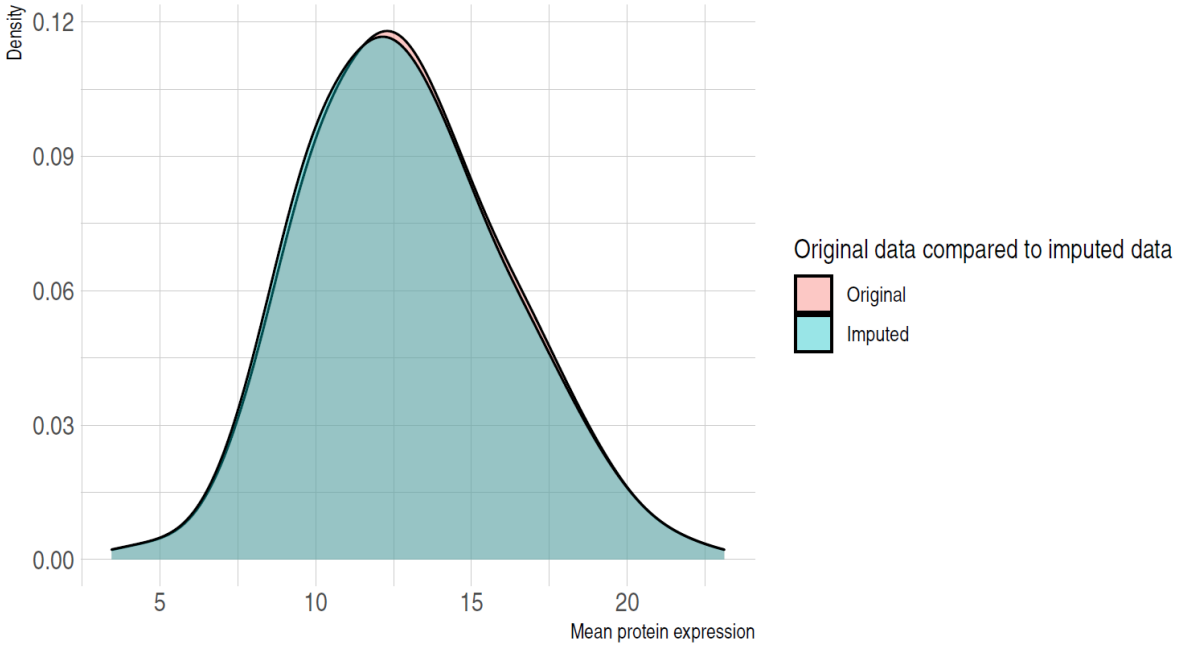


Figure 5.10. Overlaid density plot comparing original with imputed data in second batch of etanercept sub-cohort.

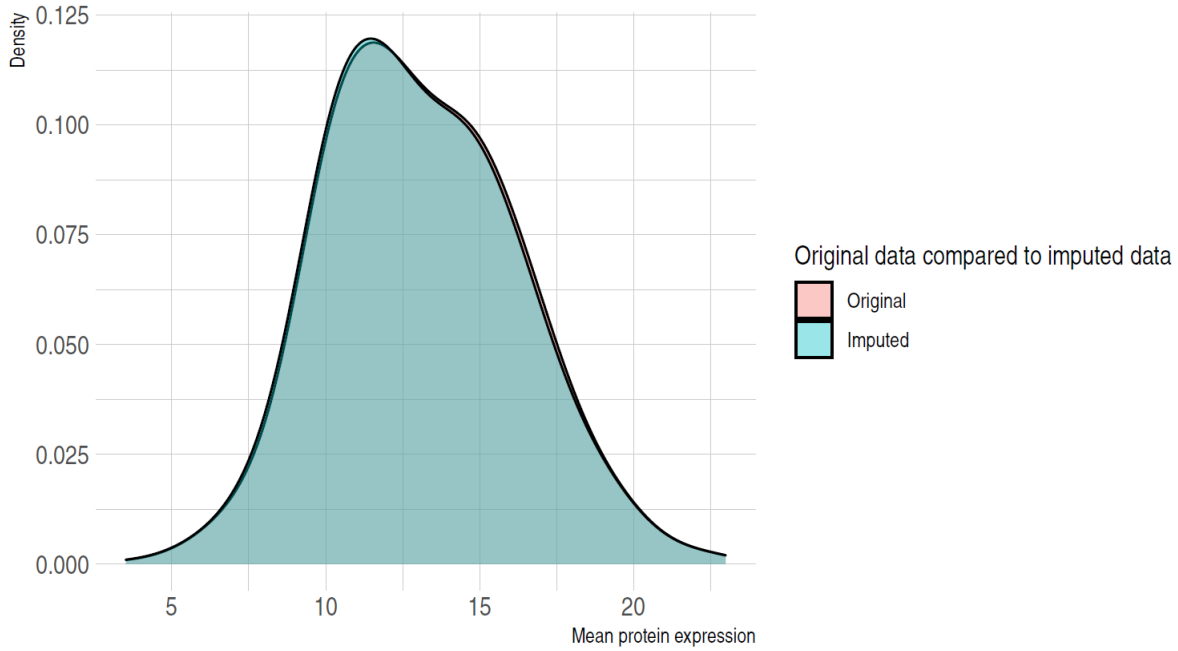


Figure 5.11. Overlaid density plot comparing original with imputed data in third batch of etanercept sub-cohort.

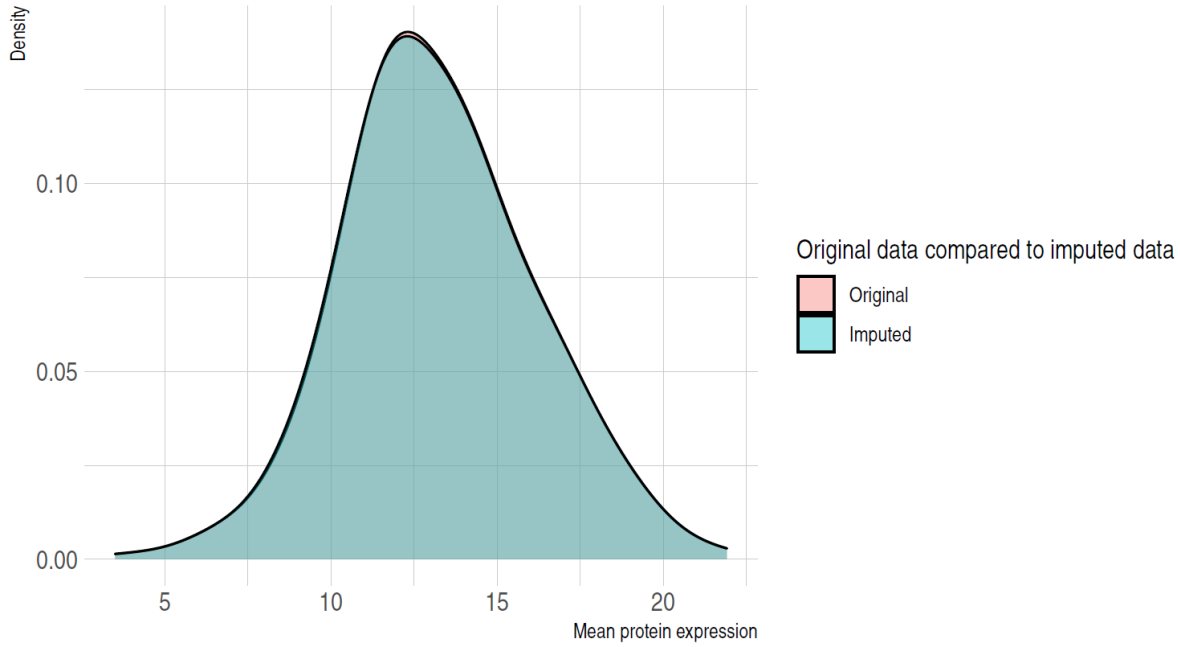
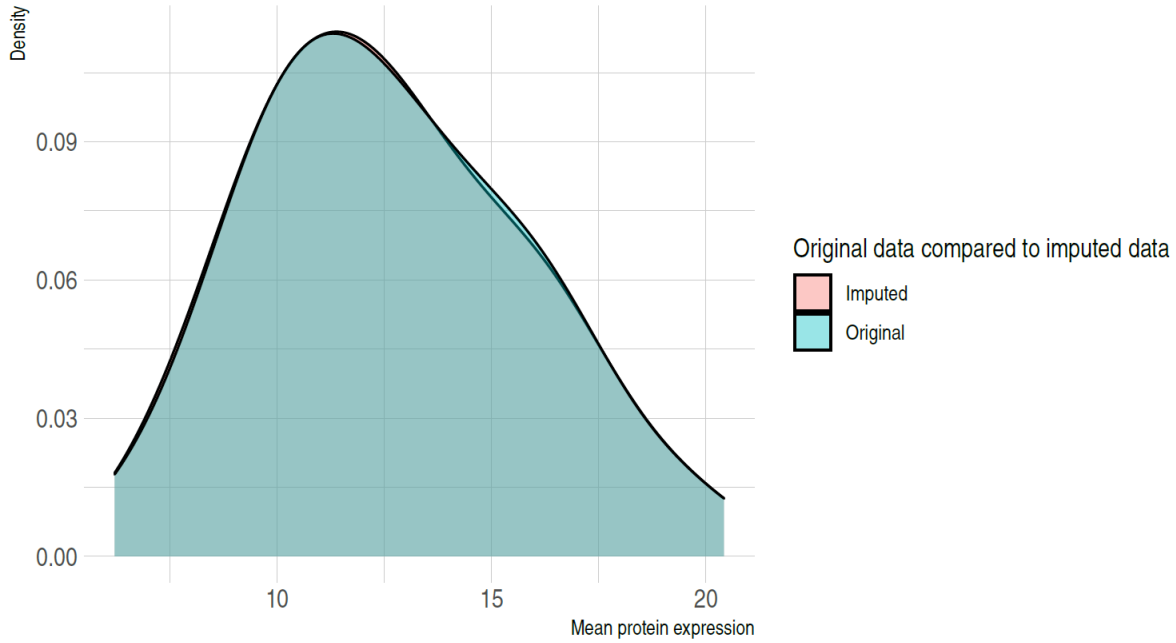


Figure 5.12. Overlaid density plot comparing original with imputed data in BRAGGSS-PD cohort.



5.4.5. Assessment for outliers and batch effect

Hierarchical clustering was carried out for the etanercept cohort samples at each time point and in all BRAGGSS-PD samples combined. Outlying samples were identified and removed. The resultant cluster dendrograms are presented in Figures 5.13 – 5.16. As Figure 5.13 demonstrates, one outlying sample was identified and removed from the etanercept sub-cohort at baseline; all other samples were retained for analysis.

PCA was then carried out for the etanercept sub-cohort and the BRAGGSS-PD cohort to determine whether there was any batch effect; the resultant plots are presented in Figures 5.17 and 5.18. Both cohorts demonstrated clear separation between batches that were processed for SWATH-MS at different times, but no separation by sampling time point. Therefore, samples from both cohorts underwent batch correction using a parametric empirical Bayes framework. Figures 5.19 and 5.20 show the re-run PCA after batch correction, which demonstrate no separation in samples due to batch effect.

Figure 5.13. Hierarchical cluster dendrogram to identify outliers, etanercept sub-cohort, baseline samples.

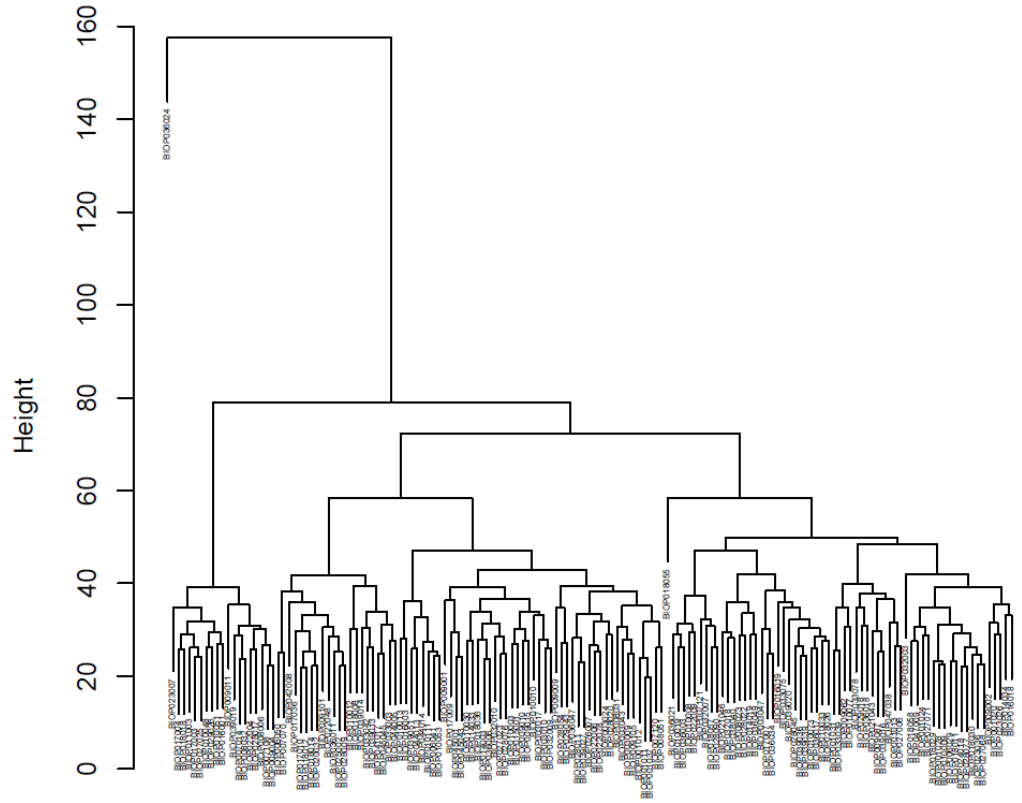


Figure 5.14. Hierarchical cluster dendrogram to identify outliers, etanercept sub-cohort, three month samples.

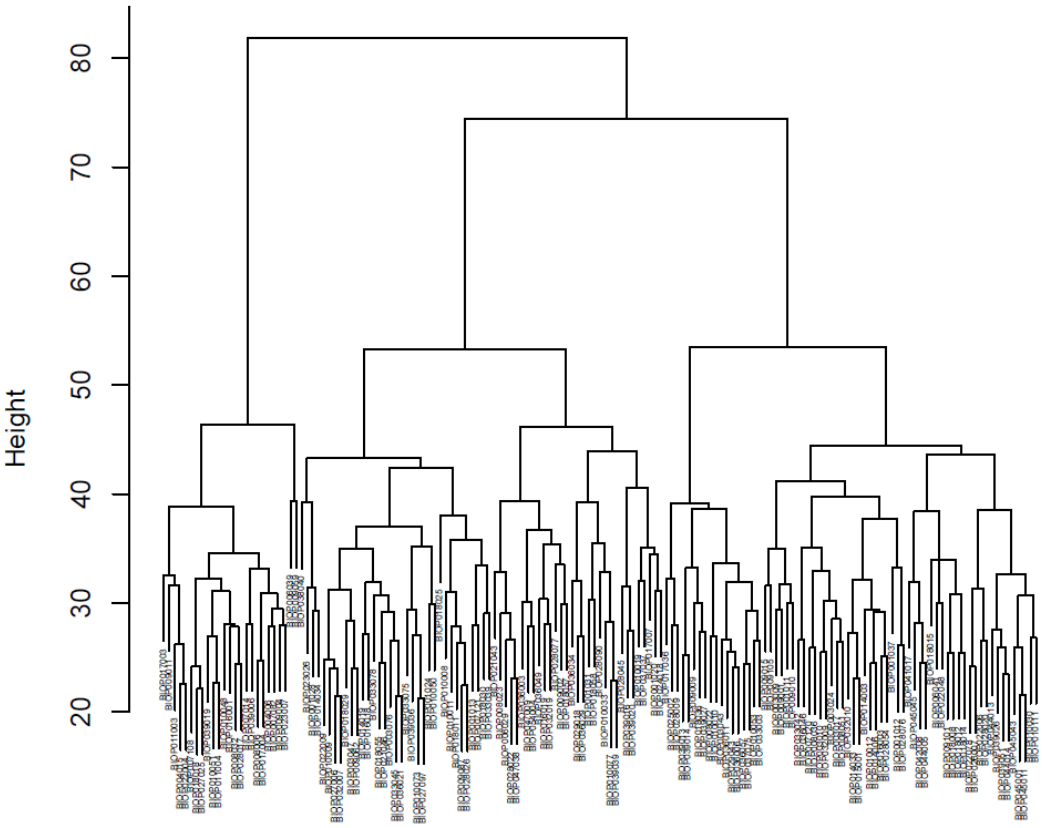


Figure 5.15. Hierarchical cluster dendrogram to identify outliers, etanercept sub-cohort, six month samples.

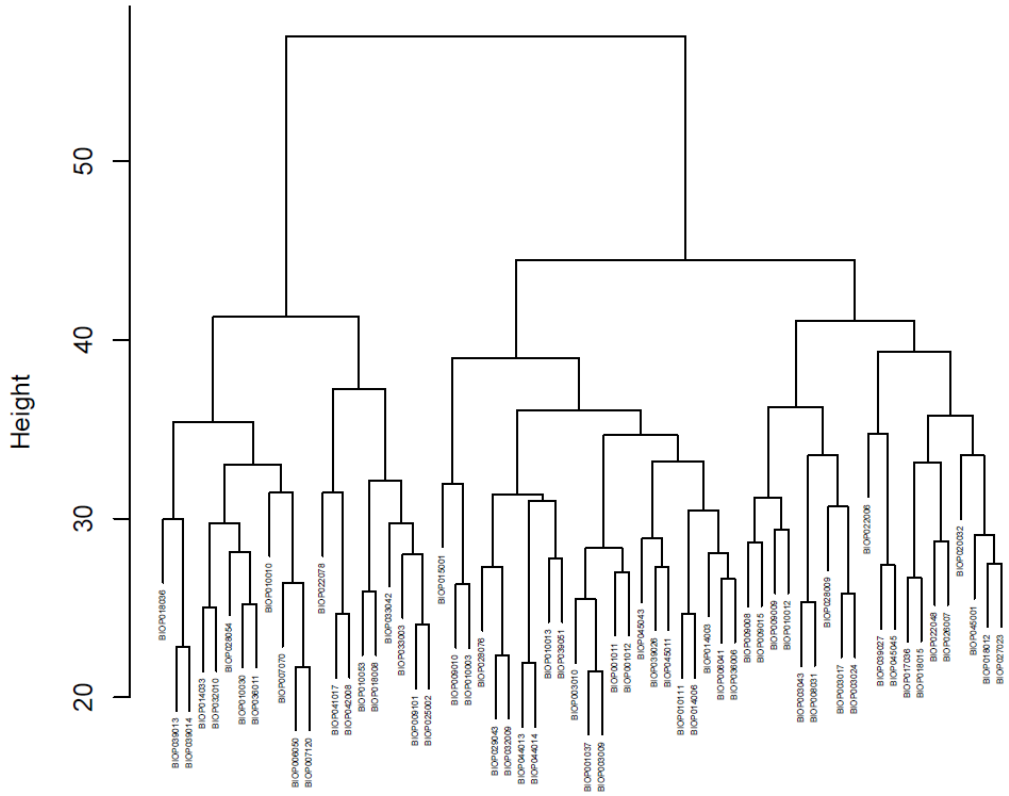


Figure 5.16. Hierarchical cluster dendrogram to identify outliers, BRAGGSS-PD cohort, all time points.

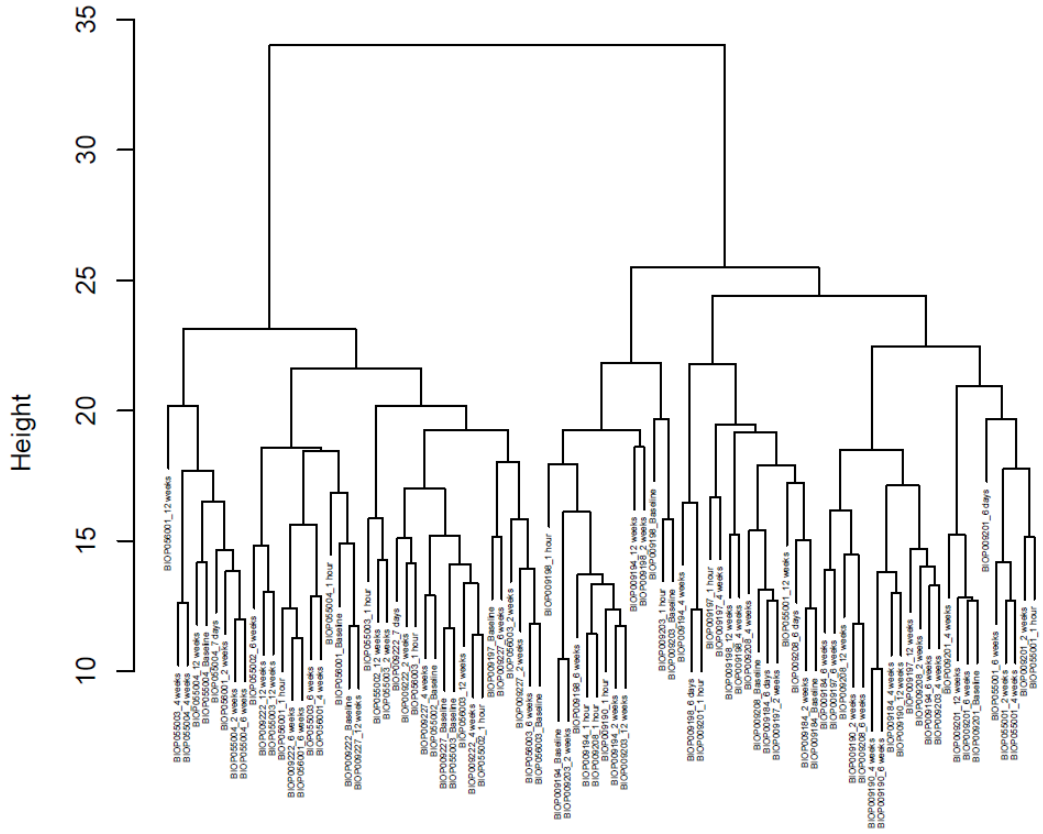


Figure 5.17. PCA plot to assess for batch or sampling time point effect, etanercept cohort.

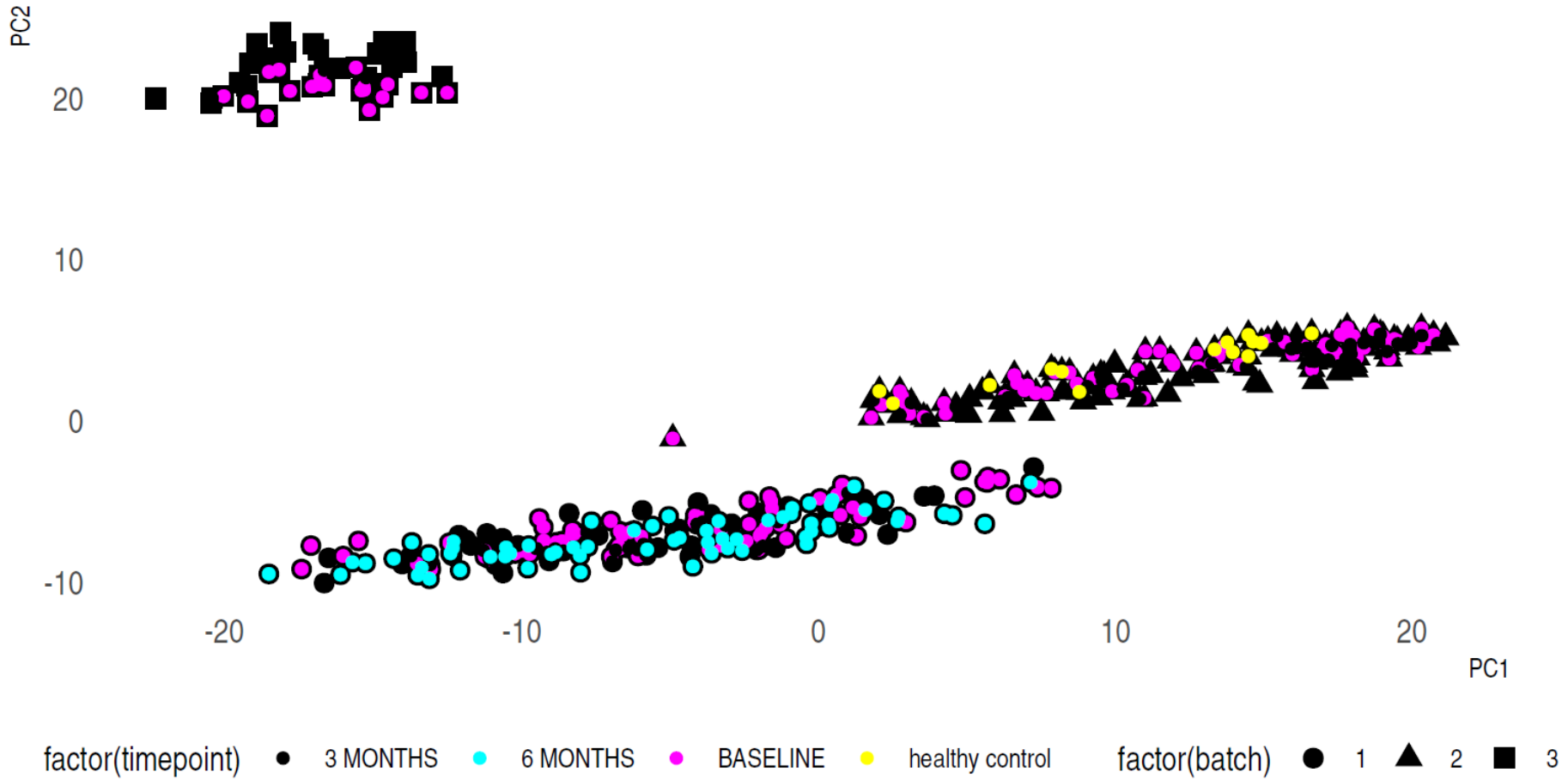


Figure 5.18. PCA plot to assess for batch or sampling time point effect, BRAGGSS-PD cohort.

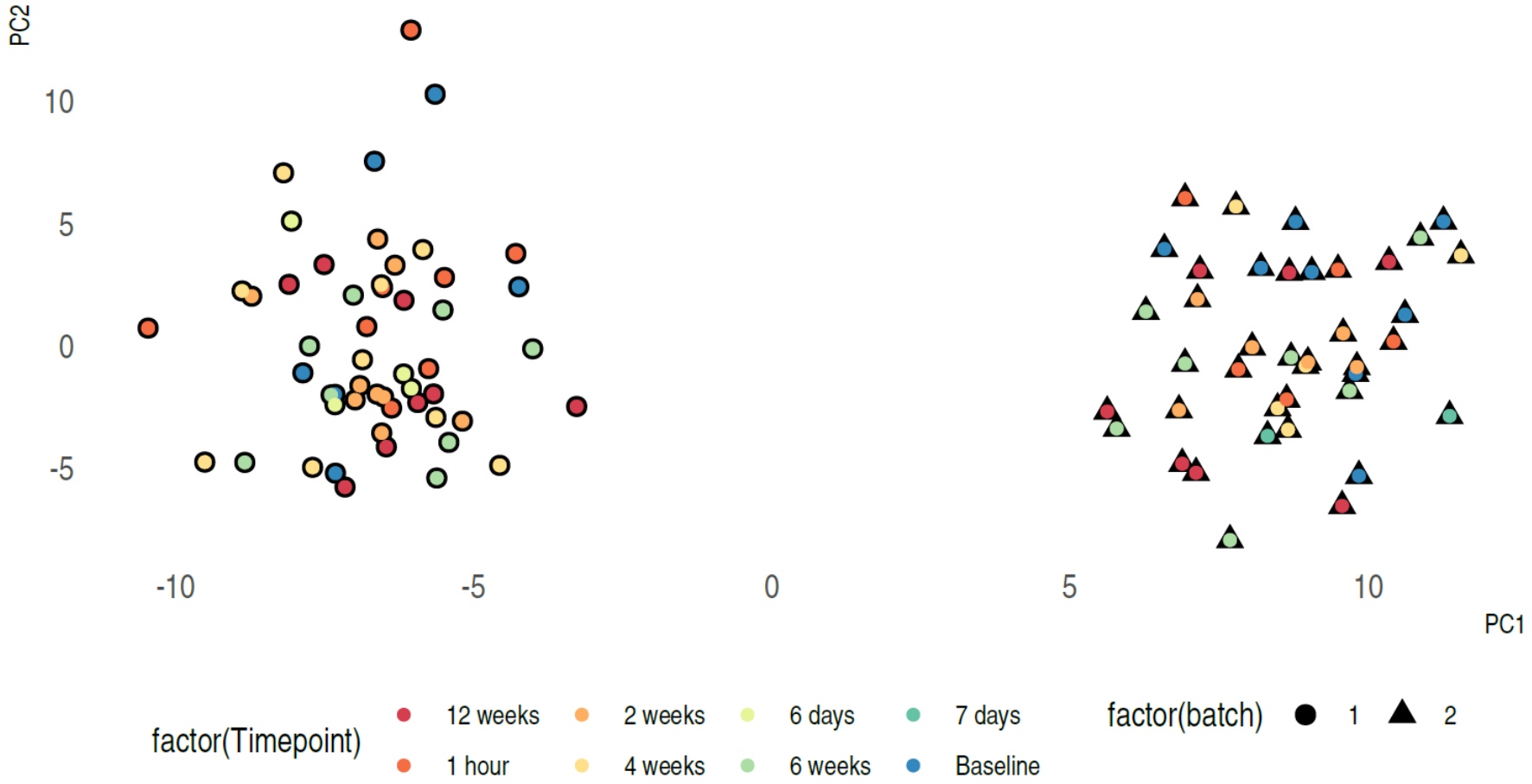


Figure 5.19. PCA plot after batch correction to assess for batch or sampling time point effect, etanercept cohort.

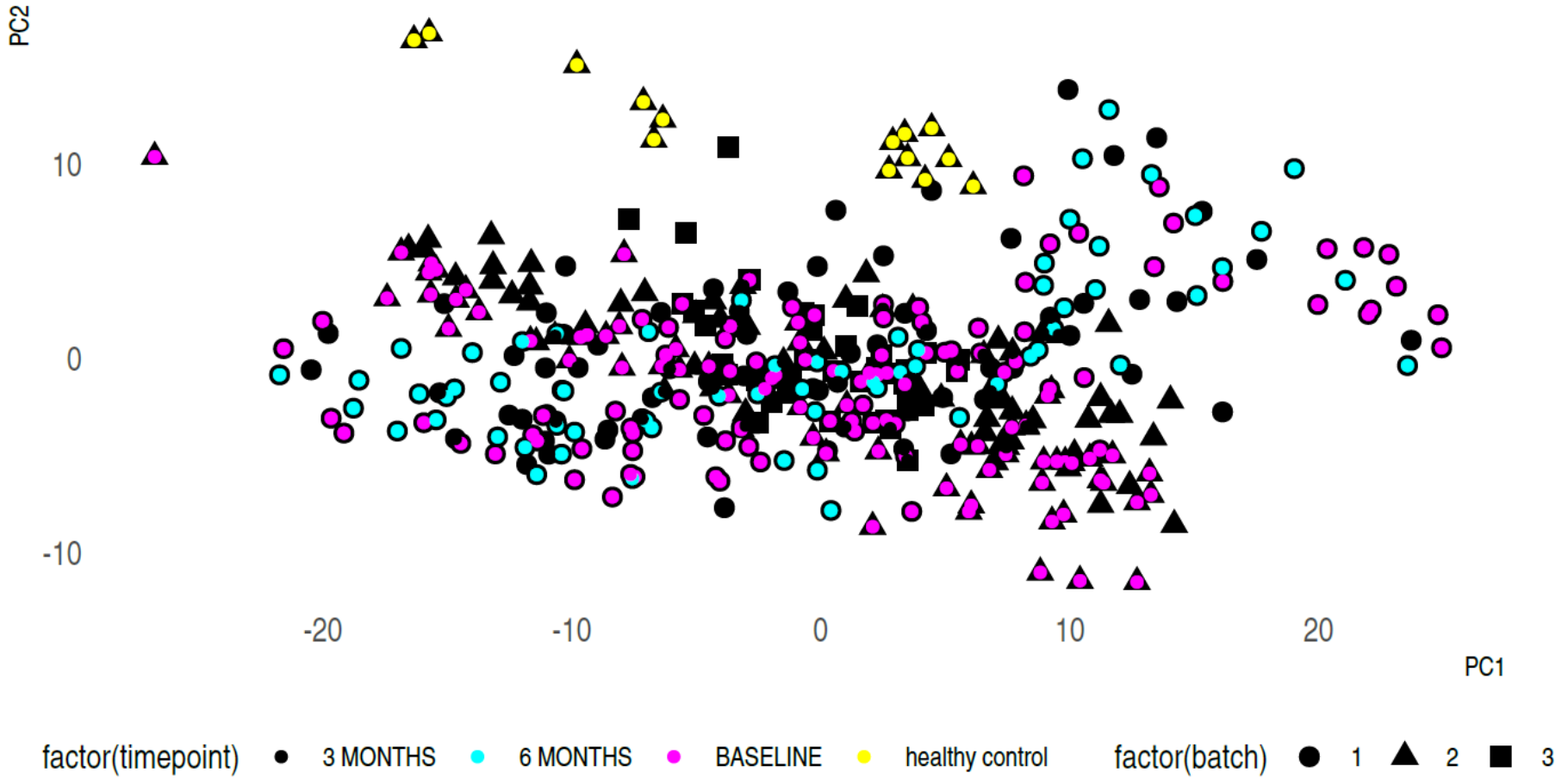
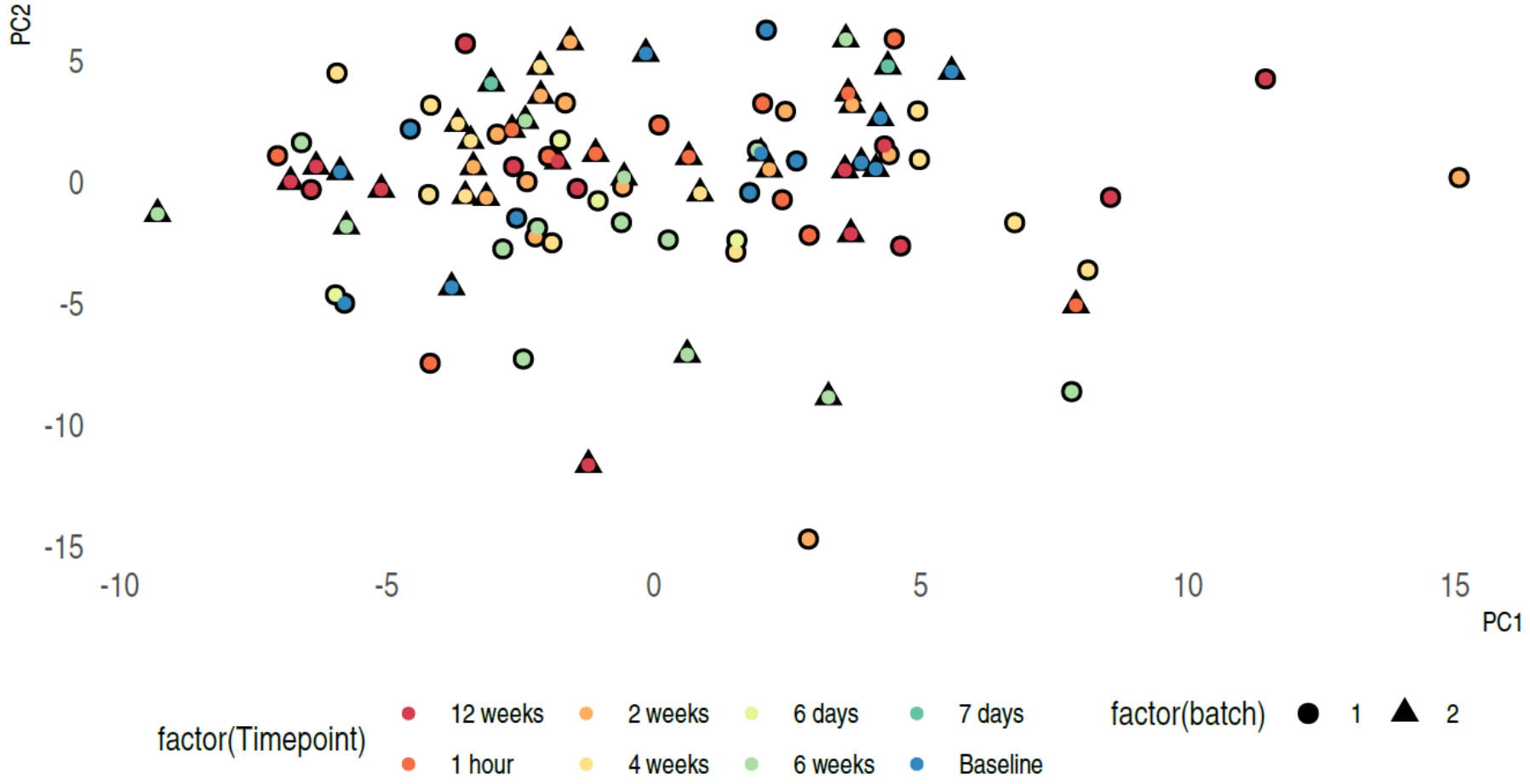


Figure 5.20. PCA plot after batch correction to assess for batch or sampling time point effect, BRAGGSS-PD cohort.



5.5. Differential expression of proteins between RA cases and HCs

Analysis was carried out in the second batch processed for SWATH-MS by the SBDC, as this batch contained a mixture of 73 patients with RA and 14 HCs. Samples from RA patients were filtered to the pre-treatment sampling time point prior to comparison with HCs, in order to reflect maximal RA disease activity.

216 proteins were found to be significantly differentially expressed in RA patients when compared with HCs. 179 proteins were from the RA protein library and the remaining 37 proteins were from the plasma protein library. 70 of the 216 proteins were down-regulated in RA patients with active disease compared to HCs, and the remaining proteins were all up-regulated. The full results are presented in Appendix Six. Proteins acquired at SWATH-MS for both etanercept and BRAGGSS-PD cohorts were filtered to include only these 216 significantly differentially expressed proteins for the remainder of analysis in order to focus analysis on protein measurements associated with RA, as well as improving power by reducing the number of comparisons made.

5.6. Longitudinal analysis of protein expression in the first 12 weeks of treatment with Amgevita or Benepali

16 patients with RA were included in analysis; their baseline characteristics are detailed in Tables 4.1 (Amgevita), 4.4 (Benepali) and 5.1 (whole BRAGGSS-PD cohort combined). Ten patients commencing Amgevita and six patients commencing Benepali were included. Serum proteomics acquired using SWATH-MS was available at the following sampling time points:

- Baseline (pre-treatment).
- 1 hour.
- 6 or 7 days (Benepali patients only).
- 2 weeks.
- 4 weeks.
- 6 weeks.
- 12 weeks.

A total of 188 of the 216 significant proteins from Section 5.1.5 were detectable in the BRAGGSS-PD cohort.

5.6.1. Linear mixed effects models of protein expression and achievement of therapeutic drug levels

All 16 BRAGGSS-PD patients were pooled for this analysis, as the dichotomous outcome variable of therapeutic drug levels made this outcome directly comparable between patients on different drugs, unlike the continuous outcome variable of drug concentration levels. Therapeutic drug levels were defined as between 5 – 8 mg/L for Amgevita (as per Pouw *et al*[219]) and between 2.1 – 4.7 mg/L for Benepali (as per Jamnitski *et al*[220]). Of the 15 patients with available drug concentration levels after 12 weeks of treatment, 11 patients (73.33%) had concentrations within the therapeutic ranges as defined above.

Five proteins were found to be significantly associated with therapeutic drug levels, following adjustment for age, biological sex, weight, concurrent csDMARD therapy, patient ID and sampling time point:

- CRP (UniProtID P02741), adjusted odds ratio (OR_{adj}) 0.45, 95% CI 0.22 – 0.94, p-value = 0.0342.
- Complement C4-B (C4B, UniProt ID P0C0L5), OR_{adj} 7.57, 95% CI 1.36 – 42.23, p-value = 0.0210.
- SAA2 (UniProt ID P0DJI9), OR_{adj} 0.23, 95% CI 0.06 – 0.95), p-value = 0.0416.
- LBP (UniProt ID P18428), OR_{adj} 0.10, 95% CI 0.02 – 0.64, p-value = 0.0145.
- Baculoviral IAP repeat-containing protein 2 (BIRC2, UniProt ID Q13490), OR_{adj} 0.18, 95% CI 0.05 – 0.66, p-value = 0.0099.

Full results for these models are detailed in Table 5.9. These five proteins were then placed into a multivariable model, detailed in Table 5.10. The only variables that remained significant when adjusted for other significant proteins were levels of C4B (OR_{adj} 11.45, 95% CI 1.42 – 92.60, p-value = 0.0223) and age at baseline (OR_{adj} 1.32, 95% CI 1.03 – 1.70, p-value = 0.0303).

Table 5.9. Protein levels significantly associated with therapeutic drug levels using linear mixed effects models in the BRAGGSS-PD cohort.

CRP (P02741)		
% missing pre-imputation: 10.51		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.2652 (3.20)	0.5149
Time point	8.0140 (96.80)	2.8309
Fixed effects		
Variable	**OR_{adj} (95% CI)**	**p-value**
P02741	0.45 (0.22 – 0.94)	0.0342
Age at baseline	1.15 (1.02 – 1.30)	0.0227
Female sex	0.96 (0.06 – 14.76)	0.9741
Weight	0.98 (0.93- 1.04)	0.4822
Concurrent csDMARD	0.79 (0.03 – 18.35)	0.8833
C4B (P0C0L5)		
% missing pre-imputation: 6.92		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.3974 (2.37)	0.6304
Time point	16.3465 (97.63)	4.0431
Fixed effects		
Variable	**OR_{adj} (95% CI)**	**p-value**
P0C0L5	7.57 (1.36 – 42.23)	0.0210
Age at baseline	1.19 (1.02 – 1.39)	0.0248
Female sex	12.46 (0.18 – 885.78)	0.2463
Weight	0.99 (0.93 – 1.05)	0.7630
Concurrent csDMARD	1.43 (0.02 – 95.47)	0.8661
SAA2 (P0DJI9)		
% missing pre-imputation: 65.90		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.3651 (2.93)	0.6042
Time point	12.0878 (97.07)	3.4768
Fixed effects		
Variable	**OR_{adj} (95% CI)**	**p-value**
P0DJI9	0.23 (0.06 – 0.95)	0.0416
Age at baseline	1.13 (0.99 – 1.28)	0.0703
Female sex	0.51 (0.03 – 9.02)	0.6429
Weight	0.97 (0.64 – 1.45)	0.2091
Concurrent csDMARD	1.69 (0.06 – 43.95)	0.7528

LBP (P18428)		
% missing pre-imputation: 0.51		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.0000 (0.00)	0.0000
Time point	10.4200 (100.00)	3.2270
Fixed effects		
Variable	**OR_{adj} (95% CI)**	**p-value**
P18428	0.10 (0.02 – 0.64)	0.0145
Age at baseline	1.27 (1.09 – 1.48)	0.0017
Female sex	3.45 (0.17 – 71.94)	0.4247
Weight	0.99 (0.94 – 1.04)	0.5867
Concurrent csDMARD	0.52 (0.02 – 11.34)	0.6774
BIRC2 (Q13490)		
% missing pre-imputation: 73.59		
Random effects		
Sample ID	2.64E-10 (0.00)	1.62E-05
Time point	16.89 (100.00)	2.1090
Fixed effects		
Variable	**OR_{adj} (95% CI)**	**p-value**
Q13490	0.18 (0.05 – 0.66)	0.0099
Age at baseline	1.23 (1.06 – 1.42)	0.0048
Female sex	1.56 (0.04 – 59.41)	0.8099
Weight	0.98 (0.93 – 1.03)	0.4418
Concurrent csDMARD	6.21 (0.19 – 205.92)	0.3069

ABBREVIATIONS: Adjusted (adj), baculoviral IAP repeat-containing protein 2 (BIRC2), complement C4-B (C4B), confidence interval (CI), conventional synthetic disease-modifying anti-rheumatic drug (csDMARD),C-reactive protein (CRP), identifier (ID), lipopolysaccharide-binding protein (LBP), odds ratio (OR), serum amyloid A2 protein (SAA2), standard deviation (SD).

Table 5.10. Multivariable model of protein levels significantly associated with therapeutic drug levels using linear mixed effects models in the BRAGGSS-PD cohort.

Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	1.211E-06 (0.00)	0.0011
Time point	14.74 (100.00)	3.8388
Fixed effects		
Variable	**OR_{adj} (95% CI)**	**p-value**
CRP (P02741)	0.47 (0.15 – 1.51)	0.2053
C4B (P0C0L5)	11.45 (1.42 – 92.60)	0.0223
SAA2 (P0DJI9)	0.50 (0.08 – 2.96)	0.4455
LBP (P18428)	0.39 (0.02 – 6.07)	0.5009
BIRC2 (Q13490)	0.24 (0.05 – 1.19)	0.0807
Age at baseline	1.32 (1.03 – 1.70)	0.0303
Female sex	23.22 (0.05 – 10158.54	0.3107
Weight	1.03 (0.95 – 1.12)	0.4702
Concurrent csDMARD	0.47 (0.00 – 99.91)	0.7812

ABBREVIATIONS: Adjusted (adj), baculoviral IAP repeat-containing protein 2 (BIRC2), complement C4-B (C4B), confidence interval (CI), conventional synthetic disease-modifying anti-rheumatic drug (csDMARD),C-reactive protein (CRP), identifier (ID), lipopolysaccharide-binding protein (LBP), odds ratio (OR), serum amyloid A2 protein (SAA2), standard deviation (SD).

5.6.2. Protein levels significantly associated with Amgevita or Benepali drug levels using linear mixed effects models in the BRAGGSS-PD cohort

Ten patients with RA commenced on Amgevita were included in this analysis. Nine proteins were found to be significantly associated with Amgevita drug concentration levels, following adjustment for age, biological sex, weight, concurrent csDMARD, patient ID and sampling time point:

- Filamin-B (FLNB, UniProt ID O75369), β-$coefficient_{adj}$ 1.10, 95% CI 0.23 – 1.96, p-value = 0.0179.
- CRP (UniProt ID 02741), β-$coefficient_{adj}$ -0.44, 95% CI -0.80 – (-0.07), p-value = 0.02311.
- α-1-acid glycoprotein 1 (A1AG1, UniProt ID P02763), β-$coefficient_{adj}$ -0.54, 95% CI -1.05 – (-0.03), p-value = 0.0494.
- Collagen α-2(VI) chain (COL6A2, UniProt ID P12110), β-$coefficient_{adj}$ -2.95, 95% CI -5.46 – (-0.43), p-value – 0.0429.

- LBP (UniProt ID P18428), β-coefficient$_{adj}$ -1.13, 95% CI -1.96 – (-0.41), p-value = 0.0037.
- Lysozyme C (LYZ, UniProt ID P61626), β-coefficient$_{adj}$ -0.83, 95% CI -1.61 – (-0.05), p-value = 0.0467.
- CD166 antigen (UniProt ID Q13740), β-coefficient 0.84$_{adj}$, 95% CI 0.26 – 1.41, p-value = 0.0070.
- Ester hydrolase C11orf54 (C11orf54, UniProt ID Q9H0W9), β-coefficient$_{adj}$ 1.55, 95% CI 0.34 – 2.70), p-value = 0.0160.
- Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit δ isoform (PIK3CD, UniProt ID O00329), β-coefficient$_{adj}$ 1.41, 95% CI 0.10 – 2.72, p-value = 0.0422.

Full results for these proteins are detailed in Table 5.11. These nine proteins were then placed into a multivariable model, detailed in Table 5.12. The only variables that remained significant when adjusted for other significant proteins were levels of A1AG1 (β-coefficient$_{adj}$ -0.76, 95% CI -1.31 – (-0.22), p-value = 0.0097) and female sex (β-coefficient$_{adj}$ 3.93, 95% CI 1.38 – 6.47, p-value = 0.0047).

Table 5.11. Protein levels significantly associated with Amgevita drug levels using linear mixed effects models in the BRAGGSS-PD cohort.

FLNB (O75369)		
% missing pre-imputation: 52.82		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.3472 (2.57)	0.5892
Time point	9.8832 (73.22)	3.1438
Residual	3.2684 (24.21)	1.8079
Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
O75369	1.10 (0.23 – 1.96)	0.0179
Age at baseline	0.06 (-0.01 – 0.14)	0.1646
Female sex	2.87 (-0.33 – 6.08)	0.1452
Weight	-0.02 (-0.05 – 0.01)	0.2778
Concurrent csDMARD	-0.91 (-3.51 – 1.70)	0.5236

CRP (P02741)		
% missing pre-imputation: 10.51		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.0000 (0.00)	0.0000
Time point	7.7450 (67.92)	2.7830
Residual	3.6580 (32.08)	1.9130
Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
P02741	-0.44 (-0.80 – (-0.07))	0.0231
Age at baseline	0.03 (-0.03 – 0.10)	0.2962
Female sex	4.64 (1.89 – 7.38)	0.0019
Weight	-0.01 (-0.04 – 0.02)	0.5805
Concurrent csDMARD	-2.35 (-4.53 – (-0.16))	0.0418
A1AG1 (P02763)		
% missing pre-imputation: 0.00		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.1570 (1.32)	0.3963
Time point	8.1120 (68.35)	2.8481
Residual	3.6000 (30.33)	1.8973
Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
P02763	-0.54 (-1.05 – (-0.03))	0.0494
Age at baseline	0.06 (-0.01 – 0.13)	0.1774
Female sex	4.93 (1.80 – 8.06)	0.0263
Weight	-0.01 (-0.05 – 0.02)	0.4533
Concurrent csDMARD	-1.80 (-4.15 – 0.54)	0.1920
COL6A2 (P12110)		
% missing pre-imputation: 60.77		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.0038 (0.03)	0.0616
Time point	9.0265 (71.41)	3.0044
Residual	3.6110 (28.56)	1.9003

Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
P12110	-2.95 (-5.46 – (-0.43))	0.0429
Age at baseline	0.04 (-0.02 – 0.10)	0.2872
Female sex	5.03 (2.17 – 7.89)	0.0219
Weight	-0.03 (-0.07 – 0.00)	0.0793
Concurrent csDMARD	-3.01 (-5.39 – (-0.64))	0.0515
LBP (P18428)		
% missing pre-imputation: 0.51		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.0000 (0.00)	0.0000
Time point	8.0820 (70.63)	2.8430
Residual	3.3610 (29.37)	1.8300
Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
P18428	-1.13 (-1.86 – (-0.41))	0.0037
Age at baseline	0.09 (0.03 – 0.16)	0.0098
Female sex	3.90 (1.37 – 6.44)	0.0043
Weight	-0.00 (-0.03 – 0.03)	0.9608
Concurrent csDMARD	-1.91 (-3.96 – 0.14)	0.0754
LYZ (P61626)		
% missing pre-imputation: 4.10		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.4880 (3.89)	0.6986
Time point	8.6800 (69.12)	0.29461
Residual	3.3900 (26.99)	1.8412
Random effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
P61626	-0.83 (-1.61 – (-0.05))	0.0467
Age at baseline	0.04 (-0.04 – 0.12)	0.3661
Female sex	4.59 (1.09 – 8.09)	0.0484
Weight	-0.02 (-0.05 – 0.02)	0.4347
Concurrent csDMARD	-2.75 (-5.60 – 0.10)	0.1066

CD166 antigen (Q13740)		
% missing pre-imputation: 50.77		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.6494 (4.69)	0.8059
Time point	10.1747 (73.52)	3.1898
Residual	3.0144 (21.78)	1.7362
Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
Q13740	0.84 (0.26 – 1.41)	0.0070
Age at baseline	0.03 (-0.55 – 0.60)	0.5210
Female sex	3.08 (-0.48 – 6.64)	0.1535
Weight	-0.02 (-0.06 – 0.02)	0.3541
Concurrent csDMARD	-1.72 (-4.53 – 1.09)	0.2839
C11orf54 (Q9H0W9)		
% missing pre-imputation: 49.23		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.8414 (6.21)	0.9173
Time point	9.5952 (70.87)	3.0976
Residual	3.1022 (22.91)	1.7613
Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
Q9H0W9	1.55 (0.34 – 2.70)	0.0160
Age at baseline	0.05 (-0.04 – 0.14)	0.3442
Female sex	2.51 (-1.43 – 6.45)	0.2722
Weight	-0.02 (-0.06 – 0.02)	0.4240
Concurrent csDMARD	-1.31 (-4.37 – 1.75)	0.4437
PIK3CD (O00329)		
% missing pre-imputation: 38.72		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.1991 (1.49)	0.4462
Time point	9.6740 (72.42)	3.1103
Residual	3.4855 (26.09)	1.8670

Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
O00329	1.41 (0.10 – 2.72)	0.0422
Age at baseline	0.04 (-0.03 – 0.11)	0.3735
Female sex	3.24 (0.25 – 6.23)	0.1037
Weight	-0.02 (-0.06 – 0.01)	0.1903
Concurrent csDMARD	-1.49 (-3.88 – 0.90)	0.2855

ABBREVIATIONS: α-1-acid glycoprotein 1 (A1AG1), adjusted (adj), cluster of differentiation (CD), collagen α-2(VI) chain (COL6A2), confidence interval (CI), conventional synthetic disease-modifying anti-rheumatic drug (csDMARD), C-reactive protein (CRP), ester hydrolase C11orf54 (C11orf54), filamin-B (FLNB), identifier (ID), lipopolysaccharide-binding protein (LBP), lysozyme C (LYZ), phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit δ isoform (PIK3CD), standard deviation (SD).

Six patients with RA commenced on Benepali were included in this analysis. Three proteins were found to be significantly associated with Benepali drug concentration levels, following adjustment for age, biological sex, weight, concurrent csDMARD, patient ID and sampling time point:

- Mannan-binding lectin serine protease 1 (MASP1, UniProt ID P48740), β-coefficient$_{adj}$ 0.87, 95% CI 0.11 – 1.64, p-value = 0.0371.
- Inhibitor of NF-κB kinase subunit α (IKKA, UniProt ID O15111), β-coefficient$_{adj}$ 0.92, 95% CI 0.26 – 1.59, p-value = 0.01317.
- Far upstream element-binding protein 1 (FUBP1, UniProt ID Q96AE4), β-coefficient$_{adj}$ 0.96, 95% CI 0.08 – 1.83), p-value = 0.0446).

Table 5.12. Multivariable model of protein levels significantly associated with Amgevita drug levels using linear mixed effects models in the BRAGGSS-PD cohort.

Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.0000 (0.00)	0.0000
Time point	7.2360 (75.44)	2.6900
Residual	2.3560 (24.56)	1.5350
Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
FLNB (O75369)	0.73 (-0.08 – 1.54)	0.08593
CRP (P02741)	0.11 (-0.37 – 0.59)	0.64838
A1AG1 (P02763)	-0.76 (-1.31 – (-0.22))	0.00970
COL6A2 (P12110)	-0.17 (-2.71 – 2.36)	0.89523
LBP (P18428)	-0.04 (-1.00 – 0.93)	0.94338
LYZ (P61626)	-0.72 (-1.46 – 0.03)	0.06912
CD166 antigen (Q13740)	0.49 (-0.13 – 1.12)	0.13160
C11orf54 (Q9H0W9)	1.22 (-0.00- 2.44)	0.05818
PIK3CD (O00329)	0.76 (-0.53 – 2.05)	0.25862
Age at baseline	0.07 (-0.00 – 0.15)	0.06404
Female sex	3.93 (1.38 – 6.47)	0.00472
Weight	0.02 (-0.02 – 0.05)	0.31179
Concurrent csDMARD	-1.31 (-3.46 – 0.85)	0.24237

ABBREVIATIONS: α-1-acid glycoprotein 1 (A1AG1), adjusted (adj), cluster of differentiation (CD), collagen α-2(VI) chain (COL6A2), confidence interval (CI), conventional synthetic disease-modifying anti-rheumatic drug (csDMARD), C-reactive protein (CRP), ester hydrolase C11orf54 (C11orf54), filamin-B (FLNB), identifier (ID), lipopolysaccharide-binding protein (LBP), lysozyme C (LYZ), phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit δ isoform (PIK3CD), standard deviation (SD).

Full results for these proteins are detailed in Table 5.13. Concurrent csDMARD therapy was dropped from all models due to rank deficiency because all patients with available information regarding this variable were receiving csDMARDs. The three significant proteins were then placed into a multivariable model, detailed in Table 5.14. The only variables that remained significant when adjusted for other significant proteins were levels of IKKA (β-coefficient$_{adj}$ 0.88, 95% CI 0.31 – 1.45, p=value = 0.0075), female sex (β-coefficient$_{adj}$ -4.21, 95% CI -7.24 – (-1.18), p-value = 0.0140) and weight (β-coefficient$_{adj}$ -0.18, 95% CI -0.26 – (-0.10), p-value = 0.0004).

Table 5.13. Protein levels significantly associated with Benepali drug levels using linear mixed effects models in the BRAGGSS-PD cohort.

MASP1 (P48740)		
% missing pre-imputation: 6.15		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.0000 (0.00)	0.0000
Time point	8.8410 (80.68)	2.9730
Residual	2.1170 (19.32)	1.4550
Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
P48740	0.87 (0.11 – 1.64)	0.0371
Age at baseline	-0.27 (-0.82 – 0.28)	0.3449
Female sex	-3.53 (-7.23 – 0.16)	0.0754
Weight	-0.15 (-0.24 – (-0.06))	0.0046
IKKA (O15111)		
48.46		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.0000 (0.00)	0.0000
Time point	9.9040 (83.90)	3.1470
Residual	1.9000 (16.10)	1.3780
Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
O15111	0.92 (0.26 – 1.59)	0.0132
Age at baseline	-0.14 (-0.62 – 0.35)	0.5826
Female sex	-2.23 (-5.50 – 1.04)	0.1963
Weight	-0.15 (-0.23 – (-0.06))	0.0032
FUBP1 (Q96AE4)		
% missing pre-imputation: 64.10		
Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.0000 (0.00)	0.0000
Time point	6.9760 (75.45)	2.6410
Residual	2.2700 (24.55)	1.5070
Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
Q96AE4	0.96 (0.08 – 1.83)	0.0446
Age at baseline	-0.05 (-0.57 – 0.47)	0.8471
Female sex	-3.23 (-6.98 – 0.52)	0.1032
Weight	-0.20 (-0.30 – (-0.09))	0.8012

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), identifier (ID), far upstream element-binding protein 1 (FUBP1), inhibitor of nuclear factor-κB kinase subunit α (IKKA), mannan-binding lectin serine protease 1 (MASP1), standard deviation (SD).

Table 5.14. Multivariable model of protein levels significantly associated with Benepali drug levels using linear mixed effects models in the BRAGGSS-PD cohort.

Random effects		
Variable	**Variance (% variance)**	**SD**
Sample ID	0.0000 (0.00)	0.0000
Time point	9.9610 (87.96)	3.1560
Residual	1.3640 (12.04)	1.1680
Fixed effects		
Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**
MASP (P48740)	0.59 (-0.07 – 1.24)	0.0955
IKKA (O15111)	0.88 (0.31 – 1.45)	0.0075
FUBP1 (Q96AE4)	0.71 (-0.01 – 1.43)	0.0687
Age at baseline	-0.32 (-0.77 – 0.12)	0.1701
Female sex	-4.21 (-7.24 – (-1.18))	0.0140
Weight	-0.18 (-0.26 – (-0.10))	0.0004

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), identifier (ID), far upstream element-binding protein 1 (FUBP1), inhibitor of nuclear factor-κB kinase subunit α (IKKA), mannan-binding lectin serine protease 1 (MASP1), standard deviation (SD).

5.7. Analysis of protein expression and association with RA clinical outcome measures following treatment with etanercept

180 patients with RA commencing etanercept were included in this analysis; baseline characteristics are detailed in Table 5.2. All patients had baseline and follow-up data recorded at three months, and 176/180 patients had follow-up data recorded at six months. Protein levels as detected using SWATH-MS were available on 152 patients at baseline, 159 patients at three months and 64 patients at six months.

5.7.1. Linear regression between protein levels and RA clinical outcome measures, continuous variables

5.7.1.1. Primary outcomes: DAS28 and ΔDAS28

5.7.1.1.1. DAS28

15 proteins measured at baseline were associated with DAS28 at that time point after adjustment for multiple testing and are presented in Appendix Seven, Table 1. 16 proteins measured at baseline were associated with DAS28 at that time point after adjustment for age, biological sex and RA disease duration and are presented in Appendix Seven, Table 2. Significant proteins from confounder-adjusted analysis were included in a multivariable model, along with age, sex and disease duration. Four proteins remained significant:

- Transferrin receptor protein 1 (TFRC, UniProt ID P02786), β-coefficient$_{adj}$ -0.15, 95% CI -0.26 – (-0.04), p-value = 0.0080.
- 14-3-3 protein η (YWHAH, UniProt ID Q04917), β-coefficient$_{adj}$ -0.18, 95% CI -0.32 – (-0.04), p-value = 0.0123.
- SAA1 (UniProt ID P0DJI8), β-coefficient$_{adj}$ 0.09, 95% CI 0.00 – 0.17, p-value = 0.0496.
- MAP2K3 (UniProt ID P46734), β-coefficient$_{adj}$ 0.36, 95% CI 0.09 – 0.64, p-value = 0.0110.

Full results of the multivariable analysis are presented in Appendix Seven, Table 3.

One protein measured at baseline were associated with DAS28 at three months following adjustment for multiple testing: EH domain-containing protein 1 (EHD1, UniProt ID Q9H4M9). However, after adjustment for age, sex and disease duration, two baseline proteins (including EHD1) were associated with DAS28 at three months and are presented in Appendix Seven, Table 4. These two proteins were then included in a multivariable model along with age, sex and disease duration and both remained significantly associated with DAS28 at three months:

- EHD1 (UniProt ID Q9H4M9), β-coefficient$_{adj}$ 0.25, 95% CI 0.09 – 0.41, p-value 0.0030.
- T-complex protein 1 subunit η (TCPH, UniProt ID Q99832), β-coefficient$_{adj}$ 0.67, 95% CI 0.21 – 1.13, p-value = 0.0050.

The full multivariable model is presented in Appendix Seven, Table 5.

No proteins detected at baseline were associated with DAS28 at six months following adjustment for multiple testing. There were no significant associations after adjustment for age, sex and disease duration.

36 proteins detected at three months were associated with DAS28 at the same time point after adjustment for multiple testing and are presented in Appendix Seven, Table 6. 28 proteins detected at three months were associated with DAS28 at this time point after adjustment for age, sex and disease duration and are presented in Appendix Seven, Table 7. Significant proteins from confounder-adjusted analysis were included in a multivariable model, alongside age, sex and disease duration. Four proteins remained significantly associated with DAS28:

- EHD1 (UniProt ID Q9H4M9), β-coefficient$_{adj}$ 0.17, 95% CI 0.02 – 0.32, p-value = 0.0239.
- Aspartyl/asparaginyl β-hydroxylase (ASPH, UniProt ID Q12797), β-coefficient$_{adj}$ -0.28, 95% CI -0.53 – (-0.04), p-value = 0.0249.
- CFHR3 (UniProt ID Q02985), β-coefficient$_{adj}$ 0.64, 95% CI 0.10 – 1.17, p-value = 0.0209.
- IGF1 (UniProt ID P05019), β-coefficient$_{adj}$ -0.28, 95% CI -0.54 – (-0.01), p-value = 0.0423.

The full results of the multivariable model are presented in Appendix Seven, Table 8.

10 proteins detected at three months were associated with DAS28 at six months after adjustment for multiple testing and are presented in Appendix Seven, Table 9. Seven proteins detected at three months were associated with DAS8 at six months after adjustment for age, sex and disease duration and are presented in Appendix Seven, Table 10. Significant proteins from confounder-adjusted analysis were included in a multivariable model, alongside age, sex and disease duration. Two proteins remained significantly associated with DAS28:

- ASPH (UniProt ID Q12797), β-coefficient$_{adj}$ -0.34, 95% CI -0.59 – (-0.10), p-value = 0.0076.
- CFHR3 (UniProt ID Q02985), β-coefficient$_{adj}$ 0.72, 95% CI 0.15 – 1.28, p-value = 0.0145.

The full results of the multivariable model are presented in Appendix Seven, Table 11.

One protein detected after six months of treatment with etanercept was associated with DAS28 at this time point after adjustment for multiple testing: CRP (UniProt ID P02741), β-coefficient 0.67, 95% CI 0.38 – 0.96, p-value = 2.37E-05, adjusted p-value = 0.0051. CRP remained significantly associated after adjustment for age, sex and disease duration: β-coefficient$_{adj}$ 0.67, 95% CI 0.37 – 0.96, p-value = 4.94E-05, adjusted p-value = 0.0107.

5.7.1.1.2. ΔDAS28

One protein detected before treatment with etanercept was associated with change in DAS28 (ΔDAS28) at three months after adjustment for multiple testing: EHD1 (UniProt ID Q9H4M9), β-coefficient -0.27, 95% CI -0.41 – (-0.14), p-value = 6.09E-05, adjusted p-value = 0.0132. A positive value for ΔDAS28 represents improvement in DAS28 and a negative number represents worsening DAS28 after treatment. EHD1 remained significantly associated after adjustment for age, sex and disease duration: β-coefficient$_{adj}$ -0.28, 95% CI -0.41 – (-0.15), p-value = 5.38E-05, adjusted p-value = 0.0012.

No proteins detected at baseline were associated with ΔDAS28 at six months following adjustment for multiple testing. There were still no significant associations after adjustment for age, sex and disease duration.

Eight proteins detected at three months were associated with ΔDAS28 at three months after adjustment for multiple testing and are presented in Appendix Seven, Table 12. Two proteins detected at three months were significantly associated with ΔDAS28 at three months after adjustment for age, sex and disease duration and are presented in Appendix Seven, Table 13. These two proteins were included in a multivariable model alongside age, sex and disease duration, and both remained significantly associated with ΔDAS28:

- EHD1 (UniProt ID Q9H4M9), β-coefficient$_{adj}$ -0.38, 95% CI -0.54 – (-0.22), p-value = 4.66E-06.
- CRP (UniProt ID P02741), β-coefficient$_{adj}$ -0.17, 95% CI -0.29 – (-0.05), p-value = 0.0047.

Full results are presented in Appendix Seven, Table 14.

No proteins detected at three months or six months were associated with ΔDAS28 at six months after adjustment for multiple testing. There were still no significant associations after adjustment for age, sex and disease duration.

5.7.1.2. Secondary outcomes: DAS28 sub-components

5.7.1.2.1. TJC

No proteins detected at baseline were associated with TJC at baseline, three months or six months after Benjamini-Hochberg adjustment for multiple testing. There were still no associated proteins after adjustment for age, biological sex and RA disease duration.

11 proteins detected at three months were associated with TJC at this time point following adjustment for multiple testing and are presented in Appendix Seven, Table 15. After adjustment for the potential confounders of age, sex and disease duration, eight proteins remained significantly associated following adjustment for multiple testing and are presented in Appendix Seven, Table 16. Significant proteins from analysis adjusted for potential confounders were then included in a multivariable model alongside age, sex and disease duration; three proteins remained significantly associated:

- TNF (UniProt ID P01375), β-coefficient$_{adj}$ -1.49, 95% CI -2.58 – (-0.41), p-value = 0.0079.
- Macrophage migration inhibitory factor (MIF, UniProt ID P14174), β-coefficient$_{adj}$ -4.21, 95% CI -6.45 – (-1.98), p-value = 0.0003.
- EHD1 (UniProt ID Q9H4M9), β-coefficient$_{adj}$ 0.82, 95% CI 0.16 – 1.48), p-value = 0.0158.

The full multivariable model is presented in Appendix Seven, Table 17.

No proteins detected at three months or six months were associated with TJC at six months following adjustment for multiple testing. There were still no associations after adjustment for age, sex and disease duration.

5.7.1.2.2. SJC

No proteins detected at baseline were associated with SJC at this time point or at six months following adjustment for multiple testing. There were still no associations after adjustment for age, biological sex and RA disease duration.

No proteins detected at baseline were associated with SJC at three months following adjustment for multiple testing. One protein was significant following adjustment for age, sex and disease duration: TCPH (UniProt ID Q99832), β-coefficient$_{adj}$ 0.05, 95% CI 0.02 – 0.07, p-value = 0.0002, adjusted p-value = 0.0421.

11 proteins detected at three months were associated with SJC at this time point following adjustment for multiple testing and are presented in Appendix Seven, Table 18. Six proteins remained significantly associated following adjustment for age, sex and disease duration and are presented in Appendix Seven, Table 19. Significant proteins from confounder-adjusted analysis were then included in a multivariable model alongside age, sex and disease duration. Five proteins remained significantly associated with SJC:

- Interleukin enhancer-binding factor 3 (ILF3, UniProt ID Q12906), β-coefficient$_{adj}$ -1.75, 95% CI -2.92 – (-0.69), p-value = 0.0016.
- Complement factor H-related protein 3 (CFHR3, UniProt ID Q02985), β-coefficient$_{adj}$ 1.39, 95% CI 0.29 – 2.49, p-value = 0.0143.
- CRP (UniProt ID P02741), β-coefficient$_{adj}$ 0.34, 95% CI 0.10 – 0.59, p-value = 0.0075.
- Insulin-like growth factor I (IGF1, UniProt ID P05019), β-coefficient$_{adj}$ -0.81, 95% CI -1.41 – (-0.21), p-value = 0.0091.
- TNF (UniProt ID P01375), β-coefficient$_{adj}$ -0.63, 95% CI -1.19 – (-0.07), p-value = 0.0290).

The full multivariable model is presented in Appendix Seven, Table 20.

No proteins detected at three months or six months were associated with SJC at six months following adjustment for multiple testing. There were still no associations after adjustment for age, sex and disease duration.

5.7.1.2.3. Patient-reported VAS of global health

No proteins detected at baseline were associated with patient-reported VAS of global health at baseline, three months or six months following adjustment for multiple testing. There were still no associations after adjustment for age, biological sex and RA disease duration.

Two proteins detected at three months were associated with VAS at this time point following adjustment for multiple testing:

- EHD1 (UniProt ID Q9H4M9), β-coefficient 0.02, 95% CI 0.01 – 0.03, p-value = 1.51E-05, adjusted p-value = 0.0033.
- ASPH (UniProt ID Q12797), β-coefficient -0.01, 95% CI -0.02 – (-0.01), p-value = 0.0004, adjusted p-value = 0.0417.

One protein remained significant following adjustment for age, sex and disease duration: EHD1 (UniProt ID Q9H4M9), β-coefficient$_{adj}$ 0.02, 95% CI 0.01 – 0.03, p-value = 4.71E-05, adjusted p-value = 0.0102.

One protein detected at three months was associated with VAS at six months following adjustment for multiple testing: ASPH (UniProt ID Q12797), β-coefficient -0.01, 95% CI -0.02 – (-0.01), p-value = 4.10E-05, adjusted p-value = 0.0089. ASPH remained significantly associated following adjustment for age, sex and disease duration: β-coefficient$_{adj}$ -0.01, 95% CI -0.02 – (-0.01), p-value = 0.0002, adjusted p-value = 0.0345.

No proteins detected at six months were associated with VAS at this time point following adjustment for multiple testing. There were still no associations after adjustment for age, sex and disease duration.

5.7.1.2.4. CRP measured using ELISA

CRP detected using SWATH-MS was strongly correlated with CRP measured using ELISA (β-coefficient 7.36, 95% CI 6.15 – 8.57, p-value < 2.2E-16), and so, was excluded from analysis. 49 proteins detected at baseline (excluding CRP detected using SWATH-MS) were associated with CRP measured using ELISA at this time point following adjustment for multiple testing, and are presented in Appendix Seven, Table 21. 50 proteins detected at baseline were associated with CRP measured using ELISA at this time point after adjustment for age, biological sex and RA disease duration and are presented in Appendix Seven, Table 22. Significant proteins from confounder-adjusted analysis were then included in a multivariable model alongside age, sex and disease duration. Nine proteins remained significantly associated with CRP and are presented in Table 5.15; the full multivariable model is presented in Appendix Seven, Table 23.

One protein detected at baseline was associated with CRP measured using ELISA at 3 months following adjustment for multiple testing: dual specificity mitogen-activated protein kinase kinase 3 (MAP2K3, UniProt ID P46734), β-coefficient 0.02, 95% CI 0.01 – 0.02, p-value = 0.0004, adjusted p-value = 0.0417. This significant association was lost following adjustment for age, sex and disease duration.

Table 5.15. Proteins measured before treatment with etanercept significantly associated with CRP at baseline after inclusion in a multivariable model.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**% missing values before imputation**
SAA1 (P0DJI8)	2.11 (0.40 – 3.82)	0.0172	26.92
PRS6A (P17980)	1.17 (5.82 – 17.66)	0.0002	55.13
SAA2 (P0DJI9)	3.48 (0.99 – 5.98)	0.0073	65.90
APOA4 (P06727)	2.87 (0.42 – 5.32)	0.0241	13.33
RBP4 (P02753)	-9.47 (-14.33 – (-4.61))	0.0002	0.00
CASP10 (Q92851)	-1.99 (-3.83 – (-0.15))	0.0365	27.18
TPP2 (P29144)	-6.30 (-10.63 – (-1.98))	0.0052	38.72
CFHR3 (Q02985)	-7.04 (-12.96 – (-1.12))	0.0217	66.67
ELANE (P08246)	2.70 (0.33 – 5.08)	0.0279	62.31

ABBREVIATIONS: 26S proteasome regulatory subunit 6A (PRS6A), adjusted (adj), apolipoprotein A-IV (APOA4), caspase-10 (CASP10), complement factor H-related protein 3 (CFHR3), confidence interval (CI), C-reactive protein (CRP), identifier (ID), neutrophil elastase (ELANE), retinol-binding protein 4 (RBP4), serum amyloid A-1 protein (SAA1), serum amyloid A-2 protein (SAA2), tripeptidyl-peptidase 2 (TPP2).

Three proteins detected at baseline were associated with CRP measured using ELISA at six months following adjustment for multiple testing and are presented in Appendix Seven, Table 24. The same three proteins from univariate analysis remained significant following adjustment for age, sex and disease duration and are presented in Appendix Seven, Table 25. These three proteins were then included in a multivariable model alongside age, sex and disease duration. All three proteins remained significant:

- Clathrin heavy chain 1 (CLTC, UniProt ID Q00610), β-coefficient$_{adj}$ 4.55, 95% CI 0.63 – 8.47, p-value = 0.0244.
- MAP2K3 (UniProt ID P46734), β-coefficient$_{adj}$ 7.26, 95% CI 3.09 – 11.43, p-value = 0.0008.
- Selenoprotein P (SELENOP, UnitProt ID P49908), β-coefficient$_{adj}$ -5.50 (-8.95 – (-2.05)), p-value = 0.0022.

Full results of the multivariable model are presented in Appendix Seven, Table 26.

20 proteins detected at three months were associated with CRP measured using ELISA at three months following adjustment for multiple testing and are presented in Appendix Seven,

Table 27. 21 proteins detected at three months were associated with CRP measured using ELISA at three months after adjustment for age, sex and disease duration and are presented in Appendix Seven, Table 28. Significant proteins from confounder-adjusted analysis were then included in a multivariable model alongside age, sex and disease duration. Two proteins remained significant in the multivariable model:

- LBP (UniProt ID P18428), β-coefficient$_{adj}$ 4.12, 95% CI 1.60 – 6.64, p-value 0.0017.
- N-α-acetyltransferase 25, NatB auxiliary subunit (NAA25, UniProt ID Q14CX7), β-coefficient$_{adj}$ -1.39, 95% CI -2.63 – (-0.15), p-value = 0.0303.

The full results of the multivariable model are presented in Appendix Seven, Table 29.

15 proteins detected at three months were associated with CRP measured using ELISA following adjustment for multiple testing and are presented in Appendix Seven, Table 30. Nine proteins detected at three months were associated with CRP measured using ELISA at six months after adjustment for age, sex and disease duration and are presented in Appendix Seven, Table 31. Significant proteins from confounder-adjusted analysis were included in a multivariable model, along with age, sex and disease duration. Two proteins remained significant in the multivariable model:

- MAP2K3 (UniProt ID P46734), β-coefficient$_{adj}$ 6.29, 95% CI 1.88 – 10.70, p-value = 0.0060.
- SAA1 (UniProt ID P0DJI8), β-coefficient$_{adj}$ 2.90, 95% CI 0.94 – 4.87, p-value = 0.0044.

Full results of the multivariable model are presented in Appendix Seven, Table 32.

24 proteins detected at six months were associated with CRP measured using ELISA at six months following adjustment for multiple testing and are presented in Appendix Seven, Table 33. 22 proteins detected at six months were associated with CRP measured using ELISA at six months after adjustment for age, sex and disease duration and are presented in Appendix Seven, Table 34. Significant proteins from confounder-adjusted analysis were included in a multivariable model, along with age, sex and disease duration. Four proteins remained significant in the multivariable model:

- CFHR5 (UniProt ID Q9BXR6), β-coefficient$_{adj}$ 6.28, 95% CI 1.15 – 11.42, p-value = 0.0215.

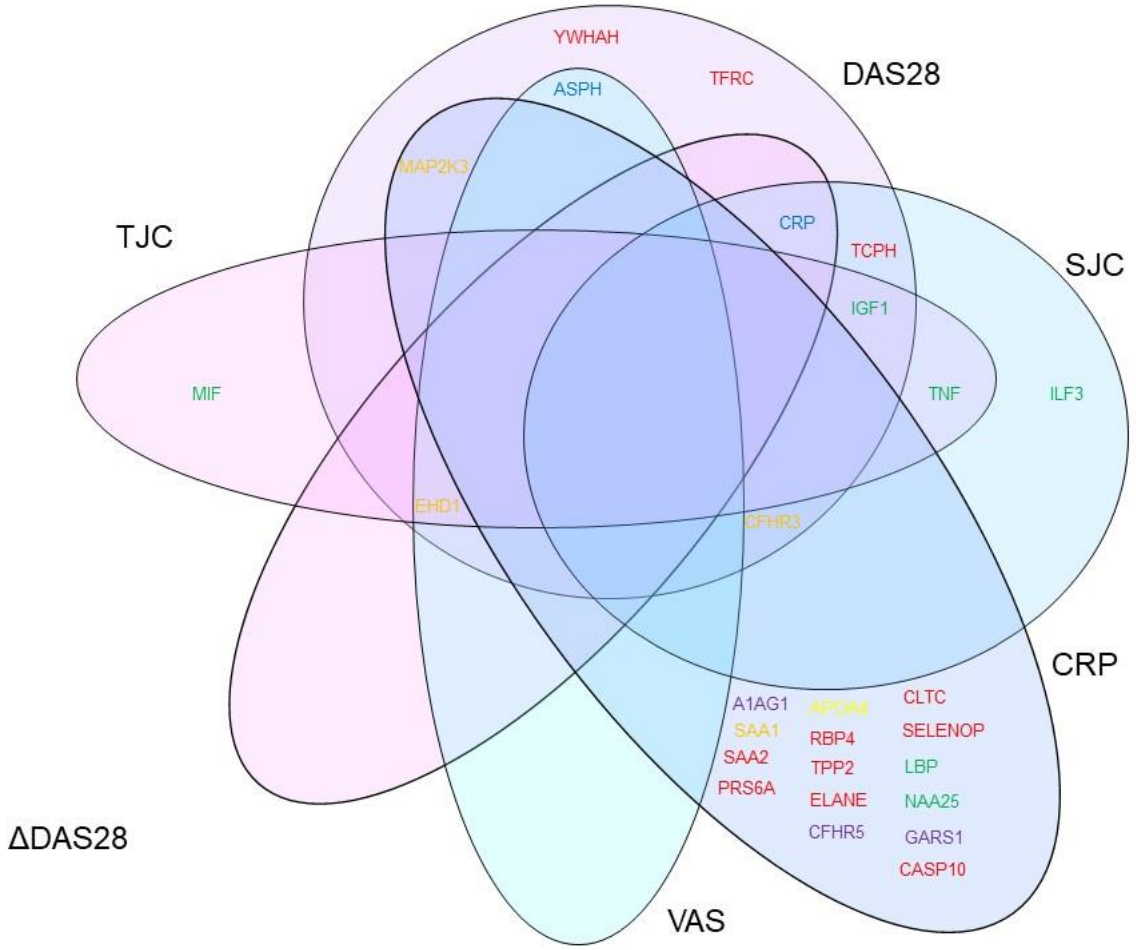
- APOA4 (UniProt ID P06727), β-coefficient$_{adj}$ 3.37, 95% CI 0.23 – 6.51, p-value = 0.0422.
- Glycine—tRNA ligase (GARS, UniProt ID P41250), β-coefficient$_{adj}$ -18.02, 95% CI -26.21 – (-9.83), p-value 0.0001.
- A1AG1 (UnitProt ID P02763), β-coefficient$_{adj}$ 2.67, 95% CI 0.90 – 6.71, p-value = 0.0143.

The full results of the multivariable model are presented in Appendix Seven, Table 5.35.

5.7.1.3. Summary of proteins associated with DAS28, ΔDAS28 and DAS28 components after linear regression analysis

After adjustment for multiple testing, potential confounders (age, biological sex, RA disease duration) and within multivariable models, a number of proteins were associated with continuous RA disease outcome measures, and are summarised in Figure 5.19 and Table 5.23.

Figure 5.21. Venn diagram of proteins associated with DAS28, ΔDAS28 and DAS28 components after linear regression analysis.



LEGEND: Proteins detected at baseline, proteins detected at baseline and three months, proteins detected at baseline and six months, proteins detected at three months, proteins detected at three and six months, proteins detected at six months.

ABBREVIATIONS: 14-3-3 protein η (YWHAH), aspartyl/asparaginyl β-hydroxylase (ASPH), caspase-10 (CASP10), complement factor H-related protein 3 (CFHR3), C-reactive protein (CRP), dual specificity mitogen-activated protein kinase kinase 3 (MAP2K3), EH domain-containing protein 1 (EHD1), insulin-like growth factor I (IGF1), interleukin enhancer-binding factor 3 (ILF3), macrophage migration inhibitory factor (MIF), T-complex protein 1 subunit η (TCPH), transferrin receptor protein 1 (TFRC), tumour necrosis factor (TNF).

5.7.2. Logistic regression between protein levels and RA disease outcomes, categorical variables

5.7.2.1. Poor EULAR response

One protein detected before treatment with etanercept was associated with a poor EULAR response at three months after adjustment for multiple testing: PKP3 (UniProt ID Q9Y446), OR 0.48, 95% CI 0.33 – 0.70, p-value = 0.0001, adjusted p-value = 0.0273. PKP3 remained

significant after adjustment for age, biological sex and disease duration: OR_{adj} 0.48, 95% CI 0.33 – 0.71, p-value = 0.0002, adjusted p-value = 0.0394.

No proteins detected at baseline were associated with a poor EULAR response at six months following adjustment for multiple testing. There were still no significant associations after adjustment for age, sex and disease duration.

Nine proteins detected at three months were associated with a poor EULAR response at that time point following adjustment for multiple testing and are presented in Appendix Seven, Table 36. Five proteins detected at three months were associated with a poor EULAR response at that time point following adjustment for age, sex and disease duration and are presented in Appendix Seven, Table 37. Significant proteins from confounder-adjusted analysis were included in a multivariable model alongside age, sex and disease duration and three proteins remained significant:

- ILF3 (UniProt ID Q12906), OR_{adj} 0.69, 95% CI 0.06 – 0.64, p-value = 0.0080.
- IGF1 (UniProt ID P05019), OR_{adj} 0.44, 95% CI 0.21 – 0.82, p-value = 0.0166.
- PKP3 (UniProt ID Q9Y446), OR_{adj} 0.57, 95% CI 0.36 – 0.87, p-value = 0.0127.

The full results of the multivariable analysis are presented in Appendix Seven, Table 38.

No proteins detected after three months or six months were associated with poor EULAR response at six months following adjustment for multiple testing. There were still no significant associations after adjustment for age, sex and disease duration.

5.7.2.2. Failure to achieve MCID in DAS28 (>-1.2)

One protein detected before treatment with etanercept was associated with failure to achieve an MCID in DAS28 at three months after adjustment for multiple testing: PKP3 (UniProt ID Q9Y446), OR 0.49, 95% CI 0.34 – 0.69, p-value = 5.62E-05, adjusted p-value = 0.0121). PKP3 remained significantly associated after adjustment for age, biological sex and RA disease duration: OR_{adj} 0.48, 95% CI 0.34 – 0.69, p-value = 8.40E-05, adjusted p-value = 0.0181.

No proteins detected at baseline were associated with failure to achieve an MCID in DAS28 at six months after adjustment for multiple testing. There were still no significant associations after adjustment for age, sex and disease duration.

Two proteins detected after three months of treatment with etanercept were associated with failure to achieve an MCID in DAS28 at that time point after adjustment for multiple testing and are presented in Appendix Seven, Table 39. The same two proteins remained significantly associated after adjustment for age, sex and disease duration and are presented in Appendix Seven, Table 40. These two proteins were included in a multivariable model with age, sex and disease duration. One protein remained significant: ASPH, OR_{adj} 0.61, 95% CI 0.36 – 0.96, p-value = 0.0451. The full results of the multivariable model are presented in Appendix Seven, Table 41.

One protein detected at three months was associated with failure to achieve an MCID in DAS28 at six months after adjustment for multiple testing: 26S proteasome regulatory subunit 6A (PRS6A, UniProt ID P17980), OR 3.48, 95% CI 1.79 – 6.75, p-value = 0.0002, adjusted p-value = 0.0487. The significant association was lost after adjustment for age, sex and disease duration.

No proteins detected at six months were associated with failure to achieve an MCID in DAS28 at that time point, following correction for multiple testing. There were still no significant associations after adjustment for age, sex and disease duration.

5.8. Differential expression of proteins over time following treatment with etanercept

Patients from the BRAGGSS etanercept sub-cohort were used in this analysis. Phenotype data from the BRAGGSS database was incorporated with protein expression data generated using SWATH-MS at the SBDC to carry out differential expression analysis, adjusted for age, biological sex, RA disease duration and pre-treatment DAS28. Demographic details of this cohort are detailed in Table 5.2.

5.8.1. Differential expression of proteins between sampling time points

5.8.1.1. Differential expression of proteins before treatment and after three months of treatment

133 patients with RA had available paired protein expression data between baseline (pre-treatment) and at three months follow-up (after/during treatment). 31 patients (23.31%) had a poor EULAR response after three months of treatment with etanercept, 37 patients (27.82%) had a moderate response and 65 patients (48.87%) had a good response. The good

and moderate responders were pooled into a good/moderate group for analysis, as predictors of poor response were the variables of interest.

11 proteins were significantly differentially expressed in poor responders between baseline and three months following Benjamini-Hochberg adjustment and are presented in Table 5.16. A positive-fold change indicates that a protein had increased expression at three months when compared to baseline. 49 proteins were differentially expressed in good/moderate responders between baseline and three months and are presented in Appendix Eight, Table 1. When protein expression was compared between baseline and three months in all patients, regardless of EULAR response, four proteins were differentially expressed and are presented in Appendix Eight, Table 2.

Table 5.16. Differentially expressed proteins between baseline and 3 months of treatment with etanercept in EULAR poor responders.

Protein (UniProt ID)	**Log-fold change**	**Average expression**	**p-value**	**Adjusted p-value**	**% missing values before imputation**
OGN (P20774)	-0.47	8.57	4.77E-07	0.0001	77.69
PARK7 (Q99497)	-1.67	13.41	1.28E-06	0.0001	72.05
PRDX3 (P30048)	-2.33	14.96	3.58E-06	0.0003	13.59
UGGT1 (Q9NYU2)	-1.19	9.84	7.53E-06	0.0004	69.74
AKT1 (P31749)	-0.76	9.66	3.32E-05	0.0014	67.44
CALD1 (Q05682)	1.51	11.21	9.28E-05	0.0033	76.67
CAMK1 (Q14012)	1.37	13.50	0.0001	0.0056	62.82
MIF (P14174)	-0.35	`0.52	0.0005	0.0116	71.54
NAMPT (P43490)	-1.73	12.03	0.0005	0.0116	53.08
LCN2 (P80188)	0.55	10.45	0.0009	0.0202	76.15
FLII (Q13045)	0.60	14.26	0.0013	0.0260	71.28

ABBREVIATIONS: Calcium/calmodulin-dependent protein kinase type 1 (CAMK1), caldesmon (CALD1), European League Against Rheumatism (EULAR), macrophage migration inhibitory factor (MIF), mimecan (OGN), neutrophil gelatinase-associated lipocalin (LCN2), nicotinamide phosphoribosyltransferase (NAMPT), Parkinson disease protein 7 (PARK7), protein flightless-1 homologue (FLII), RAC-α serine-threonine-protein kinase (AKT1), thioredoxin-dependent peroxide reductase, mitochondrial (PRDX3), UDP-glucose:glycoprotein glucosyltransferase 1 (UGGT1).

5.8.1.2. Differential expression of proteins before treatment and after six months of treatment

59 patients with RA had paired protein expression data between baseline and after six months of treatment. 15 patients (25.42%) had a poor EULAR response at this time point, 18 patients (30.51%) had a moderate response and 26 patients (44.07%) had a good response. Again, good and moderate responders were pooled into one categorical variable for comparison with poor responders.

Eight proteins were significantly differentially expressed in poor responders between baseline and six months following Benjamini-Hochberg adjustment and are presented in Appendix Eight, Table 3. 35 proteins were significantly differentially expressed in good/moderate responders between baseline and six months and are presented in Appendix Eight, Table 4. No proteins were significantly differentially expressed between baseline and six months when all patients were pooled.

5.8.1.3. Differential expression of proteins between 3 and 6 months of treatment

56 patients had paired protein expression data between three and six months of treatment with etanercept. 15 patients (26.79%) had a poor EULAR response at six months, 16 patients (28.57%) were moderate responders and 25 patients (44.64%) were good responders. Good and moderate responders were again pooled and analysed against poor responders.

Three proteins were differentially expressed between three and six months of treatment in poor responders, following Benejamini-Hochberg adjustment for multiple testing:

- Band 3 anion transport protein (B3AT, UniProt ID P02730), log-fold change 2.02, average expression 10.68, p-value = 4.10E-09, adjusted p-value = 8.86E-07.
- MAP2K3 (UniProt ID P46734), log-fold change 0.98, mean expression 10.80, p-value 8.08E-06, adjusted p-value = 0.0009.
- Leucine-rich repeat flightless-interacting protein 1 (LRRFIP1, UniProt ID Q32MZ4), log-fold change 0.91, average expression 14.56, p-value = 9.23E-05, adjusted p-value = 0.0066.

16 proteins were differentially expressed between three and six months of treatment in good/moderate responders and are presented in Appendix Eight, Table 5. No proteins were

significantly differentially expressed between three and six months of treatment when all patients were included in analysis, regardless of EULAR response.

5.8.2. Differential expression of proteins between responder statuses at the same time point

5.8.2.1. Proteins at baseline correlated with EULAR response by three months

152 patients with RA had protein expression data available at baseline with EULAR response data at three months. 34 patients (22.37%) had a poor EULAR response, 39 patients (25.66%) had a moderate response and 79 patients (51.97%) had a good response; good and moderate responders were again pooled for analysis. Nine proteins were significantly differentially expressed between EULAR good/moderate responders and poor responders; these are presented in Table 5.17. A positive-fold change indicates increased protein expression in the good/moderate response group.

The same 152 patients with protein expression data at baseline also had EULAR response data at six months. 39 patients (25.66%) were poor responders at six months, 31 patients (20.39%) were moderate responders and 82 patients (53.95%) were good responders. No proteins at baseline were significantly differentially expressed between EULAR good/moderate responders and poor responders at six months.

Table 5.17. Differentially expressed proteins at baseline between 3-month good/moderate and poor EULAR responders.

Protein (UniProt ID)	Log-fold change	Average expression	p-value	Adjusted p-value	% missing values before imputation
PKP3 (Q9Y446)	0.92	13.87	8.08E-05	0.0174	43.85
TCPH (Q99832)	-0.36	14.69	0.0002	0.0245	51.28
KRT1 (P04264)	0.51	13.45	0.0008	0.0460	56.41
CALD1 (Q05682)	1.09	10.75	0.0011	0.0460	76.67
LTBP1 (Q14766)	0.36	13.09	0.0012	0.0460	60.00
LRRFIP1 (Q32MZ4)	0.37	14.40	0.0013	0.0460	72.56
PDIA6 (Q15084)	0.55	12.67	0.0018	0.0460	58.46
CAMK1 (Q14012)	1.07	13.18	0.0019	0.0460	62.82
PRDX3 (P30048)	-1.06	15.52	0.0019	0.0460	13.59

ABBREVIATIONS: Calcium/calmodulin-dependent protein kinase type 1 (CAMK1), caldesmon (CALD1), European League Against Rheumatism (EULAR), keratin, type II cytoskeletal 1 (KRT1), latent-transforming growth factor β-binding protein 1 (LTBP1), leucine-rich repeat flightless-interacting protein 1 (LRRFIP1), plakophilin-3 (PKP3), protein disulphide-isomerase A6 (PDIA6), T-complex protein 1 subunit η (TCPH), thioredoxin-dependent peroxide reductase, mitochondrial (PRDX3).

5.8.2.2. Proteins at 3 months correlated with EULAR response by 6 months

159 patients with protein expression data after three months of treatment with etanercept had EULAR response data at six months available for analysis. 43 patients (27.04%) had a poor EULAR response, 33 patients (20.75%) had a moderate response and 83 patients (52.20%) had a good response. Good and moderate responders were pooled for analysis. 16 proteins were significantly differentially expressed between good/moderate responders and poor responders and are presented in Table 5.18.

Table 5.18. Differentially expressed proteins at 3 months between 6-month good/moderate and poor EULAR responders.

Protein (UniProt ID)	Log-fold change	Average expression	p-value	Adjusted p-value	% missing values before imputation
NAMPT (P43490)	1.87	11.69	2.36E-07	5.10E-05	53.08
ASPH (Q12797)	0.77	12.24	3.60E-05	0.0039	41.54
IGF1 (P05019)	0.55	9.49	0.0002	0.0131	78.97
PKP3 (Q9Y446)	0.82	14.23	0.0003	0.0131	43.85
TNF (P01375)	0.58	11.07	0.0003	0.0131	40.77
ILF3 (Q12906)	0.31	10.61	0.0004	0.0131	63.08
CRP (P02741)	-1.30	12.55	0.0005	0.0141	10.51
XRCC6 (P12956)	0.82	11.74	0.0010	0.0257	60.26
CFHR3 (Q02985)	-0.28	9.13	0.0011	0.0257	66.67
SAA1 (P0DJI8)	-0.91	9.97	0.0012	0.0257	26.92
CALM2 (P62158)	1.55	13.43	0.0016	0.0293	5.64
COL6A2 (P12110)	0.48	14.65	0.0016	0.0293	60.77
EHD1 (Q9H4M9)	-0.87	11.89	0.0018	0.0304	72.82
LBP (P18428)	-0.47	13.12	0.0021	0.0327	0.51
PPIA (P62937)	0.90	12.64	0.0026	0.0381	60.77
RSU1 (Q15404)	0.78	11.65	0.0030	0.0410	58.21

ABBREVIATIONS: Aspartyl/asparaginyl β-hydroxylase (ASPH), calmodulin-2 (CALM2), collagen α-2(VI) chain (COL6A2), complement factor H-related protein 3 (CFHR3), C-reactive protein (CRP), EH domain-containing protein 1 (EHD1), European League Against Rheumatism (EULAR), insulin-like growth factor I (IGF1), interleukin enhancer-binding factor 3 (ILF3), lipopolysaccharide-binding protein (LBP), nicotinamide phosphoribosyltransferase (NAMPT), peptidyl-prolyl cis-trans isomerase A (PPIA), plakophilin-3 (PKP3), Ras suppressor protein 1 (RSU1), serum amyloid A-1 protein (SAA1), tumour necrosis factor (TNF), X-ray repair cross-complementing protein 6 (XRCC6).

5.9. Machine learning methods to determine proteomic predictors of treatment response

Four different machine learning algorithms were used to build and benchmark predictive models using a combination of clinical variables and protein expression to predict either poor EULAR response or failure to achieve an MCID in DAS28. Baseline values were used

to predict outcomes at three and six months, and three-month values were used to predict outcomes at six months. During benchmarking in training datasets using nested resampling, the support vector machine algorithm was found to be the algorithm that produced the best model fits, as judged by the lowest MMCE (Table 5.19).

Support vector machine models were then built using full training datasets at each time point without nested resampling to obtain optimum hyperparameters. These optimum hyperparameters were then used to test predictive performance of each model on an independent test dataset at each time point. Predictive performance was poor, with the best predictive accuracy of 60.60% in the prediction of poor EULAR response at six months using baseline data. Full results are detailed in Table 5.20.

Calibration was checked on all models tested on independent datasets and was poor across all models. Detailed classifier and calibration plots for each model are detailed in Appendix Nine. Furthermore, the Hosmer-Lemeshow goodness-of-fit test was carried out on each model fitted on test datasets to assess model fit. Low p-values indicate poor fit. All models returned p-values of < 2.2E-16; full results are detailed in Table 5.21.

Table 5.19. Benchmarking results to assess selection of machine learning algorithms in training data.

Baseline variables and poor EULAR response at 3 months	
Model	**MMCE**
Penalised regression	0.1156
K-nearest neighbours	0.2263
Random forest	0.0968
Support vector machine	0.0137
Baseline variables and poor EULAR response at 6 months	
Model	**MMCE**
Penalised regression	0.2426
K-nearest neighbours	0.2862
Random forest	0.1890
Support vector machine	0.1333
Baseline variables and failure to achieve MCID in DAS28 at 3 months	
Model	**MMCE**
Penalised regression	0.0944
K-nearest neighbours	0.2511
Random forest	0.0980
Support vector machine	0.0320
Baseline variables and failure to achieve MCID in DAS28 at 6 months	
Model	**MMCE**
Penalised regression	0.1754
K-nearest neighbours	0.2879
Random forest	0.1603
Support vector machine	0.1027
3-month variables and poor EULAR response at 6 months	
Model	**MMCE**
Penalised regression	0.1887
K-nearest neighbours	0.1931
Random forest	0.1526
Support vector machine	0.0789
3-month variables and failure to achieve MCID in DAS28 at 6 months	
Model	**MMCE**
Penalised regression	0.2068
K-nearest neighbours	0.2573
Random forest	0.1586
Support vector machine	0.0862

ABBREVIATIONS: Disease Activity Score of 28 Joints (DAS28), European League Against Rheumatism (EULAR), mean misclassification error (MMCE), minimally clinically important difference (MCID).

Table 5.20. Performance of each support vector machine model in prediction on independent test datasets.

Baseline variables and poor EULAR response at 3 months	
Performance metric	
AUC	0.8318
Predictive accuracy	0.5238
MMCE	0.4762
Baseline variables and poor EULAR response at 6 months	
Performance metric	
AUC	0.6316
Predictive accuracy	0.6061
MMCE	0.3939
Baseline variables and failure to achieve MCID in DAS28 at 3 months	
Performance metric	
AUC	0.2722
Accuracy	0.4737
MMCE	0.5263
Baseline variables and failure to achieve MCID in DAS28 at 6 months	
Performance metric	
AUC	0.6754
Predictive accuracy	0.4595
MMCE	0.5405
3-month variables and poor EULAR response at 6 months	
Performance metric	
AUC	0.4737
Predictive accuracy	0.5152
MMCE	0.4848
3-month variables and failure to achieve MCID in DAS28 at 6 months	
Performance metric	
AUC	0.5188
Predictive accuracy	0.5455
MMCE	0.4545

ABBREVIATIONS: Area under the receiver operating characteristic curve (AUC), Disease Activity Score of 28 Joints (DAS28), European League Against Rheumatism (EULAR), minimally clinically important difference (MCID), mean misclassification error (MMCE).

Table 5.21. Hosmer-Lemeshow test results for each predictive model.

Model	χ^2	**Degrees of freedom**	**p-value**
Baseline variables, poor EULAR response at 3 months	39747	8	<2.2E-16
Baseline variables, poor EULAR response at 6 months	830.77	8	<2.2E-16
Baseline variables, failure to achieve MCID in DAS28 at 3 months	3073422	8	<2.2E-16
Baseline variables, failure to achieve MCID in DAS28 at 6 months	1224.7	8	<2.2E-16
3-month variables, poor EULAR response at 6 months	567.32	8	<2.2E-16
3-month variables, failure to achieve MCID in DAS28 at 6 months	1263.3	8	<2.2E-16

ABBREVIATIONS: Disease Activity Score of 28 Joints (DAS28), European League Against Rheumatism (EULAR), minimally clinically important difference (MCID).

5.10. Network analysis of proteins significantly associated with treatment response to adalimumab and/or etanercept

Proteins found to be significantly associated with treatment response were entered into the STRING Database v.11.5[232]. Only proteins that were significant following adjustment in multivariable models were included. 51 unique proteins were entered, and seven proteins were not joined to any other proteins via an interaction:

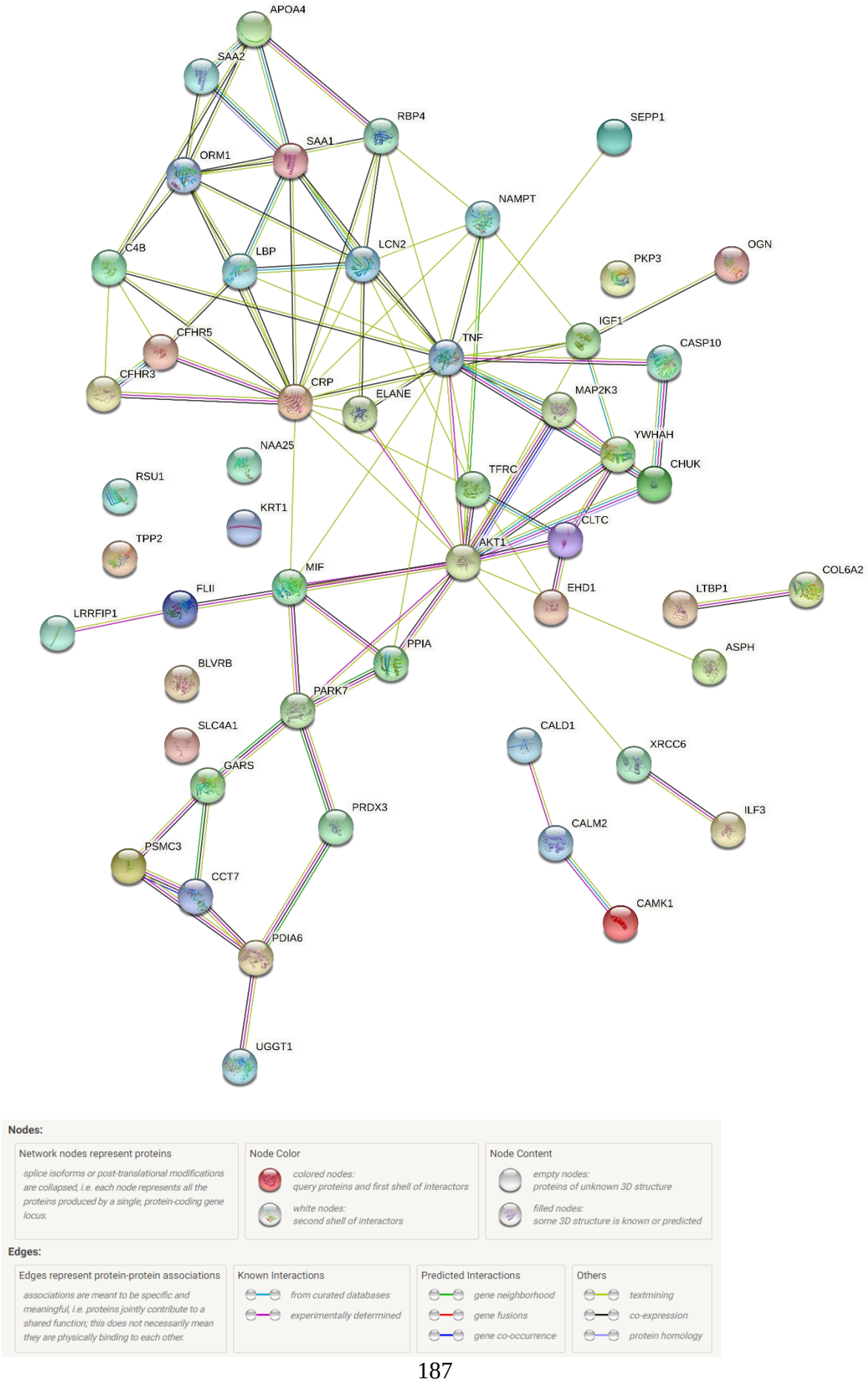
- Tripeptidyl-peptidase 2 (TPP2, UniProt ID P29144).
- NAA25 (UniProt ID Q14CX7).
- PKP3 (UniProt ID Q9Y446).
- Keratin, type II cytoskeletal 1 (KRT1, UniProt ID P04264).
- Ras suppressor protein 1 (RSU1, UniProt ID Q15404).
- Flavin reductase (NADPH) (BLVRB, UniProt ID P30043).
- Band 3 anion transport protein (SLC4A1, UniProt ID P02730).

All other proteins were joined in a network by a combination of known interactions (curated databases and published literature), predicted interactions (from gene neighbourhoods, gene fusions and gene co-occurrence), text mining, co-expression and protein homology. The full results of network analysis are presented in Figure 5.20.

5.11. Chapter summary

SWATH-MS was used to quantify proteins in 180 patients with RA and 14 HCs. Rigorous QC of protein data was carried out. Case-control analysis between RA patients and HCs identified proteins that were differentially expressed in RA patients, and further analysis focused on these proteins only. Using a variety of regression techniques, 53 proteins were found to be significantly associated with several different variables associated with RA disease activity, such as drug levels, DAS28 and its sub-components, EULAR response and achieving an MCID in DAS28. These proteins and their associations are summarised in Table 5.22. These proteins were then entered into the STRING database, and apart from seven proteins, all other proteins were found to have associations in a network linked to other proteins from the list of 53, indicating that there could be an underlying biological pathway behind most of the significant associations identified through analysis in this chapter.

Figure 5.22. Network analysis of proteins found to be significantly associated with treatment response to adalimumab and/or etanercept.



ABBREVIATIONS: 14-3-3 protein η (YWHAH), 26S proteasome regulatory subunit 6A (PSMC3, also known as PRS6A), α-1-acid glycoprotein 1 (ORM1, also known as A1AG1), apolipoprotein A-IV (APOA4), aspartyl/asparaginyl β-hydroxylase (ASPH), band 3 anion transport protein (SLC4A1), calcium/calmodulin-dependent protein kinase type 1 (CAMK1), caldesmon (CALD1), calmodulin-2 (CALM2), caspase-10 (CASP10), clathrin heavy chain 1 (CLTC), complement C4-B (C4B), complement factor H-related protein (CFHR), C-reactive protein (CRP), dual specificity mitogen-activated protein kinase kinase 3 (MAP2K3), EH domain-containing protein 1 (EHD1), flavin reductase (NADPH) (BLVRB), glycine—tRNA ligase (GARS), inhibitor of NF-κB kinase subunit α (CHUK, also known as IKKA), insulin-like growth factor I (IGF1), interleukin enhancer-binding factor 3 (ILF3), keratin, type II cytoskeletal 1 (KRT1), latent-transforming growth factor β-binding protein 1 (LTBP1), leucine-rich repeat flightless-interacting protein 1 (LRRFIP1), lipopolysaccharide-binding protein (LBP), macrophage migration inhibitory factor (MIF), mimecan (OGN), N-α-acetyltransferase 25, NatB auxiliary subunit (NAA25), neutrophil elastase (ELANE), neutrophil gelatinase-associated lipocalin (LCN2), nicotinamide phosphoribosyltransferase (NAMPT), Parkinson disease protein 7 (PARK7), peptidyl-prolyl cis-trans isomerase A (PPIA), plakophilin-3 (PKP3), protein disulfide-isomerase A6 (PDIA6), protein flightless-1 homologue (FLII), RAC-α serine/threonine-protein kinase (AKT1), Ras suppressor protein 1 (RSU1), retinol-binding protein 4 (RBP4), selenoprotein P (SEPP1), serum amyloid A (SAA), T-complex protein 1 subunit η (CCT7, also known as TCPH), thioredoxin-dependent peroxide reductase mitochondrial (PRDX3), transferrin receptor protein 1 (TFRC), tripeptidyl-peptidase 2 (TPP2), tumour necrosis factor (TNF), UDP-gluocose:glycoprotein glycosyltransferase 1 (UGGT1), X-ray repair cross-complementing protein 6 (XRCC6).

Table 5.22. Summary of proteins identified as significantly associated with RA disease outcome and activity measures.

Protein	Therapeutic drug levels	Amgevita levels	Benepali levels	TJC	SJC	VAS	CRP	DAS28	ΔDAS28	Poor EULAR response	ΔDAS28 <1.2	Expression baseline→3 months Poor Response	Expression baseline→6 months Poor Response	Expression 3→6 months, Poor Response	Baseline Good/moderate vs Poor Response	3 months Good/moderate vs Poor Response
C4B (P0C0L5)	X															
A1AG1 (P02763)		X					X									
IKKA (O15111)			X													
TNF (P01375)				X	X											X
MIF (P14174)				X								X				
EHD1 (Q9H4M9)				X		X		X	X							X
TCPH (Q99832)					X			X							X	
ILF3 (Q12906)					X					X						X
IGF1 (P05019)					X			X		X						X

Protein	Therapeutic drug levels	Amgevita levels	Benepali levels	TJC	SJC	VAS	CRP	DAS28	ΔDAS28	Poor EULAR response	ΔDAS28 <1.2	Expression baseline→3 months Poor Response	Expression baseline→6 months Poor Response	Expression 3→6 months, Poor Response	Baseline Good/moderate vs Poor Response	3 months Good/moderate vs Poor Response
CFHR3 (Q02985)					X		X	X					X			X
CRP (P02741)					X			X	X							X
ASPH (Q12797)						X		X			X					X
SAA1 (P0DJI8)							X									X
PRS6A (P17980)							X									
SAA2 (P0DJI9)							X									
APOA4 (P06727)							X									
RBP4 (P02753)							X									
CASP10 (Q92851)							X									
TPP2 (P29144)							X									

Protein	Therapeutic drug levels	Amgevita levels	Benepali levels	TJC	SJC	VAS	CRP	DAS28	ΔDAS28	Poor EULAR response	ΔDAS28 <1.2	Expression baseline→3 months Poor Response	Expression baseline→6 months Poor Response	Expression 3→6 months, Poor Response	Baseline Good/moderate vs Poor Response	3 months Good/moderate vs Poor Response
ELANE (P08246)							X									
MAP2K3 (P46734)							X	X					X	X		
CLTC (Q00610)							X									
SELENOP (P49908)							X									
LBP (P18428)							X									X
NAA25 (Q14CX7)							X									
CFHR5 (Q9BXR6)							X									
GARS1 (P41250)							X									
YWHAH (Q04917)								X								
TFRC (P02786)								X								

Protein	Therapeutic drug levels	Amgevita levels	Benepali levels	TJC	SJC	VAS	CRP	DAS28	ΔDAS28	Poor EULAR response	ΔDAS28 <1.2	Expression baseline→3 months Poor Response	Expression baseline→6 months Poor Response	Expression 3→6 months, Poor Response	Baseline Good/moderate vs Poor Response	3 months Good/moderate vs Poor Response
PKP3 (Q9Y446)										X	X				X	X
NAMPT (P43490)												X				X
OGN (P20774)												X				
PARK7 (Q99497)												X	X			
PRDX3 (P30048)												X			X	
UGGT1 (Q9NYU2)												X				
AKT1 (P31749)												X				
CALD1 (Q05682)												X	X		X	
CAMK1 (Q14012)												X			X	
LCN2 (P80188)												X				

Protein	Therapeutic drug levels	Amgevita levels	Benepali levels	TJC	SJC	VAS	CRP	DAS28	ΔDAS28	Poor EULAR response	ΔDAS28 <1.2	Expression baseline→3 months Poor Response	Expression baseline→6 months Poor Response	Expression 3→6 months, Poor Response	Baseline Good/moderate vs Poor Response	3 months Good/moderate vs Poor Response
FLII (Q13045)												X				
SLC4A1 (P02730)													X			
LRRFIP1 (Q32MZ4)													X	X	X	
KRT1 (P04264)													X		X	
BLVRB (P30043)													X			
LTBP1 (Q14766)															X	
PDIA6 (Q15084)															X	
XRCC6 (P12956)																X
CALM2 (P62158)																X
COL6A2 (P12110)																X

Protein	Therapeutic drug levels	Amgevita levels	Benepali levels	TJC	SJC	VAS	CRP	DAS28	ΔDAS28	Poor EULAR response	ΔDAS28 <1.2	Expression baseline→3 months Poor Response	Expression baseline→6 months Poor Response	Expression 3→6 months, Poor Response	Baseline Good/moderate vs Poor Response	3 months Good/moderate vs Poor Response
PPIA (P62937)																X
RSU1 (Q15404)																X

LEGEND: Protein expressed/differentially expressed at baseline, protein expressed/differentially expressed at baseline and three months, protein expressed/differentially expressed at baseline and six months, protein expressed/differentially expressed at three months, protein expressed/differentially expressed at three and six months, protein expressed/differentially expressed at six months, protein expressed/differentially expressed at all time points or significant in linear mixed effects model.

ABBREVIATIONS: 14-3-3 protein η (YWHAH), 26S proteasome regulatory subunit 6A (PSMC3, also known as PRS6A), α-1-acid glycoprotein 1 (ORM1, also known as A1AG1), ΔDAS28 (change in DAS28 between measurement time points), apolipoprotein A-IV (APOA4), aspartyl/asparaginyl β-hydroxylase (ASPH), band 3 anion transport protein (SLC4A1), calcium/calmodulin-dependent protein kinase type 1 (CAMK1), caldesmon (CALD1), calmodulin-2 (CALM2), caspase-10 (CASP10), clathrin heavy chain 1 (CLTC), complement C4-B (C4B), complement factor H-related protein (CFHR), C-reactive protein (CRP), Disease Activity Score of 28 Joints (DAS28), dual specificity mitogen-activated protein kinase kinase 3 (MAP2K3), EH domain-containing protein 1 (EHD1), European League Against Rheumatism (EULAR), flavin reductase (NADPH) (BLVRB), glycine—tRNA ligase (GARS), Health Assessment Questionnaire (HAQ), inhibitor of NF-κB kinase subunit α (CHUK, also known as IKKA), insulin-like growth factor I (IGF1), interleukin enhancer-binding factor 3 (ILF3), keratin, type II cytoskeletal 1 (KRT1), latent-transforming growth factor β-binding protein 1 (LTBP1), leucine-rich repeat flightless-interacting protein 1 (LRRFIP1), lipopolysaccharide-binding protein (LBP), macrophage migration inhibitory factor (MIF), mimecan (OGN), N-α-acetyltransferase 25, NatB auxiliary subunit (NAA25), neutrophil elastase (ELANE), neutrophil

gelatinase-associated lipocalin (LCN2), nicotinamide phosphoribosyltransferase (NAMPT), Parkinson disease protein 7 (PARK7), peptidyl-prolyl cis-trans isomerase A (PPIA), plakophilin-3 (PKP3), protein disulfide-isomerase A6 (PDIA6), protein flightless-1 homologue (FLII), RAC-α serine/threonine-protein kinase (AKT1), Ras suppressor protein 1 (RSU1), retinol-binding protein 4 (RBP4), selenoprotein P (SEPP1), serum amyloid A (SAA), swollen joint count (SJC), T-complex protein 1 subunit η (CCT7, also known as TCPH), tender joint count (TJC), thioredoxin-dependent peroxide reductase mitochondrial (PRDX3), transferrin receptor protein 1 (TFRC), tripeptidyl-peptidase 2 (TPP2), tumour necrosis factor (TNF), UDP-gluocose:glycoprotein glycosyltransferase 1 (UGGT1), visual analogue score of patient global health (VAS), X-ray repair cross-complementing protein 6 (XRCC6).

CHAPTER SIX: PROTEOMIC PREDICTORS OF TREATMENT RESPONSE TO ETANERCEPT IN PATIENTS WITH RHEUMATOID ARTHRITIS – DISCUSSION

Summary of chapter contents:

6.1. Acquisition of data using SWATH-MS and data QC

In the etanercept sub-cohort, three batches of patient samples were processed by the SBDC using SWATH-MS to acquire protein expression data. The first two batches were processed using the same methods, but immunodepletion of the most abundant proteins was carried out using a different method for the third batch. Subsequent PCA demonstrated separation by batch, and batch correction was carried out statistically using the sva[217] package in R. In the BRAGGSS-PD cohort, two batches of patient samples were processed using SWATH-MS using the same methods for both batches, but again, PCA showed clear separation by batch. Statistical batch correction was again performed using the sva package. One outlying sample from the baseline (pre-treatment) time point in the etanercept sub-cohort was also removed from analysis following hierarchical cluster dendrogram analysis.

Across both cohorts, a large proportion of proteins detected had missing values. Missing values are a known limitation of DIA proteomics techniques (such as SWATH-MS), and missingess is multifactorial e.g. biological/chemical factors (such as sample degradation during storage, mis-cleavage during digestion, ion suppression) or bioinformatics factors (e.g. peptide misidentification, ambiguous precursor matching during quantitation)[102]. Missing values can range between 10 – 90% in gel-based MS techniques[100], but as yet, no statistics for missing values have been compiled for SWATH-MS specifically. However,

Krasny *et al* demonstrated a 15 – 20% increase in peptides identified compared with DDA MS and were able to identify 54% more hepatic extracellular matrix proteins with SWATH-MS than with DDA MS[295]. It is known from previous studies that peptides which are less abundant are more challenging to detect using DIA techniques, so these peptides are more likely to be recorded as "missing"[296]. In this work, multi-density plots showed reduced protein expression with increasing missing values, which fits with prior knowledge regarding DIA proteomics, so it was hypothesised that values were likely to be missing because of lack of detection, as opposed to due to a patient's physiological state during sampling. Heatmaps comparing protein missingness with sampling time point and EULAR response status of patients showed no clear pattern causing proteins to be missing, so missing values were imputed using a random forest algorithm. Repeat density plots after imputation showed good agreement of imputed values with the original datasets.

The clear batch effect in SWATH-MS results is likely to have been due to processing factors at the SBDC, rather than sampling issues. Samples sent for processing at the SBDC were collected over a time period of several years from multiple different centres across the UK, but no batch effect was seen in terms of sampling time point. Furthermore, HCs processed as part of the second batch were seen grouped with other samples from this batch, and did not separate out on their own, indicating that the effect is due to processing at the SBDC, and not due to physiological differences between HCs and patients with RA. Sample processing and SWATH-MS techniques are continually evolving at SBDC, and the four batches sent for SWATH-MS were processed over a time period of three years, so it is unsurprising that there is separation in PCA between all batches. The first and second batches processed had different processing protocols from the third and fourth batches, and yet, all four batches still separated from one another. This could have been due to any number of variables, such as time to processing, use of a different spectrometer, differences in technique and differences in staff processing samples, but this can only be speculated as these samples were processed externally from the CfMR. However, because these differences were identified, and batch was not correlated with clinical response or sampling time point, statistical methods were able to be used to correct for batch effect.

Imputation of missing protein values was based on the hypothesis that protein values were missing because they were not detected during MS, and not because they were not present in a sample physiologically. Treating missing values as missing, instead of imputing them, could have generated different results, and this would have influenced findings of analysis

of the SWATH-MS data. However, as post-imputation multi-density plots remained almost unchanged compared to the original data, it is unlikely that positive findings would have significantly changed if this alternative approach had been used. Furthermore, heatmaps of missing and non-missing proteins compared to sampling time point and treatment response showed no patterns of missingness, so these proteins were likely to be missing at random due to non-detection at SWATH-MS, as opposed to due to a physiological cause.

SWATH-MS provides a robust method of acquisition of protein expression data. Because it is a DIA MS method, it enables increased and more reproducible proteome coverage with fewer missing values than DDA methods[295]. With fewer missing values relative to DDA MS, imputation accuracy is likely to be improved due to an increased number of values in the original dataset to contribute to imputation. Multiple imputation techniques were assessed on the same subset of data and directly compared; the most accurate method (random forest) was chosen based on the lowest RMSE. Following imputation with the most objectively accurate method, repeat multi-density plots showed that imputation did not alter the mean protein expression profile across each imputed dataset, indicating good accuracy of imputation. Statistical batch correction was carried out with a robust and validated methodology (surrogate variable analysis)[297], leading to confidence in the analysis of the combined data set.

Batch effects would have been negated if samples were all processed at the same time. However, to due limitations in budget and the time taken to collect BRAGGSS-PD samples, samples were sent for processing in multiple batches over the course of three years. If samples had been sent in one large batch towards the end of this project, while batch effect might have been eliminated, this would have reduced analysis time and preliminary work to determine the validity of the SWATH-MS data on earlier batches would not have been carried out in order to justify the processing of later batches. The decision to treat data as missing at random and not biologically missing may have led to different results and conclusions being drawn. However, due to relatively low levels of missingness leading to improved imputation accuracy and negligible alteration in mean protein expression distribution post-imputation, this is unlikely to be a major weakness of this analysis.

Future improvements could involve sending all samples to be analysed in one single large batch, if budget allows, as well as validation with a second proteomics quantitation technique (e.g. MRM) to determine whether values are missing at random or biologically missing.

Multiplexed methods of protein quantification (e.g. SomaScan®, Olink®) could also be considered, as these techniques provide excellent quantification accuracy, although proteins not included on proprietary panels would obviously not be detected, unlike using a bottom-up proteomics quantitation technique such as SWATH-MS.

In conclusion, outliers and batch effect have been corrected during SWATH-MS data QC and missing values have been imputed with an empirically accurate technique, leading to a comprehensive proteomics dataset for further analysis against RA disease outcomes.

6.2. Differential expression of proteins between RA cases and HCs

216 proteins were found to be significantly differentially expressed in pre-treatment RA patients with high disease activity, in comparison to HCs. Four of these proteins overlapped with those included on the commercially-available Sectra DA panel, which measures 12 protein biomarkers to determine RA disease activity[298]. The four overlapping proteins are:

- Chitinase-3-like protein 1 (CHI3L1, UniProt ID P36222), also known as YKL-40.
- SAA1 (UniProt ID P0DJI8) and SAA2 (UniProt ID P0DJI9) – it is unclear which isoform of SAA is included in the Vectra DA panel as the published manuscript states only "SAA".
- CRP (UniProt ID P02741).

While this overlap with some proteins already included in a commercially-available biomarker panel aimed at measuring RA disease activity is reassuring, it also opens up avenues for future research. Subsequent sections in this thesis discuss potential candidate biomarkers for future studies of treatment response, but these could also be developed into a novel, sophisticated panel to assess RA disease activity. This would require a clinical trial in an inception cohort of patients with RA with detailed clinical outcome measurements carried out with each blood sampling time point to determine whether the panel accurately reflects measurable changes in disease activity.

179/216 proteins were from the RA protein library, and the remaining 37 proteins were from the plasma protein library. This has important implications for future research because proteins previously found to be important in RA have been replicated here, and carrying out a literature review to build a bespoke RA protein library instead of using a generic plasma library alone has been a vital step in this process. Researchers should continue to build bespoke disease-specific protein libraries based on literature reviews in future studies,

despite the low power of some studies that have been included in the RA protein library in this thesis.

The majority of proteins (146/216) were found to be increased in RA patients, and the remaining 70 proteins had decreased expression. Because protein expression values had already been log2-transformed and normalised as part of data pre-processing, a Welch's t-test was used to compare expression of each protein between cases and controls. These findings imply that proteins that decrease in expression after successful treatment are potentially of most interest. For example, Tasaki *et al* used multi-omics to carry out high-dimensional phenotyping of patients with RA to identify molecular signatures of both treatment response and non-response[138]. Future work could include defining overall signatures of response and non-response within the proteomics data generated during the course of this thesis.

This case-control approach led to the creation of a more streamlined dataset, with fewer proteins included in analysis. This was so that analysis would focus on proteins that were more strongly associated with an active RA disease state, as well as the issue of the presence of many potential predictors in comparison to relatively few patients at any given time point. Dimensionality reduction of the dataset meant that there was an improved likelihood of determining meaningful associations between detected proteins and RA disease outcomes, as well as reducing the chances of false-positives due to multiple comparisons between many hundreds of proteins.

However, this approach may also have led to inadvertently losing any signal for potential relationships between proteins without differential expression between RA cases and HCs that may still be associated with RA disease outcomes. Future work could involve repeating this analysis with the full protein dataset and not the abridged dataset containing only significant results from the case-control analysis. Analyses could then be compared to determine whether any additional proteins are significantly associated with RA disease outcomes from the unabridged dataset.

6.3. Longitudinal analysis of protein expression in the first 12 weeks of treatment with Amgevita or Benepali

Longitudinal analysis of protein expression during the first 12 weeks of treatment was carried out on 16 patients with RA from the BRAGGSS-PD cohort, receiving a combination

of Amgevita (10 patients) and Benepali (six patients). All 16 patients were included in linear mixed effects modelling to determine any associations between protein expression and achievement of therapeutic drug levels. In the univariate analysis, five proteins were found to be significantly associated with therapeutic drug levels. These proteins were then placed in a multivariable linear mixed effects model, with age, biological sex, weight and concurrent csDMARD use included as fixed effects and patient ID and sampling time point included as random effects. In the multivariable model, only C4B (OR_{adj} 11.4, 95% CI 1.42 – 92.60) and age at baseline (OR_{adj} 1.32, 95% CI 1.03 – 1.70) were associated with increased odds of achieving therapeutic drug levels.

Protein expression was then analysed against individual therapeutic agent drug concentrations to determine whether any associations existed, again using a linear mixed effects modelling approach. Nine proteins were found to be significantly associated with Amgevita drug concentration levels, but after these were all included in a multivariable model, only A1AG1 (β-coefficient $_{adj}$ -0.76, 95% CI -1.31 – (-0.22)) and female sex (β-coefficient$_{adj}$ 3.93, 95% CI 1.38 – 6.47) remained significantly associated. Three proteins were significantly associated with Benepali drug concentration levels; after inclusion in a multivariable model, only IKKA (β-coefficient$_{adj}$ 0.88, 95% CI 0.31 – 1.45), female sex (β-coefficient -4.21, 95% CI -7.24 – (-1.18)) and weight (β-coefficient -0.18, 95% CI -0.26 – (-0.10)) remained significant. Protein expression might be different between the two drugs due to a number of reasons, such as drug structure (Amgevita is a humanised mAb, whereas Benepali is a dimeric fusion protein) or differences in study populations (e.g. 10 patients, of whom nine were female, in the Amgevita cohort, versus six patients, of whom four were female, in the Benepali cohort).

C4B is a C3 and C5 convertase that is an essential participant in the classical complement pathway[299]. It also binds to Igs and immune complexes (ICs), which augments their solubility and promotes clearance of ICs via erythrocyte complement receptor type 1 (CR1). This suggests a role for C4B in the mediation of immune-regulated conditions, such as RA. Rigby *et al* conducted a study of C4A and C4B gene copy numbers in a cohort of 160 RA patients, 88 non-RA patients and 51 HCs[300]. They found that RA patients had an approximately two-fold increase in the frequency of either homo- or heterozygous C4B deficiency compared to non-RA patients and HCs. In addition, C4B deficiency in their study cohort was found to be associated with the shared epitope in seropositive RA patients, so the authors suggested a role for C4B deficiency and interaction with the shared epitope in

seropositive RA pathogenesis. Conversely, Holers *et al* found reduced C4B deposition on mannan in a mouse model of collagen antibody-induced arthritis (CAIA); mice were injected with MASP2 duplexes prior to induction of CAIA, and these were associated with reduced clinical disease activity as well as decreased *ex vivo* C4B deposition on mannan[301]. However, this study was carried out in a mouse model and the Rigby study was carried out in humans, so the findings in this thesis would agree with the human genetic study in that increased levels of C4B were associated with the achievement of therapeutic drug levels. This suggests that patients without C4B deficiency (whether genetic or caused by increased RA disease activity) could be more likely to clear ICs associated with RA or even ADAbs, which could contribute to reaching therapeutic drug levels in these patients. It should be noted that the 95% CI of this association is very wide (1.42 – 92.60), so any conclusions regarding C4B from this study cannot be drawn with a high level of certainty. However, only 6.92% of protein values were missing for C4B in this study, so findings are unlikely to be affected by protein missingness or imputation errors.

A1AG1 is a hepatically-produced transport protein that binds multiple ligands, as well as synthetic drugs; drug-binding alters drug disposition and bioavailability[302]. Interestingly, neither csDMARDs nor bDMARDs are on the list of known medications that interact with A1AG1. However, a number of studies have found increased levels of A1AG1 in RA patients compared with HCs, in serum[303], plasma[304] and urine[294]. Rydén *et al* demonstrated increased fucosylation (a form of glycosylation constituting the addition of fucose sugar units to a molecule) of A1AG1 in patients with RA compared to HCs, but there was only a weak association with DAS28 at baseline and one year in male patients, and not in female patients, indicating that it did not represent an accurate indicator of disease activity[305]. The Rydén study included fewer patients than the current study, consisting of 130 patients with recently-diagnosed RA from six separate clinical rheumatology departments in Sweden, but it did include a higher number of HCs, consisting of 120 healthy blood donors (60 male and 60 female) from a single blood donation centre. Haston *et al* showed that A1AG1 detected in RA patients had increased fucosylation and sialylation (the covalent addition of sialic acid to a protein) compared to previously published levels found in health[306]. This altered A1AG1 in RA showed less efficient inhibition of callegenase-2 catalysis compared to normal plasma A1AG1, and it was hypothesised that A1AG1 found in RA synovial fluid may have insufficient function to prevent the enhanced cartilage destructive characteristic of active RA. Furthermore, A1AG1 has demonstrated varied glycosylation patterns, and this was shown to be different even between serum and synovial fluid within the same RA patient;

this structural diversity of A1AG1 showed that A1AG1 in serum was able to inhibit binding to the cell adhesion molecule E-selectin, but not A1AG1 in synovial fluid[307]. Using a different approach, Fischer *et al* used a rat model to elucidate the role of the DNA-sensing Toll-like receptor (TLR) 9 in inflammatory arthritis pathogenesis[308]. Rats with pristine-induced arthritis had TLR9 inhibited prior to disease induction, and alongside reduced inflammatory arthritis and almost complete lack of bone erosions, serum levels of A1AG1, IL-6 and RF were found to be decreased.

The current study found a negative association between A1AG1 protein expression and Amgevita levels. Although most previous studies have focused on differences in A1AG1 expression between RA patients and HCs, Rydén *et al* did find only a weak association between this protein and DAS28 in male patients only[305]. Previous studies seem to indicate that this protein may play a role in RA pathogenesis and active inflammation at the synovium, so increased levels could fit with reduced drug levels in the current cohort of patients, particularly given that A1AG1 is known to affect drug-binding and bioavailability, although this has not previously been documented for therapeutic agents such as csDMARDs or TNFi such as adalimumab or etanercept. Further mechanistic studies designed to assess A1AG1 in the setting of bDMARD therapeutic efficacy in RA patients would need to be carried out to assess this protein as a potential predictive biomarker. It should also be noted that this protein was not significantly associated with either achieving therapeutic drug levels in the combined BRAGGSS-PD cohort (both Amgevita and Benepali patients) or with Benepali drug concentrations. This could be due to a lack of signal in the Benepali patients given the low study numbers (n = 6), or it could be a genuine association with only Amgevita. This discrepancy could be due to the different structure and pharmacology of these two drugs (e.g. Amgevita is a humanised mAb with a longer $t_{1/2}$, Benepali is a dimerised fusion protein with a shorter $t_{1/2}$. This finding requires replication in a larger cohort, with consideration of measurement of glycosylation and sialylation of A1AG1 in the study design.

IKKA is a serine kinase that has been shown to regulate negative feedback of NF-κB[309 310]; NF-κB is known to influence the inflammatory state in RA, such as promoting T helper 1 responses, activation, proliferation and atypical apoptosis of RA fibroblast-like synoviocytes (FLS), and enhancing the bone resorption activity of osteoclasts[311]. A previous candidate gene study in BRAGGSS (using a cohort independent to those included in this thesis) showed an association between a SNP for *CHUK* (the gene that controls IKKA expression) and treatment response to both etanercept, and etanercept, adalimumab and infliximab in a

wider cohort of patients[312]. In a later candidate gene study of 755 RA patients, Ferreiro-Iglesias *et al* identified three genes associated with TNFi response, replicating the association with *CHUK*[313]. IKKA expression was associated with Benepali levels, and increased drug levels likely indicate an increased probability of treatment response. Increased IKKA could indicate reduced inflammation via the NF-κB pathway, but it is unclear whether patients had increased drug levels due to a predisposed reduced inflammatory burden, or whether IKKA levels were increased because increased drug levels were treating active RA more effectively.

Strengths of this study include sampling over multiple time points within the same patient; this provides detailed information on protein expression in each patient over the first 12 weeks of treatment, as opposed to many other studies, where only one-to-three sampling time points are used. This increased sampling provides a detailed reflection of protein expression over a short period of time and gives an increased chance of identifying biomarkers of treatment response very early in treatment with expensive therapeutic agents. If non-responder status can be predicted earlier in treatment, this could lead to escalation/switching of medication before more cost is spent on an ineffective treatment. Another advantage to this study is the use of linear mixed-effects modelling in analysis. Linear mixed-effects models are able to incorporate variables such as within-patient sampling, as well as confounding variables such as age and biological sex into analysis, enabling a reduction in false-positive associations by incorporation of relatedness into model structure[314]. Power is increased because corrections applied to each model incorporate its specific structure.

As with the popPK study, this study would have benefitted from additional patient recruitment to increase power (discussed in Section 4.2.1). Another disadvantage is that proteomics samples were processed in two separate batches, leading to batch processing effect; this has been discussed in Section 6.1.

Future work could include recruitment of additional patients so that power can be increased to improve confidence in results. Furthermore, significant proteins from this analysis require replication in an independent cohort to ensure the transferability of findings and that findings are not specific to this single cohort. Replication could be carried out with lower-throughput protein quantitation methods, such as targeted ELISA or bespoke multiplexed panels.

In conclusion, linear mixed-effects modelling of protein expression in patients commencing Amgevita or Benepali identified several proteins associated with either the achievement of therapeutic drug levels or with drug concentration levels themselves. These findings could provide the basis for future validation to determine whether these are viable biomarkers of early treatment response in patients commencing these costly drugs. Biomarkers that are measurable in blood would have an advantage over radiological determination of RA disease activity, as the latter is time-consuming and costly. It should also be determined whether these biomarkers confer additional benefit over simply measuring CRP or drug levels.

6.4. Proteins associated with RA disease outcomes following treatment with etanercept

Linear regression was used to determine associations between protein expression and the following continuous RA disease outcomes:

- Primary outcomes:
  - DAS28.
  - Change in DAS28 between measurement time points.
- Secondary outcomes: DAS28 sub-components i.e. TJC, SJC, VAS of patient global health and high-sensitivity CRP measured using ELISA.

Over various time points, 10 proteins were associated with DAS28 and two proteins were associated with ΔDAS28, three proteins were associated with TJC, six proteins were associated with SJC, two proteins were associated with VAS and 17 proteins were associated with CRP. This is summarised in Figure 5.19 and Table 5.22. Three proteins overlapped with the Vectra DA panel[298]: SAA1 (UniProt ID P0DJI8), SAA2 (UniProt ID P0DJI9) and CRP (P02741), providing a degree of external agreement.

Logistic regression was used to determine associations between protein expression and the following categorical RA disease outcomes:

- Poor EULAR response.
- Failure to achieve an MCID in DAS28.

Over various time points, three proteins were associated with poor EULAR response and two proteins were associated with failure to achieve MCID in DAS28.

EHD1 (UniProt ID Q9H4M9) is an ATP- and membrane-binding protein involved in membrane reorganisation during ATP hydrolysis and vesiculation of endocytic

membranes[315]; this protein was positively associated with TJC in the analysis, but it is unclear whether this is a true biological association until more mechanistic studies can be carried out. If this were the case, one method to improve power would be to ensure the same person or small group of people were measuring DAS28 and calibrating their techniques, to account for variance in inter-observer scorings. EHD1 was also associated with VAS of patient global health, DAS28 and ΔDAS28, and it would be interesting to determine in the future whether this is a spurious association or whether TJC and VAS could be confirmed as surrogate measures of RA disease activity, in contrast to the findings of Hensor *et al*[52].

TCPH (UniProt ID Q99832), a component of the molecular chaperone complex chaperonin-containing T-complex that facilitates protein folding during ATP hydrolysis[316], was positively associated with SJC and DAS28, although the association with DAS28 was likely driven by its association with SJC, indicating active joint inflammation. CFHR3 (UniProt ID Q02985) is a protein thought to be involved in complement regulation that is associated with atypical haemolytic uraemic syndrome[317]; increased levels could be associated with raised SJC, but little is known about this protein's function from published literature[318]. Interestingly, CFHR3 was also associated with CRP measured using ELISA and DAS28, which could correspond with a pro-inflammatory function. Reassuringly, CRP (UniProt ID P02741) was found to have a positive association with SJC, DAS28 and ΔDAS28 (the latter two likely mediated via raised SJC). This is an expected association, as CRP is an indicator of active inflammation, as is a raised SJC, so this finding could be considered a positive control that true positive associations exist within this dataset.

IGF1 (UniProt ID P05019) is a growth factor similar and related to insulin, but with accentuated growth-promoting activity; it is thought to possibly regulate glucose transport and glycogen synthesis in osteoblasts[319]. In RA, osteoblast maturation is inhibited by circulating pro-inflammatory cytokines[320], so decreased osteoblast maturation due to increased disease activity (with concurrent raised SJC) could potentially fit with decreased IGF1. IGF1 was also negatively associated with DAS28, although this could be driven by its association with SJC. Concurrently, IGF1 was also negatively associated with poor EULAR response, which fits with its negative association with DAS28. ILF3 (UniProt ID Q12906) is a RNA-binding protein that contributes to the innate antiviral response during acute infection[321]; expression was negatively associated with SJC and poor EULAR response and could indicate an anti-inflammatory action, possibly induced by treatment with etanercept.

ASPH (UniProt ID Q12797) is a protein with two known isoforms: isoform 1 is involved in hydroxylation of Asp/Asn residues in specific epidermal growth factor-like domains[322] and isoform 8 is a membrane-bound calcium ion-sensing protein that is part of the endoplasmic reticulum[323]. According to the STRING database (text mining evidence only), this protein interacts with AKT1, a central node in the network of significant proteins in Figure 5.20. It was negatively associated with VAS of patient global health and DAS28 in analysis, and it was also negatively associated with failure to achieve an MCID. This may represent an anti-inflammatory role for ASPH.

TNF (UniProt ID P01375) was associated with TJC and SJC, which would be unsurprising given that it is the therapeutic target of etanercept (a TNFi drug), apart from the direction of association: increased TNF expression was associated with reduced TJC and SJC, which would not correlate with our current understanding of inflammation and perpetuated disease activity in patients with RA. Given that TJC can potentially be a subjective measure of disease activity due to factors such as superimposed OA or fibromyalgia, then this may indicate a false-positive signal, despite adjustment for potential confounders and multiple testing, as well as correction in a multivariable model. SJC as palpated by a clinician and not assessed as objective synovitis using imaging could also potentially be subjective, for example, if a patient has synovial hypertrophy due to persistent uncontrolled disease, or due to adiposity around joints. Again, this could indicate that the negative association with TNF is a false-positive. However, MIF (UniProt ID P14174), a pro-inflammatory cytokine[324], was negatively associated with TJC, and this would correspond with reduced symptoms and an anti-inflammatory effect from commencement of treatment with etanercept.

The function of A1AG1 has already been discussed in Section 6.3; it is positively associated with CRP measured using ELISA. SAA1 (UniProt ID P0DJI8) and SAA2 (UniProt ID P0DJI9) were both also associated with CRP, which is unsurprising as they are both known to be major acute phase reactants[325 326]. PRS6A (UniProt ID P17980) removes damaged or defunct proteins, participating in protein homeostasis[327]; expression was positively associated with CRP, and this could reflect ongoing cell protein damage caused by the inflammation during an active RA disease state. GARS1 is a protein involved in glycine ligation[328] and diadenosine tetraphosphate (Ap4A) production; Ap4A is a pleiotropic cell signalling molecule required in cell regulation pathways[329]. GARS1 expression was negatively associated with CRP, and it was shown to interact with PRS6A in Figure 5.24,

but not directly with CRP, so these two proteins could reflect ongoing RA disease activity, but from a pathway alternative to CRP.

APOA4 (UniProt ID P06727) is a protein involved in lipid physiology[330]; this protein was found to be positively associated with CRP, which could indicate an element of lipid dysregulation during acute inflammation. Previous literature has described the phenomenon of rheumatoid cachexia, a process of loss of fat-free mass and preservation or even increase of fat mass associated with uncontrolled RA[331], and proteins associated with lipid homeostasis could represent a novel source of disease activity biomarkers if associations can be replicated in independent cohorts.

Neutrophil elastase (ELANE, UniProt ID P08246) is a proteinase secreted by neutrophils during inflammation that influences the function of monocytes, granulocytes and natural killer (NK) cells and inhibits enzyme release and chemotaxis of neutrophils that is mediated via C5a[332]. ELANE expression was significantly positively associated with CRP, and this could be a feasible association, given that there is a body of evidence that neutrophils are involved in active RA[333]. ELANE could represent a novel biomarker of RA disease activity, particularly as it lies in an interaction network with CRP and TNF (Figure 5.20). MAP2K3 is a dual-specificity kinase that is actuated via cytokines and environmental stress, and could also reflect active RA disease[334]. This protein was positively associated with both CRP and DAS28, and likely increased levels reflect ongoing systemic inflammation.

CLTC (UniProt ID Q00610) is a constituent of clathrin, the predominant structural protein in the coat of coated pits and vesicles[335]; clathrin-coated pits are areas of the cell membrane where receptor-mediated endocytosis can occur[336]. CLTC expression was significantly associated with CRP, and this could represent increased endocytosis and cell-to-cell protein transport. LBP (UniProt ID P18428) is a protein that plays a role in innate immune response, initially by binding to lipopolysaccharide (also known as bacterial endotoxin)[337]; expression was positively associated with CRP. These proteins associated with cell function and immunity could also feasibly be potential biomarkers of RA disease activity and treatment response i.e. reduced levels could indicate treatment response. CFHR5 (UniProt ID Q9BXR6) is involved in complement regulation[338], and may play a role in RA in a similar way to C4A (discussed in Section 6.3). Expression was also positively associated with CRP.

RBP4 (UniProt ID P02753) participates in retinol (vitamin A) binding in plasma and was negatively associated with CRP[339]. SEPP1 (UniProt ID P49908) is thought to possibly be responsible for some of the antioxidant properties of selenium and was also negatively associated with CRP[340]; reduced levels with increasing CRP would fit with an antioxidant role. However, CASP10 (UniProt ID Q92851) was also negatively associated with CRP; this protein is part of the caspase cascade that contributes to apoptosis[341] and downregulation would not necessarily fit with systemic inflammation, so this may be a false positive finding. Similarly, TPP2 (UniProt ID P29144), which participates in a proteolytic cascade downstream of the 26S proteasome[342], and NAA25 (UniProt ID Q14CX7), which catalyses acetylation of certain peptide residues[343], are also negatively associated with CRP when perhaps this would be unexpected. However, both TPP2 and NAA25 do not have any interactions with any other proteins in the network map in Figure 5.20, so these could be false-positive associations.

YWHAH (UniProt ID Q04917) is an adapter protein that is thought to modulate a wide variety of signalling pathways[344] and is at the centre of a network node in Figure 5.20 including IGF1, EHD1 and CLTC. Given its diverse functions, it could feasibly interact with other proteins involved in the systemic inflammation of active RA. It was found to be negatively associated with DAS28. The final protein completing this network node is TFRC (UniProt ID P02786), which participates in iron homeostasis[345], and this was also found to be negatively associated with DAS28. The network node described here could represent a future area of research into biomarkers predictive of treatment success, or even potential therapeutic targets.

PKP3 is a protein of undetermined function that possibly plays a role in junctional plaques[346]. This protein did not have any interactions with any other proteins in the network map in Figure 5.20. However, it was also associated with a better-than-poor EULAR response and with achieving a ΔDAS28 >1.2, so future work should ascertain whether this protein could be a completely novel biomarker for RA disease activity or treatment response.

One of this study's strengths is the detailed interrogation of multiple RA disease outcome measures in order to determine whether common proteins were associated with any of these measures. This was possible because of detailed patient phenotyping available from the BRAGGSS cohort. Furthermore, a wide range of proteins were available for analysis, due to the dynamic range and sensitivity of SWATH-MS. Of note, as well as using a generic

plasma library, the benefit of SWATH-MS was that a bespoke library of proteins associated with RA was able to be used to validate previous findings (Tables 5.3 – 5.6). SWATH-MS has this advantage over proprietary multiplexed panels, as bespoke libraries can be generated for DIA, and for re-interrogation of previously acquired MS data *in silico*. Network analysis demonstrated a biologically viable network of protein interactions derived from significant proteins from analysis, and only seven out of 52 proteins had no interactions with any other proteins. This could be because they are false-positive findings, or encouragingly, this could be because they represent completely novel biomarkers in the RA inflammatory process; validation in an independent cohort followed by mechanistic studies if findings are replicated would be required to state this with any certainty, however. Novel proteins could potentially aid future prognostication of RA, particularly in terms of treatment response, or even provide new drug targets.

This arm of the study did make multiple comparisons within each patient, which could lead to false-positive results due to type 1 error. All analyses were adjusted for multiple testing using the Benjamini-Hochberg method, and were also adjusted for the potential confounders of age, biological sex and RA disease duration. However, these statistical methods may still not have mitigated for the multiple-testing effect. Furthermore, because missing protein values were imputed, this may have unknowingly influenced results. An area of future research could be to repeat analysis without imputation of missing proteins in order to assess whether results replicate; this would help to determine whether imputation affects the final results of this analysis. While multiple RA clinical outcome measures were assessed in this study, 2C-DAS28 was not used as an outcome measure, nor were EULAR response nor ΔDAS28 calculated using 2C-DAS28. These are potentially more objective outcome measures as they exclude the patient-reported components of TJC and VAS of patient-reported global health. However, clinician-reported measures have been shown to demonstrate more variability than patient-reported measures[347], so conventional DAS28 remains a valid RA clinical outcome measure. Future work could involve additional analysis using 2C-DAS28 (and EULAR response and ΔDAS28 calculated using 2C-DAS28). Ideally, protein expression would be analysed using objective measures of synovitis, such as ultrasound or MRI images, but this would require a new inception cohort and would be costly and time-consuming. Re-analysing the data with re-calculated 2C-DAS28 measurements would be the next logical step in a large dataset that has already been collected.

Therefore, future work should focus on validation of all the aforementioned significant protein associations in an independent cohort. If findings are replicated, targeted mechanistic studies could confirm whether these are biologically viable targets and whether they should be brought forward for development as large-scale biomarkers to be measured in clinical practice. Cross-referencing with known therapeutic agents could also provide avenues for repurposing of currently licensed drugs. Particularly promising protein interactions could even provide novel therapeutic targets, in the most optimistic scenario.

In conclusion, a number of both established and novel proteins have been found to be associated with various RA disease outcome measures, both clinician- and patient-reported. The established associations are reassuring and act as positive controls in this analysis. The new associations could lead to future biomarker development.

6.5. Differential expression of proteins over time and between different EULAR response groups

Differential expression of proteins was carried out within the same category of EULAR responders across different time points, and between poor and good/moderate EULAR responders within the same time point. 31 patients had a poor response at three months, and 11 proteins were differentially expressed between baseline and this time point. 15 patients had a poor response at six months; eight proteins were differentially expressed between baseline and this time point and three proteins were differentially expressed between three months and this time point. 152 patients had baseline protein data: nine proteins were differentially expressed between good/moderate and poor responders at three months, but no baseline proteins were differentially expressed between the same response categories at six months. 159 patients had three-month protein data and six-month EULAR response data, and 16 proteins were differentially expressed between good/moderate and poor responders.

A number of proteins were differentially expressed between baseline and three months in poor EULAR responders. Mimecan (OGN) stimulates bone formation alongside transforming growth factor (TGF)-β1 and/or TGF-β2, although the evidence level for this is only within the UniProt database without any peer-reviewed publications[348]. UDP-gluocose:glycoprotein glycosyltransferase 1 (UGGT1) oversees and regulates protein folding in the endoplasmic reticulum[349]. RAC-α serine/threonine-protein kinase (AKT1), alongside AKT2 and 3, is involved in the control of multiple cell processes, including cell metabolism and survival[350]. Neutrophil gelatinase-associated lipocalin (LCN2) is an iron

transport protein that is involved in a number of cellular processes, including apoptosis and innate immunity[351 352]. Protein flightless-1 homologue (FLII) is believed to play a role in cytokinesis cell migration[353]. These proteins, by-and-large, appear to be involved in multiple metabolic processes, including cell death, and it would make sense that these proteins would be differentially expressed at different time points in patients who are not responding to treatment.

Nicotinamide phosphoribosyltransferase (NAMPT) is an enzyme that catalyses the formation of an intermediate in the biosynthesis of nicotinamide adenine dinucleotide (NAD), a critical enzyme involved in a multitude of metabolic processes[354]. This protein was differentially expressed between baseline and three months in poor responders, but also between good/moderate and poor responders at three months. Parkinson disease protein 7 (PARK7) has multiple functions that are much-conjectured, but it appears to protect cells against oxidative stress and apoptosis[355-358]. As well as being differentially expressed between baseline and three months in non-responders, it is also differentially expressed between baseline and six months in non-responders. Thioredoxin-dependent peroxide reductase, mitochondrial (PRDX3) has a role in protecting cells against oxidative stress via the reduction and detoxification of peroxides[359]. This protein was differentially expressed between baseline and three months in poor responders, and between good/moderate and poor responders at baseline. Caldesmon (CALD1) is an actin- and myosin-binding protein that interacts with multiple molecular components of muscle tissue[360]. In this analysis, it was differentially expressed between baseline and three and six months in poor responders, as well as between good/moderate and poor responders at baseline. Calcium/calmodulin-dependent protein kinase type 1 (CAMK1) functions within a calcium signalling cascade, and amongst several functions, regulates the cell cycle, cell differentiation and actin filament organisation (i.e. it interacts with muscle components, like CALD1)[361]. In addition to differential expression between baseline and three months in poor responders, it also demonstrates differential expression between good/moderate and poor responders at baseline. MIF was also differentially expressed between baseline and three months in poor responders; its function is described in Section 6.4.

Other proteins were differentially expressed between baseline and six months in poor responders. Flavin reductase (NADPH) (BLVRB) is a broad-specificity oxidoreductase that is involved in multiple physiological processes, including haem catabolism[362]. Band 3 anion transport protein (SLC4A1) functions as both a transporter and structural protein in

erythrocytes[363]. It was differentially expressed between baseline and six months and between three and six months in poor responders. Leucine-rich repeat flightless-interacting protein 1 (LRRFIP1) is a transcriptional repressor; amongst other proteins, it is thought to modulate expression of TNF[364]. It is also thought to participate in TLR signalling via negative competition with the negative FLII regulator for myeloid differentiation primary response protein MyD88 (MYD88) binding[365]. It was differentially expressed between baseline and six months and between three and six months in poor responders and between good/moderate and poor responders at baseline. Keratin, type II cytoskeletal 1 (KRT1) is thought to regulate kinase activity, including the molecular scaffolder, the receptor of activated protein C kinase 1 (RACK1)[366]. KRT1 was differentially expressed between baseline and six months in poor responders, as well as between good/moderate and poor responders at baseline. CFHR3 was also differentially expressed between baseline and 6 months and between good/moderate and poor responders at 3 months; its function is described in Section 6.4. MAP2K3 was differentially expressed between baseline and six months and between three and six months in poor responders and its function is also described in Section 6.4. These proteins' common functionality appears to be in metabolic processes and as structural proteins. This could be related to increased immune cell proliferation associated with the active inflammatory process of uncontrolled RA.

Further proteins were differentially expressed between good/moderate and poor responders at baseline. Latent-transforming growth factor β-binding protein 1 (LTBP1) is a crucial modulator of TGFβ[367], which regulates growth and differentiation of a range of cell types and is involved in a number of processes, including immune function[368]. LTBP1 was differentially expressed between good/moderate and poor responders at baseline. Protein disulfide-isomerase A6 (PDIA6) is thought to prevent the aggregation of misfolded proteins[369]; this protein was also differentially expressed between good/moderate and poor responders at baseline. CCT7 was differentially expressed between good/moderate and poor responders at baseline and its function is described in Section 6.4. PKP3 was also differentially expressed between good/moderate and poor responders at baseline and three months; its function was also described in Section 6.4. Again, common themes of these proteins include cell regulation and immune function.

Finally, a number of proteins were differentially expressed between good/moderate and poor responders at three months. X-ray repair cross-complementing protein 6 (XRCC6) is a helicase that plays a variety of important roles in DNA replication and repair[370].

Interestingly, patients with SLE have been found to produce large amounts of autoantibodies to both XRCC5 and XRCC6[371]. As XRCC6 was found to be differentially expressed between good/moderate and poor responders and also has an association with a related rheumatic disease, this protein could provide an interesting potential biomarker for future study. Calmodulin-2 (CALM2) controls an array of enzymes, ion channels and other proteins via calcium binding[372], and it was found to be differentially expressed between good/moderate and poor responders at three months. Collagen α-2(VI) chain (COL6A2) is a cell binding protein[373], and was also differentially expressed between good/moderate and poor responders at three months. Peptidyl-prolyl cis-trans isomerase A (PPIA) is a multi-function protein that has wide-ranging functions, including chemotactic effects on leukocytes[374] and both pro- and anti-apoptotic signalling in situations of oxidative stress[375]. PPIA was differentially expressed between good/moderate and poor responders at three months. Ras suppressor protein 1 (RSU1) is thought to participate in the Ras signal transduction pathway[376]; Ras proteins are involved in regulation of normal cell growth and control over malignant transformation[377]. RSU1 was differentially expressed between good/moderate and poor responders at three months, and this could indicate dysregulated cell proliferation in the setting of active RA.

TNF, EHD1, ILF3, IGF1, CRP, ASPH, SAA1 and LBP were also differentially expressed between good/moderate responders and poor responders at three months; their functions have already been described in Section 6.4. The proteins described here are compatible with a pro-inflammatory state, and differential expression between responders is likely to represent treatment response to etanercept. Replication of associations with linear RA clinical outcome measures also indicates that these proteins may be more robust candidate biomarkers for future replication and functional studies. Overall, it appears that change in proteins with treatment may be more informative than measurement of proteins at a single time point (e.g. pre-treatment), and this requires validation in future studies. Perhaps focusing on change in protein expression may be informative in the prediction of future or sustained drug response. In the clinic room, this could lead to measurement of protein profiles at baseline and a time point early in treatment (e.g. 6 weeks) to determine whether to discontinue or persist with treatment.

The strength of this study is that it allows protein expression profiles to be generated for the groups of interest, as opposed to identifying single proteins of interest. These profiles can give an overview of the inflammasome involved in active RA, which could then identify

novel biomarkers and/or drug targets. Furthermore, by testing multiple proteins in future replication studies, positive replication may be more reliable if multiple replicated associations are found. The finding that CRP was associated with treatment response at three months acts as a positive control as this is an expected result, and adds more validity to the findings of this study. Other differentially expressed proteins are associated with prior knowledge of active inflammation and autoimmunity as well, such as TNF and SAA1. Proteins identified from these analyses are, therefore, promising, and warrant further study in the future that is outwith the scope of this thesis.

This study would have been strengthened by increased patient numbers, particularly if there had been more SWATH-MS data available at the six-month time point. SWATH-MS protein maps were also based on pre-defined RA and plasma libraries – this means that other influential proteins may been missed from analysis due to the pre-selection process. However, a strength of data obtained using SWATH-MS is that future extractions with alternate protein libraries can be carried out *in silico*, if desired for additional analysis. Network analysis could also reveal additional proteins to extract for testing in validation studies.

In addition, samples included in this analysis were the same samples as those from Section 5.7, leading to inevitable data leakage. Because the EULAR response criteria are based on DAS28 measurements, proteins previously associated with DAS28 and its components are also influencing the differential expression profiles in this analysis. This means that there must be a caveat on findings in this analysis, as they are not a replication of findings from Section 5.7, but rather, an extension of analysis of DAS28 and its components that is being amalgamated into overall EULAR response. For example, only three proteins were associated with poor EULAR response, but these all overlapped with differential expression profiles between good/moderate and poor responders at various time points. However, this work is still valuable, because as previously mentioned, it provides a higher-level summary of the ongoing RA active inflammasome at different sampling time points.

Future work would involve replication studies in an independent cohort of patients, as well as functional laboratory studies to assess the biological feasibility of any proteins whose findings replicate. In conclusion, this analysis provides multiple differential expression profiles that define how the protein expression of poor responders changes over time, and how good/moderate responders differ from poor responders at different time points. The

proteins that make up these profiles could contribute to treatment response prediction in the future, or provide targets for novel drug development or drug repurposing.

6.6. Machine learning methods to detect proteomic predictors of treatment response

Four different machine learning methods were used to train models predictive of treatment non-response, defined as either a poor EULAR response or failure to achieve an MCID in DAS28 at a given time point. A support vector machine provided the model with the lowest MMCE after benchmarking, then optimum hyperparameters were tuned before testing on an independent validation dataset. Predictive performance was poor, with poor calibration, using both treatment failure outcomes and at all time points tested.

At the model training and benchmarking stage, MMCE values were actually relatively low, ranging from 0.0137 for a support vector machine algorithm, predicting poor EULAR response at three months, to 0.2879 for a K-nearest neighbours algorithm predicting failure to achieve an MCID in DAS28 at six months; this latter MMCE is still relatively good performance. These very low MMCE values may indicate over-fitting of the data; it was envisaged that this would be mitigated by performing nested resampling during the training and benchmarking phase. MMCE values could potentially have been improved if a more robust resampling technique such as bootstrapping had been used, but nested resampling was chosen because it is not as computationally intensive as bootstrapping, in order to reduce computation time during the model development stage. Computational time was prolonged because nested resampling was carried out and merged with the benchmarking phase as part of the implementation in the mlr[226] R package.

Pre-processing data could potentially cause data leakage affecting the model training. Missing protein values were imputed statistically, based on pre-existing values already in the dataset, but if imputed proteins were completely missing from a sample due to biological reasons and incorrectly assumed to be present and then imputed, this could also have affected results. Similarly, batch correction was performed statistically, and this may have caused the effect of some proteins that would have influenced prediction to be mitigated, leading to a less predictive model. Finally, due to class imbalance, datasets had to be synthetically up-sampled using a SMOTE algorithm, which again could have led to data bleeding, as up-sampled values were based on pre-existing values in the dataset.

However, findings during the prediction phase of this analysis, carried out on data partitioned independently from the training data, demonstrated very poor predictive accuracy. Pragmatically, this would be more in-keeping with what would have been expected pre-analysis: hundreds of variables were being used to develop predictive models with a relatively low sample of patients. Whilst there has been some success in developing machine learning models predictive of treatment response in RA[378], many of these published studies have published AUCs from model training and not from validating these trained models in an independent dataset. Therefore, many of these studies report results with an optimistic bias that are not truly reflective of a chosen model's predictive accuracy. While the use of a support vector machine algorithm might mitigate against having multiple comparisons with few samples, algorithm selection is irrelevant if there truly is not any predictive pattern in the data that is being analysed.

Another issue with this lack of predictive accuracy could be because of the treatment response measures selected. Both outcome measures were based on the conventional four-component DAS28, and not on more objective RA disease activity outcomes, such as radiological evidence of synovitis (e.g. ultrasound or MRI scores of synovitis), or even therapeutic drug levels, such as in Section 5.6. These outcomes are more objective, because they do not rely on subjective measures with high variability, such as patient-reported VAS of global health, a TJC, or even variability between assessors of TJC and SJC sub-components of the DAS28. Perhaps future predictive models will need to use more objective outcome measures in order to predict treatment response with more accuracy. This would require large-scale inception studies that integrate detailed clinical and phenotypic data with one such objective measure of treatment response. Performance could also be improved with further optimisation of the bespoke RA protein library, as informative proteins may not have been included in the final analysis. A strategy for expanding the bespoke protein library could be to include proteins or genes known to be associated with ADME of TNFi.

Strengths of this study include the benchmarking of four different machine learning algorithms that implement training using mathematically contrasting techniques, providing a broad overview of potential suitable algorithms. Furthermore, while nested resampling is not as robust a resampling method as for example, bootstrapping, it still mitigates against over-fitting and hyperparameter selection based on unequal train/test splits during resampling carried out in model training. There can also be confidence in the results of final model testing in an independent validation dataset; while these results are essentially

negative findings, data was correctly partitioned and not re-used during both training and testing, so predictive accuracy results are likely to be true values.

This analysis relied heavily on several data transformations during pre-processing, which could have caused optimistic bias during model training. However, these transformations were also carried out on the test data (but carried out separately, so values were not imputed from samples from the training dataset), and predictive accuracy in this dataset was poor, so data manipulation prior to analysis may not have had as major an effect as anticipated. Perhaps the most important causes of poor predictive accuracy could be that a more objective treatment outcome measure was not selected or that the RA and plasma protein libraries did not include proteins that are most predictive. However, it was important to demonstrate that predictive accuracy is poor with conventional DAS28-based outcome measures in order to build a case for the adoption of more objective treatment outcome measures in both research and the clinic room.

The scope for future work is compelling: larger datasets with more detailed and objective measures of both treatment outcomes and predictors could lead to more accurate prediction models. Another piece of work could be to analyse batches sent for processing at SBDC separately to test whether correction of batch effect had a significant effect on the predictive accuracy of models. All data from SBDC was pooled and batch-corrected in this analysis in an attempt to maximise the power of the data, but keeping batches independent may change the outcome of results.

In conclusion, machine learning methods demonstrated poor predictive accuracy of SWATH-MS proteomics data in determining predictors of treatment non-response to etanercept in a cohort of densely-phenotyped patients with RA. However, there is a wide latitude for future work to be developed in this field, and this may include choosing a more objective RA disease outcome measure.

6.7. Chapter summary

SWATH-MS proteomics data was obtained on 180 patients starting etanercept from the wider BRAGGSS cohort, 16 patients starting either Amgevita or Benepali from the BRAGGSS-PD sub-study cohort and on 14 HCs. Robust protein QC was carried out prior to analysis. 216 proteins were found to be differentially expressed between RA cases and HCs. Several different proteins were found to be associated with:

- Therapeutic drug concentration levels of Amgevita or Benepali.
- Actual drug concentrations of Amgevita or Benepali.
- DAS28, ΔDAS28 and DAS28 sub-components.
- Poor EULAR response and failure to achieve an MCID in DAS28.

Furthermore, several proteins were differentially expressed over time by patients who had a poor EULAR response, and between good/moderate and poor EULAR responders at a given time point. These proteins provide a hypothesis for future replication and mechanistic studies to identify biomarkers of treatment response, or provide new drug targets, either in *de novo* drug development or in repurposing of therapeutic agents that have already been licensed.

CHAPTER SEVEN: A PROTEIN QUANTITATIVE TRAIT ANALYSIS OF PATIENTS COMMENCING ETANERCEPT

Summary of chapter contents:

7.1. Results

7.1.1. Study participants

For the pQTL study, 147 patients with available genotyping data were included from the etanercept sub-cohort of BRAGGSS patients, as previously described in Section 5.1. A summary of their baseline covariate characteristics is described in Table 7.1.

Table 7.1. Baseline covariate characteristics of patients in the BRAGGSS etanercept cohort.

Covariate	**Statistic**
Age at baseline (years), median [IQR]	56.39 [49.34, 64.73]
Disease duration (years), median[IQR]	6.00 [2.00, 13.00]
Female sex, n (%)	108 (75.52)
Concurrent csDMARD, n (%)	116 (81.12)
BMI (kg/m^2), median [IQR]	27.43 [23.48, 32.67]
Seropositivity for RF/ACPA, n (%)	96 (67.13)

ABBREVIATIONS: Anti-citrullinated antibody pepides (ACPA), Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate (BRAGGSS), body mass index (BMI), conventional synthetic disease-modifying anti-rheumatic drug (csDMARD), interquartile range (IQR), rheumatoid factor (RF).

For the polygenic risk score (PRS) study, an additional 1,563 patients from BRAGGSS independent to the above etanercept cohort of patients was included. These patients were receiving a number of different TNFi drugs, including adalimumab, etanercept and infliximab. Recruitment and genotyping of this cohort have previously been described[248]. A summary of baseline covariate characteristics is described in Table 7.2; age information was not available in this cohort.

Table 7.2. Baseline covariate characteristics of patients in the polygenic risk score target data cohort.

Covariate	**Statistic**	**Number of missing values**
Female sex, n (%)	1,202 (76.95)	1
Concurrent csDMARD, n (%)	1,273 (81.50)	1
Baseline DAS28, median [IQR]	6.35 [5.72, 7.12]	0
EULAR poor response, n (%)	279 (17.76%)	0

ABBREVIATIONS: Conventional synthetic disease-modifying anti-rheumatic drug (csDMARD), Disease Activity Score of 28 Joints (DAS28), interquartile range (IQR).

7.1.2. Results of pQTL analysis

A total of 482 unique proteins that were present in all three batches processed at the SBDC (Section 5.3) after passing QC as described in Section 5.4 were included in the analysis.

At baseline (pre-treatment), a total of 2,184 *cis* pQTLs were identified for 60 unique proteins, likely reflecting that many pQTLs will be in strong LD with one another. The most associated SNPs for each protein at this time point are presented in Table 7.3. Boxes 7.1 and 7.2 detail tissues where eQTLs for SNPs rs2894255 and rs10737680 are expressed, respectively. A q-q plot for all *cis* pQTLS identified is presented in Figure 7.1. The distribution of observed p-values for *cis* pQTLs departed the diagonal sooner than those for *trans* pQTLs, indicating that the former were easier to detect than the latter, and this reflects the difference in p-value thresholds for *cis* and *trans* pQTLs. A total of 389 *trans* pQTLs were also identified for two unique proteins; the most associated SNPs for each protein at this time point are presented in Appendix Ten, Table 1.

After three months of treatment with etanercept, a total of 1,432 *cis* pQTLs were identified for 68 unique proteins, again, likely reflecting that many of these pQTLs are in strong LD with one another. The most associated SNPs for each protein at this time point are presented in Table 7.4. Boxes 7.1 and 7.3 detail tissues where eQTLs for SNPs rs2894255 and rs190820372 are expressed, respectively. A q-q plot for all *cis* pQTLs identified is presented in Figure 7.2; again, this demonstrated that *cis* pQTLs were easier to detect than *trans* pQTLs. A total of 1,570 *trans* pQTLs were also identified for eight unique proteins; the most associated SNPs for each protein at this time point are presented in Appendix Ten, Table 2.

Table 7.3. SNPs with lowest false-discovery rate-adjusted p-values for each protein included in pQTL study at baseline (before treatment with etanercept).

Most associated SNP	**Chr**	**Protein (UniProt ID)**	**% protein values missing before imputation**	**Total SNPs for protein**	**p-value**	**Adjusted p-value**	**Tissue of expression of corresponding eQTL**
rs189758989	22	APOL1 (O14791)	0.26	1	2.64E-07	0.0005	None
rs145980995	10	IKKA (O15111)	48.46	4	7.39E-08	0.0002	None
rs62269178	3	CP (P00450)	0	1	3.54E-06	0.0049	None
rs77303550	16	HPR (P00739)	0.77	3	1.09E-07	0.0002	None
rs41269133	6	PLG (P00747)	0	3	7.63E-07	0.0013	None
rs1801020	5	F12 (P00748)	0.26	15	5.54E-24	1.41E-17	Liver, brain
rs112287874	1	AGT (P01019)	0	3	6.99E-06	0.0085	None
rs142387042	9	C5 (P01031)	0	1	1.86E-06	0.0028	None
rs145688178	2	IL1B (P01584)	24.42	1	6.71E-08	0.0002	None
rs72756526	9	AMBP (P02760)	0	10	3.39E-07	0.0007	None
rs13286883	9	A1AG1 (P02763)	0	1	3.58E-06	0.005	None
rs149850735	11	HPX (P02790)	0	3	2.20E-06	0.0033	None
rs9898	3	HRG (P04196)	0	27	1.05E-09	4.76E-06	None
rs150327239	12	VWF (P04275)	54.87	4	1.63E-06	0.0025	None
rs113005658	4	CFI (P05156)	0	1	1.24E-09	5.56E-06	None
rs111686156	1	ALPL (P05186)	20.82	1	5.66E-07	0.0011	None
rs147671350	1	TPM3 (P06753)	71.54	1	3.35E-06	0.0047	None
rs117073451	14	HSP90AA1 (P07900)	66.67	7	2.99E-09	1.18E-05	None
rs77782541	1	FH (P07954)	36.76	1	5.32E-06	0.0067	None
rs148760384	4	SOD3 (P08294)	50.77	28	7.58E-08	0.0002	None
rs56393506	6	APOA (P08519)	15.42	4	1.25E-11	1.17E-07	None
rs2894255	6	C4A (P0C0L4)	0	1,560	3.07E-16	4.60E-11	See Box 7.1
rs9268145	6	C4B (P0C0L5)	6.92	21	6.46E-07	0.0011	None
rs151142563	2	HSPD1 (P10809)	48.21	2	4.41E-06	0.0059	None
rs77900687	8	CLU (P10909)	0	29	4.34E-09	1.50E-05	None

Most associated SNP	Chr	Protein (UniProt ID)	% protein values missing before imputation	Total SNPs for protein	p-value	Adjusted p-value	Tissue of expression of corresponding eQTL
rs143494653	19	GLU2B (P14314)	43.96	2	6.86E-07	0.0012	None
rs112641969	15	PKM (P14618)	65.04	4	4.71E-06	0.0062	None
rs138661388	15	AMPN (P15144)	41.28	1	2.25E-07	0.0005	None
rs141862094	13	VGFR1 (P17948)	43.85	2	1.74E-06	0.0027	None
rs117607817	10	VCL (P18206)	23.91	1	4.17E-08	0.0001	None
rs35196210	12	TNFRSF1A (P19438)	68.89	1	6.48E-06	0.008	None
rs149248320	5	LMNB1 (P20700)	21.03	1	3.55E-06	0.0049	None
rs11713634	3	CPN2 (P22792)	0	15	3.73E-22	1.58E-16	None
rs61966459	13	PROZ (P22891)	9.77	8	1.08E-06	0.0018	Liver, pituitary, stomach, breast, thyroid, whole blood
rs2173194	9	TNC (P24821)	49.36	6	1.72E-07	0.0004	None
rs11600340	11	DDX6 (P26196)	44.22	6	4.14E-09	1.43E-05	None
rs854562		PON1 (P27169)	0	9	3.23E-07	0.0006	Liver, adrenal gland, testis, prostate, pancreas, ovary
rs117919176	7	RDX (P35241)	64.27	26	1.39E-10	8.85E-07	None
rs142553639	17	KRT20 (P35900)	30.33	5	2.18E-07	0.0004	None
rs12066959	1	CFHR2 (P36980)	1.03	116	1.42E-10	8.88E-07	None
rs117050879	7	MDH2 (P40926)	67.69	7	1.54E-07	0.0003	None
rs117045022	9	PTGDS (P41222)	72.31	1	7.66E-06	0.0093	None
rs13056865	22	RANBP1 (P43487)	67.18	2	8.60E-06	0.0103	None
rs114756928	4	AFM (P43652)	0	1	2.37E-10	1.46E-06	None
rs115631137	3	UBE2E1 (P51965)	69.15	1	1.93E-09	8.16E-06	None
rs192265627	1	HSPG2 (P98160)	67.61	2	3.14E-06	0.0045	None
rs148611864	10	PFKP (Q01813)	26.41	4	3.57E-10	1.86E-06	None
rs10737680	1	CFHR1 (Q03591)	11.54	177	6.20E-13	7.90E-09	See Box 7.2
rs74918287	6	PAFA (Q13093)	48.59	6	2.40E-08	6.39E-05	None

Most associated SNP	Chr	Protein (UniProt ID)	% protein values missing before imputation	Total SNPs for protein	p-value	Adjusted p-value	Tissue of expression of corresponding eQTL
rs57565725	16	COTL1 (Q14019)	48.46	1	3.26E-06	0.0046	None
rs138943167	9	ZNF169 (Q14929)	55.01	2	5.44E-06	0.0069	None
rs67436553	12	NAA25 (Q14CX7)	48.97	1	7.42E-07	0.0013	None
rs140435033	12	KRT77 (Q7Z794)	41.03	3	9.71E-06	0.0113	None
rs142569471	5	DCP2 (Q8IU60)	70.18	1	8.52E-07	0.0014	None
rs12122975	1	NAXE (Q8NCW5)	65.13	5	3.89E-07	0.0008	None
rs118057809	14	SCFD1 (Q8WVM8)	34.1	1	1.30E-07	0.0003	None
rs2179485	20	RRBP1 (Q9P2E9)	56.67	28	5.80E-07	0.0011	None
rs141173103	10	MINPP1 (Q9UNW1)	53.33	3	5.10E-08	0.0001	None
rs277462	1	PADI2 (Q9Y2J8)	24.16	1	3.19E-06	0.0045	None
rs73061453	3	DYNC1LI1 (Q9Y6G9)	35.48	1	2.29E-06	0.0034	None

ABBREVIATIONS: 60 kDa heat shock protein, mitochondrial (HSPD1), α-1-acid glycoprotein 1 (A1AG1), afamin (AFM), alkaline phosphatase, tissue-nonspecific isozyme (ALPL), aminopeptidase N (AMPN), angiotensinogen (AGT), apolipoprotein(a) (APOA), apolipoprotein L1 (APOL1), ATP-dependent 6-phosphofructokinase, platelet type (PFKP), basement membrane-specific heparan sulfate proteoglycan core protein (HSPG2), caeruloplasmin (CP), carboxypeptidase N subunit 2 (CPN2), chromosome (chr), clusterin (CLU), coactosin-like protein (COTL1), coagulation factor XII (F12), complement C4-A (C4A), complement C4-B (C4B), complement C5 (C5), complement factor H-related protein 1 (CFHR1), complement factor H-related protein 2 (CFHR2), complement factor I (CFI), cytoplasmic dynein 1 light intermediate chain 1 (DYNC1LI1), expression quantitative trait locus/loci (eQTL), extracellular superoxide dismutase [Cu-Zn] (SOD3), fumarate hydratase, mitochondrial (FH), glucosidase 2 subunit β (GLU2B), haemopexin (HPX), haptoglobin-related protein (HRP), heat shock protein HSP 90-α (HSP90AA1), histidine-rich glycoprotein (HRG), identifier (ID), inhibitor of nuclear factor κ-B kinase subunit α (IKKA), interleukin-1 β (IL1B), keratin, type I cytoskeletal 20 (KRT20), keratin, type II cytoskeletal 1b (KRT77), lamin-B1 (LMNB1), m7GpppN-mRNA hydrolase (DCP2), malate dehydrogenase, mitochondrial (MDH2), multiple inositol polyphosphate phosphatase 1 (MINPP1), N-α-acetyltransferase 25, NAD(P)H-hydrate epimerase (NAXE), NatB auxiliary subunit (NAA25), plasminogen (PLG), platelet-activating factor acetylhydrolase (PAFA), probable ATP-dependent RNA helicase DDX6 (DDX6), prostaglandin-H2 D-isomerase (PTGDS), protein AMBP (AMBP), protein-arginine deiminase type-2 (PADI2), protein quantitative trait locus/loci (pQTL), pyruvate kinase PKM (PKM), radixin (RDX), Ran-specific GTPase-activating protein (RANBP1), ribosome-binding protein 1 (RRBP1), Sec1 family domain-containing protein 1 (SCFD1), serum paraoxonase/arylesterase 1 (PON1), single nucleotide polymorphism (SNP), tenascin (TNC), tropomyosin α-3 chain (TPM3), tumour necrosis

factor receptor superfamily member 1A (TNFRSF1A), ubiquitin-conjugating enzyme E2 E1 (UBE2E1), vascular endothelial growth factor receptor 1 (VGFR1), vinculin (VCL), vitamin K-dependent protein Z (PROZ), von Willebrand factor (VWF), zinc finger protein 169 (ZNF169).

Box 7.1. Tissues where eQTLs for rs2894255 (C4A gene) are known to be expressed.

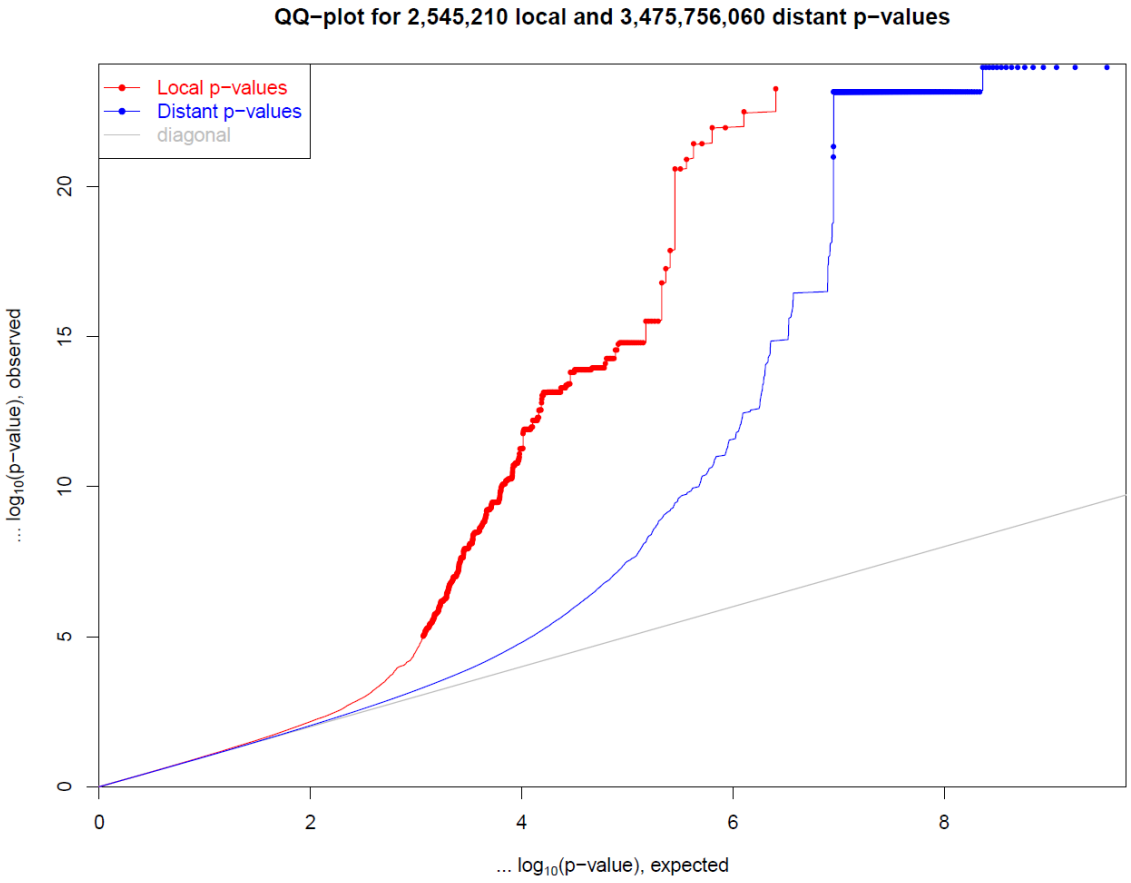
Lung	Skin	Spleen	Salivary gland
Thyroid	Testis	Prostate	Pituitary
Tibial nerve	Heart	Ovary	Spinal cord
Oesophagus	Skeletal muscle	Kidney	Vagina
Adipose tissue	Colon	Epstein Barr Virus-transformed lymphocytes	
Arteries	Liver	Uterus	Stomach
Whole blood	Brain	Adrenal gland	Breast
Cultured fibroblasts	Pancreas	Small intestine	

Box 7.2. Tissues where eQTLs for rs10737680 (CFHR1 gene) are known to be expressed.

Liver	Cultured fibroblasts	Brain	Adipose tissue
Spleen	Testis	Adrenal gland	Nerve

Overlaps between significant proteins from baseline and three-month pQTL analyses are summarised in Table 7.5 and Figure 7.3. Overlaps between significant proteins from all pQTL analyses and proteins associated with RA disease outcome measures from Sections 5.6 to 5.8 are summarised in Table 7.6. These proteins were then input into the STRING database to determine whether there were any known interactions between these proteins, and this is presented in Figure 7.4.

Figure 7.1. Q-Q plot of genome-wide pQTLs at baseline (before treatment with etanercept).



LEGEND: Q-Q plot of –log10(p-value) of genome-wide pQTL analysis before treatment with etanercept. The red points and line represent the statistical p-values for cis (or local) pQTLs. The blue points and line represent the statistical p-values for trans (or distant) pQTLS.

Box 7.3. Tissues where eQTLs for rs190820372 (CFHR1 gene) are known to be expressed.

Liver	Adipose tissue	Nerve	Spleen
Brain	Cultured fibroblasts	Adrenal gland	Breast
Colon	Small intestine	Lung	Stomach
Pancreas			

Table 7.4. SNPs with lowest false-discovery rate-adjusted p-values for each protein included in pQTL study after 3 months treatment with etanercept.

Most associated SNP	**Chr**	**Protein (UniProt ID)**	**% protein values missing before imputation**	**Total SNPs for protein**	**p-value**	**Adj p-value**	**Tissue of expression of corresponding eQTL**
rs114226893	22	APOL1 (O14791)	0.26	9	1.13E-09	1.39E-05	None
rs10510794	3	FLNB (O75369)	52.82	276	9.02E-10	1.19E-05	Cultured fibroblasts
rs34417180	16	HPR (P00739)	0.77	2	3.68E-06	0.0074	Heart, nerve
rs8191936	6	PLG (P00747)	0	2	3.30E-06	0.0067	None
rs1801020	5	F12 (P00748)	0.26	12	1.70E-23	2.87E-17	Liver, brain
rs113568276	6	CFB (P00751)	0	4	3.22E-06	0.0065	None
rs72823478	2	IL1B (P01584)	24.42	1	1.49E-07	0.0005	None
rs140942977	1	C1QC (P02747)	0	1	9.85E-07	0.0023	None
rs55825809	19	FTL (P02792)	52.19	1	9.70E-06	0.0173	None
rs150845792	16	ALDOA (P04075)	84.62	2	1.11E-06	0.0026	None
rs78370639	6	SOD2 (P04179)	44.47	1	2.05E-07	0.0006	None
rs13073829	3	HRG (P04196)	0	21	3.09E-07	0.0009	None
rs11609243	12	IGF1 (P05019)	78.97	1	3.09E-06	0.0063	None
rs41497052	1	ALPL (P05186)	20.82	5	1.70E-08	7.71E-05	None
rs144647167	14	PYGL (P06737)	69.74	1	2.03E-07	0.0006	None
rs117384764	12	LDHB (P07195)	41.39	1	9.01E-06	0.0164	None
rs79918073	11	CTSD (P07339)	54.62	1	1.33E-07	0.0004	None
rs116684782	4	SOD3 (P08294)	50.77	8	9.23E-10	1.20E-05	None
rs78924361	14	CTSG (P08311)	68.89	3	1.17E-06	0.0027	None
rs9295128	6	APOA (P08519)	15.42	22	1.05E-07	0.0003	None
rs76187874	17	A2AP (P08697)	0	22	1.86E-06	0.004	None
rs2894255	6	C4A (P0C0L4)	0	635	1.91E-11	1.62E-06	See Box 7.1
rs204883	6	C4B (P0C0L5)	6.92	26	3.27E-07	0.0009	Whole blood
rs61904443	11	HSPA8 (P11142)	28.28	2	5.18E-06	0.0102	None
rs150297179	22	XRCC6 (P12956)	60.26	1	5.29E-06	0.0104	None

Most associated SNP	Chr	Protein (UniProt ID)	% protein values missing before imputation	Total SNPs for protein	p-value	Adj p-value	Tissue of expression of corresponding eQTL
rs145827445	2	XRCC5 (P13010)	15.13	1	8.95E-06	0.0163	None
rs187550307	15	PKM (P14618)	65.04	2	8.91E-06	0.0163	None
rs9806694	15	AMPN (P15144)	41.28	2	1.44E-06	0.0032	None
rs147009906	11	CD44 (P16070)	11.83	1	6.76E-07	0.0017	None
rs79872280	3	ITIH1 (P19827)	0	3	2.96E-07	0.0009	None
rs11713634	3	CPN2 (P22792)	0	12	4.57E-11	2.75E-06	None
rs61966459	13	PROZ (P22891)	9.77	21	1.36E-07	0.0004	Liver, pituitary, stomach, breast, thyroid, whole blood
rs854562	7	PON1 (P27169)	0	8	5.28E-08	0.0002	Liver, adrenal gland, testis, prostate, pancreas, ovary
rs138816256	5	CANX (P27824)	45.5	10	1.16E-07	0.0004	None
rs12231148	12	PTPN6 (P29350)	9	20	1.35E-18	2.64E-13	None
rs139074486	19	BLVRB (P30043)	77.95	2	1.39E-09	1.65E-05	None
rs117457005	11	RDX (P35241)	64.27	1	1.79E-06	0.0038	None
rs10801582	1	CFHR2 (P36980)	1.03	94	1.92E-09	1.88E-05	None
rs148555185	1	TAGLN2 (P37802)	26.99	1	1.45E-06	0.0032	None
rs148212408	7	MDH2 (P40926)	67.69	1	1.50E-06	0.0033	None
rs10460510	2	MRPL19 (P49406)	34.7	3	4.26E-08	0.0001	None
rs115373136	1	HDGF (P51858)	56.56	1	6.15E-07	0.0015	None
rs6062997	20	PLTP (P55058)	3.6	5	1.55E-06	0.0034	None
rs145005927	20	DSTN (P60981)	58.35	1	8.31E-07	0.002	None
rs72691742	14	CALM2 (P62158)	5.64	1	3.65E-06	0.0074	None
rs6665824	1	CFHR1 (Q03591)	11.54	136	1.35E-11	1.37E-06	See Box 7.3
rs190820372	7	CALD1 (Q05682)	76.67	3	3.12E-13	4.18E-08	None
rs187671161	13	FOXO1 (Q12778)	64.1	2	2.19E-07	0.0007	None
rs11673692	19	ILF3 (Q12906)	63.08	1	9.48E-06	0.017	None

Most associated SNP	Chr	Protein (UniProt ID)	% protein values missing before imputation	Total SNPs for protein	p-value	Adj p-value	Tissue of expression of corresponding eQTL
rs696044	3	CAMK1 (Q14012)	62.82	1	1.42E-06	0.0032	None
rs79360255	16	COTL1 (Q14019)	48.46	3	1.18E-07	0.0004	None
rs9906599	17	GRB7 (Q14451)	6.68	2	4.21E-06	0.0084	None
rs142995251	17	LASP1 (Q14847)	77.89	7	5.86E-09	4.35E-05	None
rs117452594	10	RSU1 (Q15404)	58.21	2	2.87E-07	0.0008	None
rs116838200	3	MYLK (Q15746)	54.36	1	2.22E-06	0.0046	None
rs62402216	6	MAPK14 (Q16539)	53.59	3	1.87E-07	0.0006	None
rs73121704	12	CAND1 (Q86VP6)	48.72	3	8.14E-11	2.91E-06	None
rs115373136	1	NAXE (Q8NCW5)	65.13	7	1.75E-14	2.62E-09	None
rs141739774	16	DNAH3 (Q8TD57)	38.05	1	3.06E-09	2.56E-05	None
rs113930700	6	PHIP (Q8WWQ0)	47.3	1	7.54E-07	0.0019	None
rs151286550	1	PARK7 (Q99497)	72.24	1	1.13E-06	0.0026	None
rs150571376	2	TCPH (Q99832)	51.28	1	1.33E-06	0.003	None
rs145505184	11	C11orf54 (Q9H0W9)	49.23	1	3.40E-07	0.0009	None
rs188695391	11	EHD1 (Q9H4M9)	72.82	1	5.58E-06	0.0107	None
rs4687154	3	IL1RAP (Q9NPH3)	74.29	1	7.01E-06	0.0131	None
rs147277776	20	RRBP1 (Q9P2E9)	56.67	1	6.90E-07	0.0017	None
rs62290911	3	FETUB (Q9UGM5)	1.79	1	8.72E-06	0.016	None
rs147084701	10	MINPP1 (Q9UNW1)	53.33	2	7.34E-07	0.0018	None

ABBREVIATIONS: 39S ribosomal protein L19, mitochondrial (MRPL19), α-2-antiplasmin (A2AP), adjusted (adj), alkaline phosphatase, tissue-nonspecific isozyme (ALPL), aminopeptidase N (AMPN), apolipoprotein(a) (APOA), apolipoprotein L1 (APOL1), calcium/calmodulin-dependent protein kinase type 1 (CAMK1), caldesmon (CALD1), calmodulin-2 (CALM2), calnexin (CANX), carboxypeptidase N subunit 2 (CPN2), cathepsin D (CTSD), cathepsin G (CTSG), CD44 antigen (CD44), chromosome (chr), coactosin-like protein (COTL1), coagulation factor XII (F12), complement C1q subcomponent subunit C (C1QC), complement C4-A (C4A), complement C4-B (C4B), complement factor B (CFB), complement factor H-related protein 1 (CFHR1), complement factor H-related protein 2 (CFHR2), cullin-associated NEDD8-dissociated protein 1 (CAND1), destrin (DSTN), dynein axonemal heavy chain 3 (DNAH3), EH domain-containing protein 1 (EHD1), ester hydrolase C11orf54 (C11orf54), expression quantitative trait locus/loci (eQTL), extracellular superoxide dismutase [Cu-Zn] (SOD3), ferritin light chain (FTL), fetuin-B (FETUB), filamin-B

(FLNB), flavin reductase (NADPH) (BLVRB), forkhead box protein O1 (FOXO1), fructose-bisphosphate aldolase A (ALDOA), glycogen phosphorylase, liver form (PYGL), growth factor receptor-bound protein 7 (GRB7), haptoglobin-related protein (HPR), heat shock cognate 71 kDa protein (HSPA8), hepatoma-derived growth factor (HDGF), histidine-rich glycoprotein (HRG), identifier (ID), insulin-like growth factor I (IGF1), inter-α-trypsin inhibitor heavy chain H1 (ITIH1), interleukin-1 β (IL1B), interleukin-1 receptor accessory protein (IL1RAP), interleukin enhancer-binding factor 3 (ILF3), LIM and Sh3 domain protein (LASP1), L-lactate dehydrogenase B chain (LDHB), malate dehydrogenase, mitochondrial (MDH2), mitogen-activated protein kinase 14 (MAPK14), multiple inositol polyphosphate phosphatase 1 (MINPP1), myosin light chain kinase, smooth muscle (MYLK), NAD(P)H-hydrate epimerase (NAXE), Parkinson disease protein 7 (PARK7), PH-interacting protein (PHIP), phospholipid transfer protein (PLTP), plasminogen (PLG), protein quantitative trait locus/loci (pQTL), pyruvate kinase PKM (PKM), Ras suppressor protein 1 (RSU1), serum paraoxonase/arylesterase 1 (PON1), radixin (RDX), ribosome-binding protein 1 (RRBP1), single nucleotide polymorphism (SNP), superoxide dismutase [Mn], mitochondrial (SOD2), T-complex protein 1 subunit η (TCPH), transgelin-2 (TAGLN2), tyrosine-protein phosphatase non-receptor type 6 (PTPN6), vitamin K-dependent protein Z (PROZ), X-ray repair cross-complementing protein 5 (XRCC5), X-ray repair cross-complementing protein 6 (XRCC6).

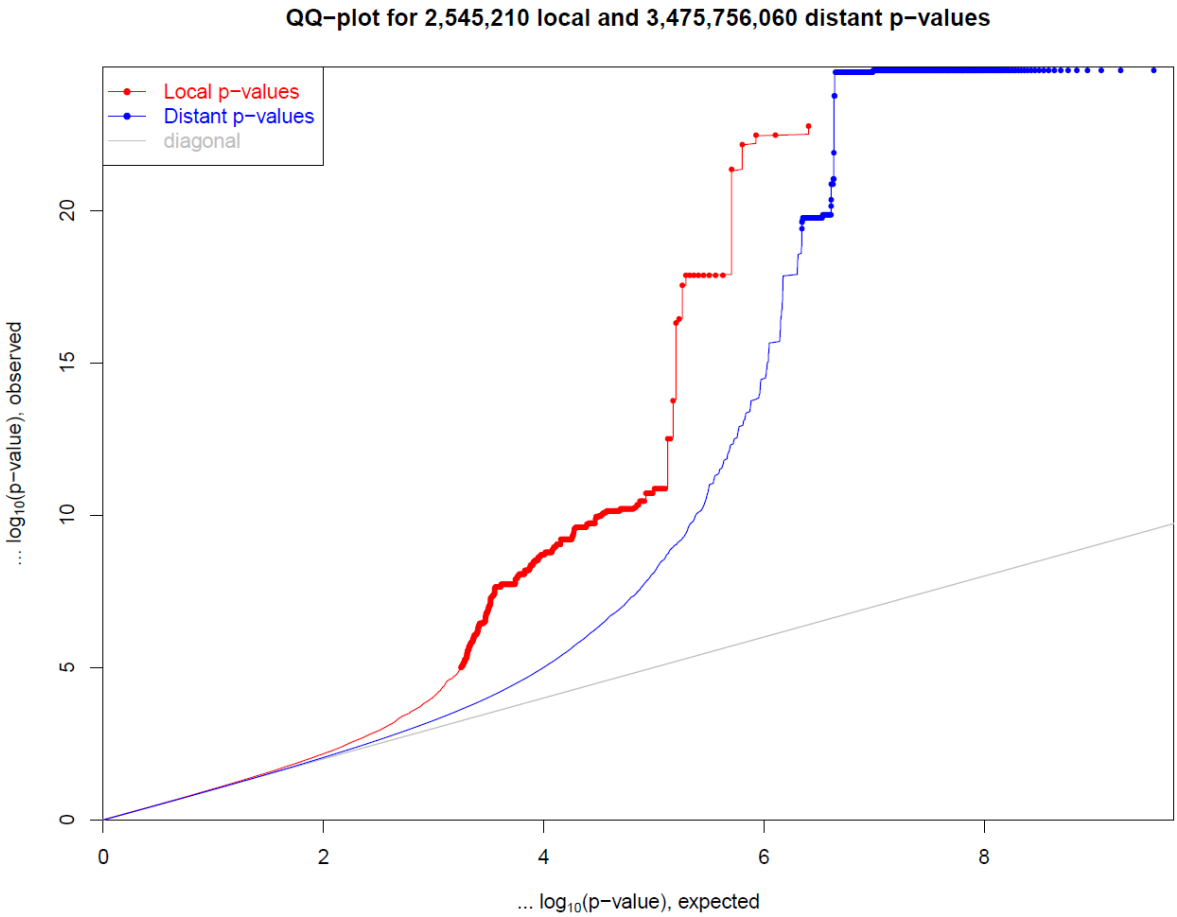
7.1.3. Derivation of polygenic risk scores predictive of poor EULAR response

Significant SNPs from the pQTL analysis were then included in PRS models to determine genetic predictors of poor EULAR response after three or six months of treatment with TNFi, using the cohort of 1,563 extant BRAGGSS genotyped samples.

At baseline (before treatment with TNFi), a PRS model was derived that included 48 significant SNPs. However, this model only explained 0.26% of variance in EULAR response in the target cohort; covariates alone explained only 2.25% of the variance. The overall p-value of the model was non-significant, at 0.1176.

After three or six months of treatment with TNFi, a polygenic risk score model was derived that included 37 significant SNPs. This model only explained 0.05% of variance in EULAR response in the target cohort; covariates alone explained 65.47% of the variance. The overall p-value of the model was non-significant, at 0.5915.

Figure 7.2. Q-Q plot of genome-wide pQTLs after 3 months of treatment with etanercept.



LEGEND: Q-Q plot of –log10(p-value) of genome-wide pQTL analysis after 3 months of treatment with etanercept. The red points and line represent the statistical p-values for cis (or local) pQTLs. The blue points and line represent the statistical p-values for trans (or distant) pQTLS.

Table 7.5. All significant proteins identified from *cis* pQTL analyses.

Protein	**pQTL at baseline**	**pQTL at 3 months**	**pQTL at baseline and 3 months**
APOL1 (O14791)			X
IKKA (O15111)	X		
FLNB (O75369)		X	
CP (P00450)	X		
HPR (P00739)			X
PLG (P00747)			X
F12 (P00748)			X
CFB (P00751)		X	
AGT (P01019)	X		
C5 (P01031)	X		
IL1B (P01584)			X
C1QC (P02747)		X	
AMBP (P02760)	X		
A1AG1 (P02763)	X		
HPX (P02790)	X		
FTL (P02792)		X	
ALDOA (P04075)		X	
SOD2 (P04179)		X	
HRG (P04196)			X
VWF (P04275)	X		
IGF1 (P05019)		X	
CFI (P05156)	X		
ALPL (P05186)			X
PYGL (P06737)		X	
TPM3 (P06753)	X		
LDHB (P07195)		X	
CTSD (P07339)		X	
HSP90AA1 (P07900)	X		
FH (P07954)	X		
SOD3 (P08294)			X
CTSG (P08311)		X	
APOA (P08519)			X
A2AP (P08697)		X	
C4A (P0C0L4)			X
C4B (P0C0L5)			X
HSPD1 (P10809)	X		
CLU (P10909)	X		
HSPA8 (P11142)		X	
XRCC6 (P12956)		X	
XRCC5 (P13010)		X	
GLU2B (P14314)	X		
PKM (P14618)			X
AMPN (P15144)			X

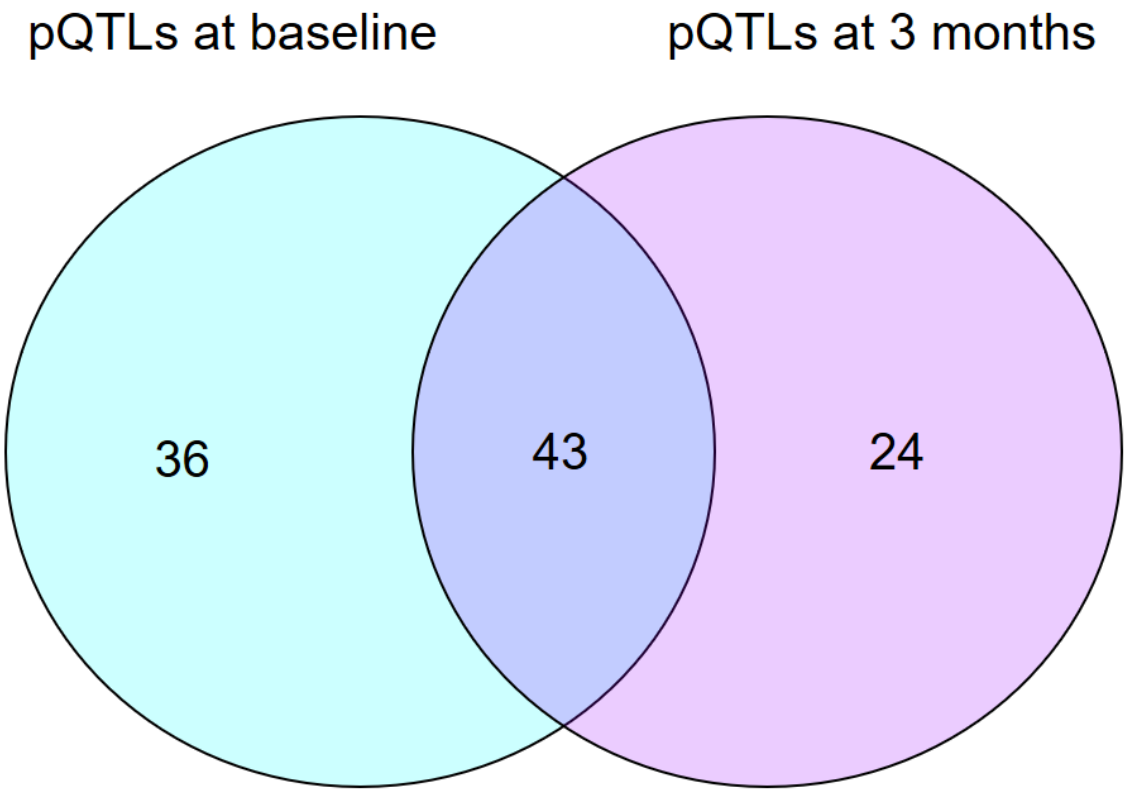
Protein	pQTL at baseline	pQTL at 3 months	pQTL at baseline and 3 months
CD44 (P16070)		X	
VGFR1 (P17948)	X		
VCL (P18206)	X		
TNFRSF1A (P19438)	X		
ITIH1 (P19827)		X	
LMNB1 (P20700)	X		
CPN2 (P22792)			X
PROZ (P22891)			X
TNC (P24821)	X		
DDX6 (P26196)	X		
PON1 (P27169)			X
CANX (P27824)		X	
PTPN6 (P29350)		X	
BLVRB (P30043)		X	
RDX (P35241)			X
KRT20 (P35900)	X		
CFHR2 (P36980)			X
TAGLN2 (P37802)		X	
MDH2 (P40926)			X
PTGDS (P41222)	X		
RANBP1 (P43487)	X		
AFM (P43652)	X		
MRPL19 (P49406)		X	
HDGF (P51858)		X	
UBE2E1 (P51965)	X		
PLTP (P55058)		X	
DSTN (P60981)		X	
CALM2 (P62158)		X	
HSPG2 (P98160)	X		
PFKP (Q01813)	X		
CFHR1 (Q03591)			X
CALD1 (Q05682)		X	
FOXO1 (Q12778)		X	
ILF3 (Q12906)		X	
PAFA (Q13093)	X		
CAMK1 (Q14012)		X	
COTL1 (Q14019)			X
GRB7 (Q14451)		X	
LASP1 (Q14847)		X	
ZNF169 (Q14929)	X		
NAA25 (Q14CX7)	X		
RSU1 (Q15404)		X	
MYLK (Q15746)		X	
MAPK14 (Q16539)		X	
KRT77 (Q7Z794)	X		

Protein	pQTL at baseline	pQTL at 3 months	pQTL at baseline and 3 months
CAND1 (Q86VP6)		X	
DCP2 (Q8IU60)	X		
NAXE (Q8NCW5)			X
DNAH3 (Q8TD57)		X	
SCFD1 (Q8WVM8)	X		
PHIP (Q8WWQ0)	X		
PARK7 (Q99497)		X	
TCPH (Q99832)		X	
C11orf54 (Q9H0W9)		X	
EHD1 (Q9H4M9)		X	
IL1RAP (Q9NPH3)		X	
RRBP1 (Q9P2E9)			X
FETUB (Q9UGM5)		X	
MINPP1 (Q9UNW1)			X
PADI2 (Q9Y2J8)	X		
DYNC1LI1 (Q9Y6G9)	X		

ABBREVIATIONS: 39S ribosomal protein L19, mitochondrial (MRPL19), 60 kDa heat shock protein, mitochondrial (HSPD1), α-1-acid glycoprotein 1 (A1AG1), α-2-antiplasmin (A2AP), afamin (AFM), alkaline phosphatase, tissue-nonspecific isozyme (ALPL), aminopeptidase N (AMPN), angiotensinogen (AGT), apolipoprotein(a) (APOA), apolipoprotein L1 (APOL1), ATP-dependent 6-phosphofructokinase, platelet type (PFKP), basement membrane-specific heparan sulfate proteoglycan core protein (HSPG2), caeruloplasmin (CP), calcium/calmodulin-dependent protein kinase type 1 (CAMK1), caldesmon (CALD1), calmodulin-2 (CALM2), calnexin (CANX), carboxypeptidase N subunit 2 (CPN2), cathepsin D (CTSD), cathepsin G (CTSG), CD44 antigen (CD44), clusterin (CLU), coactosin-like protein (COTL1), coagulation factor XII (F12), complement C1q subcomponent subunit C (C1QC), complement C4-A (C4A), complement C4-B), complement C5 (C5), complement factor B (CFB), complement factor H-related protein 1 (CFHR1), complement factor H-related protein 2 (CFHR2), complement factor I (CFI), Cullin-associated NEDD8-dissociated protein 1 (CAND1), cytoplasmic dynein 1 light intermediate chain 1 (DYNC1LI1), destrin (DSTN), dynein axonemal heavy chain 3 (DNAH3), EH domain-containing protein 1 (EHD1), ester hydrolase C11orf54 (C11orf54), extracellular superoxide dismutase [Cu-Zn] (SOD3), ferritin light chain (FTL), fetuin-B (FETUB), filamin-B (FLNB), flavin reductase (NADPH) (BLVRB), forkhead box protein O1 (FOXO1), fructose-bisphosphate aldolase A (ALDOA), fumarate hydratase, mitochondrial (FH), glucosidase 2 subunit β (GLU2B), glycogen phosphorylase, liver form (PYGL), growth factor receptor-bound protein 7 (GRB7), haemopexin (HPX), haptoglobin-related protein (HPR), heat shock cognate 71 kDa protein (HSPA8), heat shock protein HSP 90-α (HSP90AA1), hepatoma-derived growth factor (HDGF), histidine-rich glycoprotein (HRG), insulin-like growth factor 1 (IGF1), inter-α-trypsin inhibitor heavy chain H1 (ITIH1), interleukin-1 β (IL1B), interleukin-1 receptor accessory protein (IL1RAP), interleukin enhancer-binding factor 3 (ILF3), inhibitor of NF-κB kinase subunit α (IKKA), keratin, type I cytoskeletal 20 (KRT20), keratin, type II cytoskeletal 1b (KRT77), lamin-B1 (LMNB1), LIM and SH3 domain protein 1 (LASP1), L-lactate dehydrogenase B chain (LDHB), m7GpppN-mRNA hydrolase (DCP2), malate dehydrogenase, mitochondrial (MDH2), mitogen-activated protein kinase 14 (MAPK14), multiple inositol polyphosphate phosphatase 1

(MINPP1), myosin light chain kinase, smooth muscle (MYLK), N-α-acetyltransferase 25, NatB auxiliary subunit (NAA25), NAD(P)H-hydrate epimerase (NAXE), Parkinson disease protein 7 (PARK7), PH-interacting protein (PHIP), phospholipid transfer protein (PLTP), plasminogen (PLG), platelet-activating factor acetylhydrolase (PAFA), probable ATP-dependent RNA helicase DDX6 (DDX6), prostaglandin-H2 D-isomerase (PTGDS), protein AMBP (AMBP), protein-arginine deiminase type-2 (PADI2), protein quantitative trait locus/loci (pQTL), pyruvate kinase PKM (PKM), radixin (RDX), Ran-specific GTPase-activating protein (RANBP1), Ras suppressor protein 1 (RSU1), ribosome-binding protein 1 (RRBP1), Sec1 family domain-containing protein 1 (SCFD1), serum paraoxonase/arylesterase 1 (PON1), superoxide dismutase [Mn], mitochondrial (SOD2), T-complex protein 1 subunit η (TCPH), tenascin (TNC), transgelin-2 (TAGLN2), tropomyosin α-3 chain (TPM3), tumour necrosis factor receptor superfamily member 1A (TNFRSF1A), tyrosine-protein phosphatase non-receptor type 6 (PTPN6), ubiquitin-conjugating enzyme E2 E1 (UBE2E1), vascular endothelial growth factor receptor 1 (VGFR1), vinculin (VCL), vitamin K-dependent protein Z (PROZ), von Willebrand factor (VWF), X-ray repair cross-complementing protein 5 (XRCC5), X-ray repair cross-complementing protein 6 (XRCC6), zinc finger protein 169 (ZNF169).

Figure 7.3. Venn diagram illustrating number of *cis* pQTLs significant at baseline, 3 months or both time points.



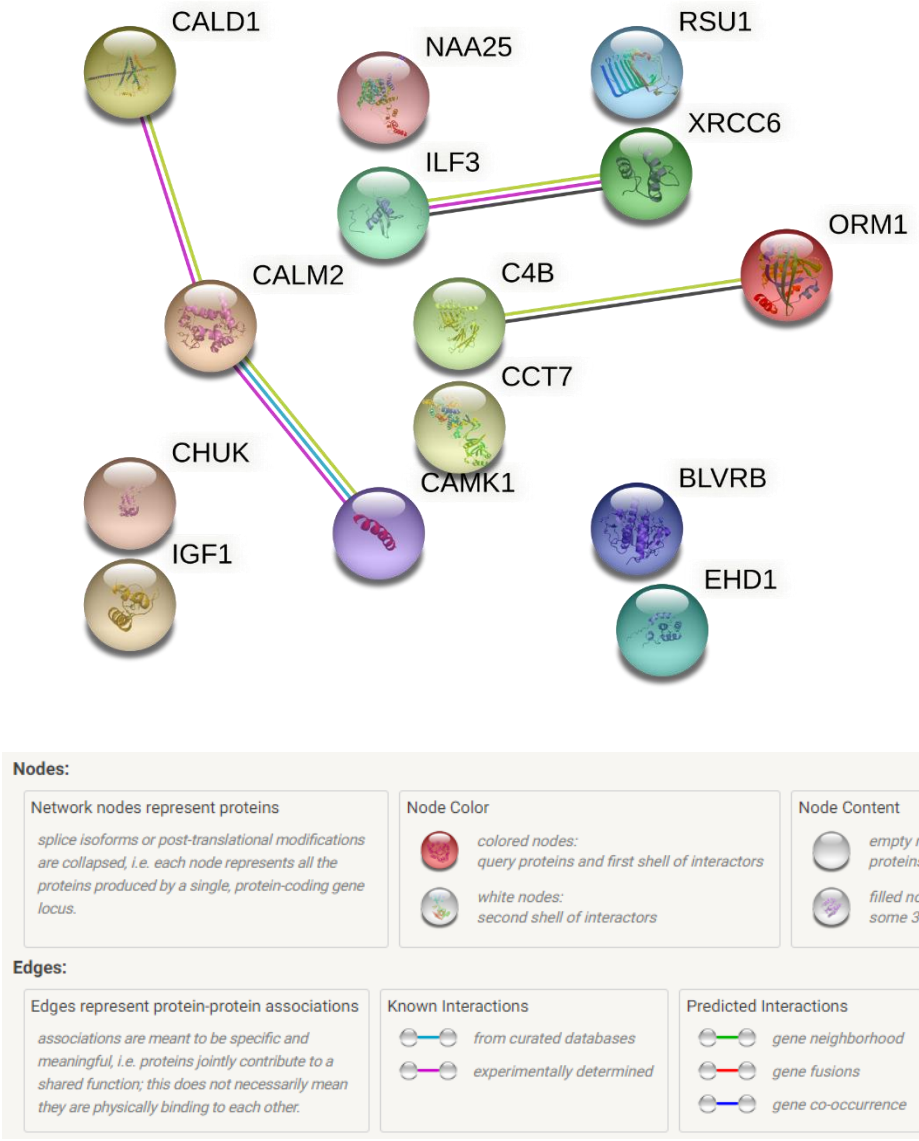
ABBREVIATIONS: Protein quantitative trait locus/loci (pQTL).

Table 7.6. Overlaps between significant proteins from pQTL analyses and proteins significantly associated with RA clinical response outcome measures.

Protein	**UniProt ID**
IKKA	O15111
A1AG1	P02763
IGF1	P05019
C4B	P0C0L5
XRCC6	P12956
BLVRB	P30043
CALM2	P62158
CALD1	Q05682
ILF3	Q12906
CAMK1	Q14012
NAA25	Q14CX7
RSU1	Q15404
TCPH	Q99832
EHD1	Q9H4M9

ABBREVIATIONS: α-1-acid glycoprotein 1 (A1AG1), calcium/calmodulin-dependent protein kinase type 1 (CAMK1), caldesmon (CALD1), calmodulin-2 (CALM2), complement C4-B (C4B), EH domain-containing protein 1 (EHD1), flavin reductase (NADPH) (BLVRB), inhibitor of nuclear factor κ-B kinase subunit α (IKKA), insulin-like growth factor I (IGF1), interleukin enhancer-binding factor 3 (ILF3), N-α-acetyltransferase 25, NatB auxiliary subunit (NAA25), Ras suppressor protein 1 (RSU1), T-complex protein 1 subunit η (TCPH), X-ray repair cross-complementing protein 6 (XRCC6).

Figure 7.4. Protein interactions between significant proteins from Table 7.6.



ABBREVIATIONS: α-1-acid glycoprotein 1 (A1AG1, also known as ORM1), calcium/calmodulin-dependent protein kinase type 1 (CAMK1), caldesmon (CALD1), calmodulin-2 (CALM2), complement C4-B (C4B), EH domain-containing protein 1 (EHD1), flavin reductase (NADPH) (BLVRB), inhibitor of nuclear factor κ-B kinase subunit α (IKKA, also known as CHUK), insulin-like growth factor I (IGF1), interleukin enhancer-binding factor 3 (ILF3), N-α-acetyltransferase 25, NatB auxiliary subunit (NAA25), Ras suppressor protein 1 (RSU1), T-complex protein 1 subunit η (TCPH, also known as CCT7), X-ray repair cross-complementing protein 6 (XRCC6).

7.2. Discussion

pQTL analyses were carried out in 147 patients with RA who were commencing etanercept with available genotyping and protein expression data at baseline and after three months of treatment. At baseline (pre-treatment), a total of 2,184 *cis* pQTLs were identified for 60 unique proteins. After three months of treatment with etanercept, a total of 1,432 cis pQTLs were identified for 68 unique proteins. 43 proteins had significant pQTL associations at both

baseline and three months. 14 proteins with significant pQTLs at either baseline or three months were also significantly associated with RA disease outcome measures as determined in Sections 5.7 and 5.8. Significant pQTLs from baseline and three months were used to attempt to develop PRS models to predict EULAR poor response at each respective time point. However, PRS derived from these pQTLs did not predict EULAR response by three or six months.

For brevity, only pQTLs with an adjusted p-value < 5E-08 will be discussed here; proteins overlapping with findings from Section 5.6 to 5.8 will be discussed in more detail in Chapter Eight. Four proteins were below this threshold at baseline. Coagulation factor XII (F12, UniProt ID P00748) is involved in the clotting cascade, and also participates in fibrinolysis and bradykinin and angiotensin production[379]. F12 has corresponding eQTLs in liver and brain. Carboxypeptidase N subunit 2 (CPN2, UniProt ID P22792) is an 83 kDa subunit that stabilises its catalytic subunit at 37^{o}C, enabling it to remain in systemic circulation[380].

C4A (UniProt ID P0C0L4), along with C4B (which was found to be significantly associated with therapeutic drug levels in Section 5.6), participates in the dissemination of the classical complement pathway[381]. C4A had multiple tissues in which corresponding eQTLs were found (see Box 7.1), and these tissues included cultured fibroblasts, skeletal muscle and adipose tissue, which could potentially participate in the active RA disease process. In addition, C4a anaphylatoxin (a proteolytic by-product of C4) is involved in local tissue inflammation, stimulating smooth muscle contraction, vascular permeability and the release of histamine from basophils and mast cells[382]. C4B also had significant pQTLs at both baseline and three months, and a genetic basis for these inflammatory mediators may give more robust evidence that C4A and B should be considered for future validation studies. Furthermore, mutations in the C4A and B genes has been shown to be causative of C4 deficiency in SLE; given that there is some clinical overlap between RA and SLE, these C4 components could represent future biomarkers of prognosis and/or disease activity that also have a genetic basis. CFHR1, along with CFHR3 (which was found to be associated with SJC, CRP and DAS28 in Section 5.7), is involved in complement regulation[338]. CFHR1 also has multiple tissues in which associated eQTLs are expressed (Box 7.2), including in cultured fibroblasts. Given that C4A and C4B genes seem to be strongly implicated in the expression of these complement proteins and the further strong association of pQTLs for CFHR1, the complement axis seems to be associated with ongoing inflammation in active RA.

After three months of treatment, there were four proteins with pQTLs with an adjusted p-value below a threshold of 5E-08, one of which was F12, whose function has already been described above. Tyrosine-protein phosphatase non-receptor type 6 (PTPN6) regulates the signalling of tyrosine phosphorylated cell surface receptors and is an important participant in haematopoiesis[383]. The function of CALD1 has been described in Section 6.5. NAD(P)H-hydrate epimerase (NAXE) is involved in the repair of a NADP(HX), a damaged form of NAD(P)H caused by enzyme or heat-dependent hydration[384]. This protein also modulates angiogenesis via acceleration of cholesterol efflux from endothelial cells to high-density lipoproteins[385].

Fourteen proteins associated with various RA disease outcome measures from Section 5.6 to 5.8 also had significant pQTLs and are presented in Table 7.6. These proteins were also input into the STRING database, and their interactions are shown in Figure 7.4. These proteins do not form one coherent network, although some on this list do have known interactions, as demonstrated in Figure 7.4. However, this may not be a disadvantage, as it means that different regions of the overall interaction network outlined in Figure 5.20 could be explored in the future, and understanding the mechanisms behind active RA and ineffective treatment response from a number of different angles could potentially lead to more targeted prediction and treatment of different subsets of patients in the future. Treatment could even be targeted using multiple pathways in patients with very active disease, to ensure rapid and comprehensive control of disease activity.

Results of the PRS modelling were disappointing, but not unexpected. As with the machine learning predictive models developed in Section 5.9, models predictive of poor EULAR response after treatment with etanercept were not statistically significant. This is likely to be due to similar reasons as discussed in Section 6.6, namely, either due to a true lack of association, or because poor EULAR response was not a sufficiently sensitive or objective treatment outcome measure. Because EULAR response is calculated from DAS28, pseudo-subjective measures that make up a total DAS28 such as TJC and patient-reported VAS of global health may not completely reflect ongoing biological processes affecting RA disease activity.

To date, there have been few published studies of pQTL analysis in patients with RA. Sun *et al* carried out a large-scale pQTL analysis and identified 1,927 pQTLs with 1,478 proteins and were able to identify a number of disease-associated loci from analysis[84]. However, all

samples in the Sun *et al* study were obtained from HCs and were not replicated in specific disease-affected cohorts, including not in patients with RA. Luo *et al* carried out a candidate gene Mendelian randomisation study in JIA patients, but this only utilised pQTLs close to two candidate genes and was not carried out in RA patients[386]. The most similar study to the work in this thesis was published recently by He *et al*: DNA methylation, eQTL and pQTL studies in PBMCs from 25 patients with RA and 18 HCs implicated the ribophorin II (RPN2) gene in RA pathogenesis[387]. However, this was a case-control study, and findings were not related back to RA clinical outcome measures. Results differ from those of the current study, which failed to identify RPN2 as a protein significantly associated with response to etanercept.

Strengths of this study include the high-quality measurement of several hundred proteins with genotype data available on the same patients to ensure a robust pQTL association study. This study has also identified a genetic basis for 14 of the proteins found to be associated with RA clinical outcome measures identified in previous sections, lending a more robust justification for these findings, as opposed to spurious false-positives due to multiple comparisons made using the SWATH-MS data. Furthermore, a moderately-sized target cohort with genotype data was available for the testing of PRS modelling using pQTLs that had been identified earlier. The pQTL-based PRS did not associate with response to treatment, however, given the sample sizes tested, large effects can be excluded.

As previously discussed, the lack of a PRS model that can accurately predict RA treatment response may well be due to the use of an outcome measure that is not closely linked to the underlying pathophysiology of RA. However, other factors could have influenced negative findings. For example, the covariates used to correct the pQTL analysis were different from those of the target data in the PRS analysis, as only limited data was available for the latter cohort. However, analysis without the use of any covariates with both cohorts (not presented) still demonstrated no predictive value in the polygenic model, so discrepancies in covariate inclusion are unlikely to have influenced results dramatically. Another contributory factor could have been protein selection: proteins that had an association could have been excluded from the original SWATH-MS extract and imputation of missing values may have influenced results. The PRS analysis may have been more sensitive if it had been focused to only include QTLs for proteins with evidence for association with response, as opposed to including all proteins passing QC from the original SWATH-MS extraction. Finally, this study has a relatively small sample size, which may have been under-powered to develop an

accurate prediction model. This does not mean that genetics should be discounted as a predictor of treatment response. Rather, that some or all of the above factors should be taken into account with future study design for genetic risk score construction.

In conclusion, pQTL analyses were carried out in RA patients commencing etanercept, and identified numerous loci that are significantly associated with paired protein expression data. 14 proteins from analyses in Chapter Five were found to have a genetic basis underlying their expression, and these could prove to be valuable candidates for future study. PRS modelling failed to fit statistically significant models predictive of RA treatment non-response.

7.3. Chapter summary

pQTL analyses were carried out at baseline and after three months of treatment with etanercept in 147 patients with RA who had paired genotype and protein expression data. At baseline, 2,184 *cis* pQTLs were identified in 60 proteins and at three months, 1,570 *cis* pQTLs were identified in 68 proteins. Significant pQTLs were included in polygenic risk score models, but failed to predict treatment non-response at three months.

CHAPTER EIGHT: DISCUSSION

Summary of chapter contents:

This thesis has sought evidence to confirm the hypothesis that biological factors, such as protein expression, contribute to variability in circulating drug levels and treatment response to biologic agents in patients with RA. By confirming this hypothesis and identifying such factors, clinical rheumatology practice can move towards a precision medicine approach, whether by personalising dosing regimens of specific agents, or by selecting therapeutic agents which are most likely to be successful in controlling disease activity in patients with RA, given their phenotypic and/or genotypic profile. By ensuring patients receive the right drug at the right time, this will maximise cost-benefit by reducing prescription of ineffective medications that have a cost to the NHS, as well as controlling a patient's active RA more rapidly, preventing the cost of treating long-term sequelae such as joint destruction, and reducing the costs of increased healthcare encounters due to disease flares. Furthermore, some patients may be able to receive biologic agents at an increased dosing interval than licensed, thereby reducing NHS drug costs by decreasing the overall number of doses, as shown with simulations of Benepali in this thesis.

This thesis has provided proof-of-concept for popPK parameter estimation in a small cohort of patients commencing either Amgevita or Benepali in a real-world setting; simulations of altered dosing intervals have provided feasibility data for a future personalised dosing trial. A number of proteins identified using SWATH-MS have been found to be associated with various RA disease outcome measures, including therapeutic Amgevita or Benepali drug levels and DAS28 and its sub-components. Several proteins were also differentially

expressed over time by patients with a poor EULAR response, and between good/moderate and poor EULAR responders, providing candidate protein expression profiles that separate these time points or responder groups and which now require independent replication. Finally, pQTL analysis has identified loci that implicate a genetic basis behind expression of proteins that are associated with RA disease outcome measures, which could prove essential in future personalised streamlining of patients onto appropriate medication.

8.1. Proof-of-concept for personalised dosing: popPK studies and parameter estimation

Amgevita and Benepali, and most other biologics prescribed for the treatment of RA, are usually dispensed as pre-filled auto-injector syringes with a set dosage. Therefore, dose increases or reductions for these agents can only be effected by alteration of the dosing interval. An etanercept dose-finding study by Breedveld *et al* consisting of *post-hoc* analysis determined that doses of 10mg twice-weekly, 50mg every two weeks and 50mg weekly were the most effective to reach a steady-state drug concentration in the target range of 0.5 – 2 mg/L[388]. Target concentrations in this study were determined using PD markers in an E_{max} model of pooled trial patients, and are lower than those proposed by Jamnitski *et al* at 2.1 – 4.7 mg/L[220]. This may be because the levels proposed by the Jamnitsky study may be random, not trough, drug concentrations, as this was not overtly stated in the manuscript. Interestingly, findings from the Breedveld study broadly agreed with those in this thesis, in that doses lower than the currently recommended dosage of Benepali 50mg weekly were also found to be efficacious, and higher doses were of no further benefit to patients. However, while lower doses were shown to be as effective as the recommended dose, this was using data from a randomised clinical trial (RCT), where different doses of the drug were administered. In routine clinical practice, only 50mg pre-filled auto-injectors are available for patients to self-administer in England. Therefore, this thesis describes methods and results for an alternate-dosing strategy that involves alteration of the standard dosing interval of every seven days, based on popPK modelling data derived from real-world patients and not just patients who met eligibility criteria for an RCT.

The importance of carrying out pharmacological studies in patients recruited directly from the clinic room instead of into the controlled situation of a clinical trial should not be underestimated. In an RCT, strict inclusion criteria exist that often exclude more unwell patients with a higher disease burden, and the drug being studied is administered by a clinician at set times using a standardised protocol, which varies from which and how patients receive these medications in routine clinical practice. In this thesis, the samples and

data were collected from patients who were injecting themselves with medications, and drugs were not being administered by study clinicians. Drugs were not being administered at precise times (although precise dates and times were documented for the purpose of popPK modelling), and this is what would be expected in a patient's daily life. Drug administration also depended on each patient's individual technique, and this may also have influenced how much drug was absorbed; overall, therefore, the study design is more in-keeping with real-world scenarios and the findings should have greater generalisability than those from an RCT.

Despite the variable factors, enough drug concentration data was collected in order to derive two popPK models of patients starting TNFi biosimilars, one each for Amgevita and Benepali. This study has shown that even outside the very controlled environment of an early-phase clinical trial, it is still possible to obtain a level of meaningful and accurate information that allows the fitting of popPK models that reflect expected PK parameters, but also form the basis for simulations for future trial design. The fact that this has been achieved with only 10 and six patients in each popPK model, respectively, with only sparse sampling over 12 weeks, highlights the power of this modelling technique. With more patients recruited and more accurate PK parameter estimation, a personalised dosing trial based on these parameters could be a very realistic and feasible next step to this research. Alongside a personalised dosing trial, health economic research could be carried out to determine the cost-benefit of increased dosing intervals to the NHS, by comparing savings in drug cost and costs incurred from delayed RA disease control (such as increased healthcare encounters due to disease flares) against the cost of extra drug doses in those patients who require reduced dosing intervals (i.e. increased doses closer together) and extra drug level sampling so that personalised doses can be recommended. Only by demonstrating improved outcomes, acceptability to patients and cost-benefit can personalised dosing translate from a research trial to routine clinical practice.

8.2. Should we use drug concentration levels as a treatment outcome measure in RA?

Previous findings have shown that random adalimumab levels, but not random etanercept levels, were predictive of EULAR response at 12 months[11]. It could be postulated that therapeutic or even absolute drug levels would provide a clinician with a more objective RA clinical outcome measure than a composite score, like DAS28 or ACR20/50/70, which includes patient-reported outcome measures. Because drug levels are not influenced by patient reporting, they could be considered an objective clinical outcome measure.

Conversely, levels could be affected by a number of factors, such as operator inaccuracy when carrying out ELISAs, faulty testing kits, sample degradation or even poor drug adherence by patients. This latter variable is less of an issue, as drug non-adherence has been shown to be predictive of treatment non-response in patients receiving TNFi[64], so low drug levels in this case would be a proxy for non-adherence and would still be predictive of drug non-response.

Drug levels could also be low due to neutralising ADAbs, although these have not been detected for etanercept[389]. Immunogenicity to TNFi is a potential cause of both primary and secondary non-response. This study did not include ADAb levels, and this is a potential area for future development as serum samples are still available in the BRAGGSS-PD patients at all sampling time points. It could be interesting to ascertain whether ADAbs develop early, within the first 21 weeks of treatment, and if so, how early they develop and how much they impact drug levels and treatment response at 3 months. Another potential confounder for drug levels is body composition and/or BMI, but in popPK analysis, weight was not shown to significantly impact the model when included as a covariate.

There are arguments against using drug concentration levels as a clinical outcome measure, however. Jani *et al* found no association between random etanercept levels and EULAR response following adjustment for the confounders of age and biological sex, and were also unable to define a clear therapeutic window of etanercept levels that corresponded to treatment response[11]. Furthermore, studies defining therapeutic windows for etanercept vary widely in estimates, with Jamnitski *et al* proposing 2.1 – 4.7 mg/L[220] (although it is unclear whether these are trough or random levels), Chen *et al* proposing trough levels 1.24 mg/L at six months and 0.80 mg/L at 12 months[390], and Sanmarti *et al* proposing trough levels of 2.3 mg/L[391]. There could be biological reasons why drug levels may not accurately reflect RA disease activity as well. For example, drug concentrations may be "adequate," in which case, a patient would be regarded as a treatment "responder," but they might have ongoing inflammation and/or synovitis, as assessed with measurement of acute-phase reactants (e.g. CRP or ESR) or with imaging (e.g. ultrasound or MRI evidence of synovitis). In this case, using drug levels as a measure of disease activity would bias biomarker discovery.

Perhaps some of the most compelling evidence for using drug levels as an outcome measure are findings from this thesis. After adjustment for confounders, three proteins were found to be associated with drug levels: C4B was associated with therapeutic drug levels, A1AG1

was negatively associated with Amgevita levels (and positively associated with CRP) and IKKA was associated with Benepali levels. While these could be chance findings due to multiple comparisons in a small cohort of patients, the same proteins have also been shown to be associated with other RA disease outcome measures in a sub-cohort of BRAGGSS patients on etanercept, independent from the BRAGGSS-PD patients. Furthermore, these proteins also had associated pQTLs: C4B had 21 associated SNPs at baseline and 26 at three months, A1AG1 had one associated SNP at baseline and four SNPs were associated with IKKA at baseline. During SWATH-MS protein map extraction, C4B was included from the plasma library, and both A1AG1[293 294] and IKKA (participant in canonical TNF pathway) were included from the bespoke RA library.

The three proteins could be examined further as predictors of drug levels. C4B, alongside C4A, is reduced in patients with SLE. It has been shown that a high copy number of C4 genes is protective against SLE[392], and this thesis found that increased levels of C4B were associated with the achievement of therapeutic drug levels. Although SLE has a distinct disease identity from RA, there are common autoinflammatory components, and increased C4B levels alongside treatment response could implicate C4 as either an anti-inflammatory driver or marker in the setting of treatment of active RA with etanercept.

A1AG1 was associated with CRP, a known acute-phase reactant, and it is known to participate in the acute-phase response itself[393]. CRP is a non-specific acute-phase reactant, and can be raised in all inflammatory states, including acute infection; it is a good indicator of ongoing systemic inflammation, but may not be specific to RA. However, A1AG1 expression was shown to have a genetic basis in the sub-cohort of etanercept patients, which could implicate a disease process more specific to RA. As A1AG1 had a pQTL at baseline, but not three months, patients with increased expression could be more likely to fail on Amgevita. This requires wider validation and experimental studies to determine the mechanism, but A1AG1 is potentially an exciting biomarker predictive of treatment non-response. IKKA inhibits the pro-inflammatory canonical NF-κ-B signalling pathway[310], so similarly, a positive association with Benepali levels and a pQTL at baseline could reflect increased levels in patients who are likely to achieve treatment response on this drug.

Whether drug levels should be used as an objective RA disease outcome measure remains a research question; while there are arguments both for and against their use, this should be confirmed against known objective measures of RA disease activity, such as radiological

evidence of synovitis. However, radiological methods are time-consuming and can be costly (in the case of MRI), so large-scale studies have thus far not been undertaken.

8.3. Evidence of proteins associated with DAS28 and its sub-components – future biomarkers?

In addition to C4B, A1AG1 and IKKA, 11 other proteins were found to be associated with DAS28 and/or its components and also had associated pQTLs. This genetic basis for expression supports these proteins as candidate biomarkers, due to the known genetics underlying RA pathogenesis[394]. These proteins had a combination of both pro- and anti-inflammatory roles.

EHD1 has multiple roles in cell membrane reorganisation during ATP hydrolysis[315]; in this thesis, it was found to be positively associated with DAS28, worsening ΔDAS28 and patient-reported VAS of global health. Therefore, this would potentially implicate EHD1 as a pro-inflammatory biomarker. TCPH is also involved processes during ATP hydrolysis, namely, assistance with protein folding[316]; and this protein was also positively associated with DAS28 and SJC. Both EHD1 and TCPH could represent biomarkers of treatment non-response via their interactions with ATP hydrolysis.

A number of proteins were negatively associated with markers of increased RA disease activity. IGF1 is known to regulate osteoblast glycogen synthesis as well as stimulate glucose transport in osteoblasts[319], and was negatively associated with DAS28, SJC and poor EULAR response. IGF1 only had a pQTL after three months of treatment, so expression seems to be more a reflection of response to etanercept, as opposed to influencing the success of the drug. ILF3 is involved in a wide range of transcriptional and post-transcriptional processes[395] and plays an important anti-viral role[321]. In this thesis, it was negatively associated with SJC and poor EULAR response, with a pQTL at three months, but not baseline, so again, expression is likely to be in response to treatment with etanercept. NAA25 is thought to participate in normal cell-cycle progression[343] and was negatively associated with CRP in this thesis. The pQTL associated with this protein was identified pre-treatment, so NAA25 could represent a baseline biomarker predictive of treatment response to etanercept.

The remaining proteins with pQTLs that were associated with RA clinical outcome measures were all differentially expressed between good/moderate and poor EULAR responders at

either baseline or three months (XRCC6, CALM2, CALD1, CAMK1 and RSU), apart from BLVRB, which was differentially expressed between baseline and six months in poor EULAR responders. These proteins seem to form expression profiles differentiating different EULAR response groups, or the same group over time. A genetic basis behind some elements of these expression profiles implies that it may be possible to predict whether a patient is likely to respond to etanercept or not.

Notable by its absence is CRP: it frequently lost significance when adjusted in multivariable models and it had no pQTLs, implying that there was no genetic basis to its expression, despite its association with DAS28 and ΔDAS28. While these associations are reassuring in that they act as a positive control for the rest of the analyses, the negative findings for CRP confirm its role as a non-specific acute-phase reactant. Therefore, measurement of CRP levels pre-treatment or early in treatment are unlikely to provide clinicians with a useful marker of prognosis or future treatment response, and simply reflect ongoing disease activity. Increased levels probably reflect treatment inefficacy and continued increased disease activity, but this protein is unlikely to inform any treatment decisions, such as whether to commence or switch patients from particular therapeutic agents. An explanation for the lack of genetic basis of CRP expression in this study could be that while numerous studies have identified a genetic effect on basal CRP, CRP variability resulting from different levels of systemic inflammation within the study cohort could be large enough to mask the influence of underlying genetics[396]. This thesis has identified other protein candidate biomarkers that are potentially more specific and more predictive of treatment response than CRP, which now require independent validation; if replicated, the testing of these proteins could have implications for future clinical practice. The candidate biomarkers must also by confirmed against objective disease activity outcomes that are not based on CRP or clinical measurements, such as radiological evidence of synovitis.

8.4. Thesis strengths

The study populations included in this thesis are, undoubtedly, one of its strengths. Although smaller than expected, the BRAGGSS-PD cohort has already yielded exciting positive findings with respect to future personalised dosing trial design. While considered sparse sampling in popPK terms, the multiple serum samples obtained over time in this cohort have provided detailed protein maps that demonstrate changes in protein levels over time with treatment with Amgevita or Benepali. Sampling was more detailed than previous studies carried out in BRAGGSS, where samples were only available at baseline and three, six and

12 months. However, because of the additional ethics obtained for the popPK study, additional serum protein mapping was carried out on the extra samples taken between baseline and three months.

Collection of samples for the BRAGGSS-PD sub-study was robust and reliable. The author collected all samples and ensured that they were all delivered to the CfMR and processed within 24 hours of sampling. Drug administration and sampling times were all accurately recorded to the nearest minute. Because drug administration was witnessed by the author, it was also assured that drug concentration measurements were true trough levels, and not random drug levels.

Use of patients from the wider BRAGGSS cohort, both in the etanercept sub-cohort and in the PRS study, was also a strength to this thesis. Patients recruited to BRAGGSS are deeply phenotyped, and have a wide range of detailed treatment outcome data that can be analysed alongside biological samples.

A large cohort of BRAGGSS patients, separate from the etanercept sub-cohort, were also included as target data as part of the PRS modelling analyses. This cohort had fully QC-d genotype data on all participants, as well as additional covariate information for adjusted analysis. Use of these additional patients was also a strength because it allowed testing of significant pQTLs in a completely independent cohort.

Another strength in this work lies in the protein acquisition technique. SWATH-MS is an accurate, high-throughput DIA MS technique that can capture thousands of proteins in each run[92]. SWATH-MS compares favourably against other MS techniques[103] and has fewer missing values than other DDA techniques[101]. In addition, because SWATH-MS captures all proteins in a sample during each run, data can be re-interrogated with an alternative protein library *in silico* at a later date, as a permanent recording of the protein makeup of each sample is made.

8.5. Thesis limitations

Whilst use of BRAGGSS patients was a strength of this thesis, this also came with inherent limitations. Because BRAGGSS is a real-world clinical observational study, there is a large scope for missing data. Many patients have missing RA disease outcome data after baseline, and this is often because of delayed outpatients appointments due to health service pressures,

leading to, for example, overdue DAS28 measurements. In addition, many aspects of BRAGGSS data collection rely on patients filling in paper questionnaires (e.g. HADS, HAQ) and returning these via post to the study coordination team. This self-directed aspect to participation in BRAGGSS does leave it open to missing data.

The work in this thesis only focused on adalimumab and etanercept and their biosimilars, and these are only two of a selection of TNFi drugs. A range of bDMARDs with differing mechanisms alternative to TNFi are available for prescription to patients with uncontrolled RA, such as abatacept (T cell blockade), tocilizumab (IL-6 receptor inhibition) and rituximab (CD20 blockade). In addition, a new generation of small-molecule tsDMARDs have been approved for treatment of RA in the UK over the past five years, which work via inhibition of Janus kinases and interfering with the JAK-STAT signalling pathway. Findings in this thesis are likely specific to the TNFi drugs studied, and need to be assessed in other drugs which work via different biological pathways to determine whether they are transferable.

While findings of significant proteins replicated across independent BRAGGSS cohorts, there was no external data for validation outside of BRAGGSS available for analysis in this thesis. BRAGGSS is a UK-based multi-centre cohort, but treatment practices and the genetics of patients are different outside of the UK. Findings from this study require external validation, both from other UK-based cohorts, as well as in internationally-recruited patients.

There are also limitations to the acquisition of data using SWATH-MS and its subsequent pre-processing prior to analysis in this thesis. Missing protein values were assumed to be missing at random and were subsequently imputed; however, if values were truly missing due to biological reasons, then this might have been the incorrect approach and may have influenced results. Proteins included in analysis were pre-selected into a generic plasma library and a bespoke RA library, but both of these protein libraries may not have included proteins that affected treatment response and hence, were not matched from the libraries during data extraction. Future studies could include an enhanced RA protein library that builds on pathways from proteins identified in this thesis, that includes proteins known to have an effect on ADME of TNFi drugs and that includes proteins identified from further functional genetics studies. Another potential issue with the SWATH-MS data was that when a batch processing effect was discovered after SWATH-MS, this was statistically corrected. However, this may have suppressed any proteins with a strong signal in the equalisation process. An alternative approach here would have been to analyse all batches separately, and

then assess for replication. The decision was made to batch correct and analyse all samples together for the machine learning analysis in order to maximise power, but this may not have been the optimal approach. Future work could include repeating the machine learning analysis with separate batches, instead of grouping them together. Proteins detected using SWATH-MS could also be confirmed using a different technology, such as multiplexed detection techniques (e.g. Olink®[82], SomaScan®[81]) or ELISA.

As discussed earlier in this thesis, negative findings could have been due to the use of RA clinical outcome measures that do not adequately reflect active synovitis. The DAS28 (and EULAR response, which is based on DAS28) is a composite score that is not agnostic of non-inflammatory factors, such as a patient's perception of their health state outwith their RA (VAS of global health) or chronic widespread syndromes such as fibromyalgia (as reflected by increased TJC). According to a patient's body habitus, it can also sometimes be challenging to palpate for swollen joints. However, objective measures of disease activity such as ultrasound or MRI synovitis are not practical for rapid assessment in the clinic room, and so, although the DAS28 is imperfect, it is the current measure of choice for assessing RA disease activity in UK practice.

8.6. Clinical implications

The research carried out in this thesis has utilised real-world clinical patients recruited from rheumatology clinics across the UK. As such, findings are applicable to patients receiving routine clinical care and should be more relevant for from translation to day-to-day practice. The popPK study has provided a basis for future personalised dosing studies. Commencing patients on a longer dosing interval of a TNFi could have substantial cost-saving implications, while also reducing patient drug exposure and risk of adverse events, such as infection. Furthermore, commencing patients with more aggressive disease on a shorter dosing interval may ensure more rapid control of symptoms and halt the processes of joint destruction, and this also has cost-saving potential, from reduced burden on healthcare services (from reduced flares) as well as ensuring patients are more likely to respond to treatment.

A number of candidate biomarkers of treatment response and non-response have been identified during the course of this thesis. Currently, patients are commenced on biologics according to local prescribing protocols, often on the least expensive drugs first, then with escalation through gradually more expensive agents if initial therapy fails. If biomarkers of

treatment response to specific therapeutic agents can be determined during experimental studies, in the future, patients may be tested for these pre-treatment or very early on in treatment to determine whether they are likely to respond to a drug, or whether its prescription is an exercise in futility. Cycling of patients through multiple biologics imparts a significant cost burden on health services, as well as needlessly exposing patients to the risk of adverse events from a drug without a clear probability of success on any given agent. Stratified drug prescription, guided by biological markers such as a patient's underlying genetics or baseline protein expression profile, could ensure patients receive the correct agent at the first attempt.

8.7. Future work

There is a wide scope for future work. Firstly, additional patients could be recruited to both popPK studies (for Amgevita and Benepali) in order to gather more data points for detailed parameter estimation. An increased number of patients would also improve power and potentially allow introduction of covariates into models. Once PK parameters are accurately estimated with confidence, this could lead to a true personalised dosing study. Patients could have drug levels measured at specific points during a trial – these levels could then be used to determine exactly what their current PK profile is, in order to advise on the next dosing interval of their biologic. This form of active monitoring and feedback could lead to more rapid disease control, but also be cost-saving if dosing intervals can later be lengthened once a patient's RA disease activity is under control. Some patient involvement and engagement work that has been carried out in the CfMR has found that patients like the concept of personalised dosing, and further qualitative work with patient focus groups could be included as part of future personalised dosing study design to evaluate patient acceptability and design feasibility.

As previously mentioned, the machine learning analysis could be re-run in a number of ways i.e. without imputed missing proteins values, without batch correction, with all proteins obtained at SWATH-MS, and not filtered for only those proteins that were significantly differentially expressed in the case-control study. A variety of different machine learning algorithms that were not included in the analysis for this thesis could also be tested. The computational time required for all of these combinations was outwith the scope of this thesis, but additional analysis could potentially yield further interesting findings in addition to those included in this body of work.

Another previously mentioned analysis would involve measurement for ADAbs in the BRAGGSS-PD patients. It would be interesting to ascertain whether ADAbs were associated with drug levels and/or treatment response outcomes in these patients, as well as how early they might occur. Given the extensive sampling before the 12-week point in these patients, this would be an accurate method to assess how early in treatment patients might develop ADAbs.

Most importantly, significant proteins identified from this study require both external validation and experimental mechanistic confirmation of association. The most logical starting point would be to validate the 14 proteins that were associated with RA disease outcome measures that also had significant pQTLs and to carry out genetic studies at the SNPs identified. Because of the low number of these proteins, these could be measured in an independent cohort to BRAGGSS using ELISA or even a bespoke panel from a provider of multiplexed protein quantification. Mechanistic studies should be carried out on any protein(s) validated in an external cohort in order to determine how each protein interacts with RA drug therapy i.e. whether protein levels predict drug response, or whether they are a consequence of drug response. Studies could also be carried out in patients with RA receiving different therapeutic agents (as previously discussed) or in patients with different diseases receiving the same drugs, in order to check the specificity of findings to RA or the drugs studied in this thesis. This would then ascertain whether significant proteins are generally prognostic of a patient's RA disease course, that is, whether they will associate with response, regardless of treatment or even disease, or whether they are predictors of specific therapeutic agents. The latter would be preferable in the practice of precision medicine, as it would aid streamlining of patients onto a drug with a mechanism targeted to their biological profile.

Further downstream, if any significant proteins are validated as biomarkers of treatment response, these could then be developed for clinical assay. The ideal outcome would be a simple ELISA-based laboratory blood test, as for CRP, where clinicians can decide that a patient requires treatment escalation, and then a test can be ordered to determine which therapy to stratify that patient onto. Ideally, candidate biomarkers identified by this study would be able to predict treatment response to TNFi either before or very early on in treatment, to ensure that patients with RA are not commenced futilely on a drug that they are destined not to respond to. Something that this thesis has touched on throughout is that perhaps treatment response should be defined differently.

Currently, composite scores such as the DAS28 and CDAI/SDAI incorporate patient-reported measures of global health. Patient VAS of global health has previously been demonstrated to correlate strongly with illness cognitions and depression as measured using the HADS[397]. However, biological variables impact on biological processes, so an objective physiological measure of disease activity would be preferential in future clinical practice, as opposed to a composite score that incorporates patient-reported elements that can be biased by psychological factors. Biomarkers of treatment response that are found to be associated with synovitis and the systemic inflammation inherent with active RA could prove to be the objective disease outcome measure and/or predictor of treatment response that could transform current bDMARD prescribing practice. Another additional piece of research that could be carried out as a consequence of this thesis is to determine whether any proteins correlate with the 2C-DAS28, which omits TJC and VAS of global health and has been shown to correlate with ultrasound evidence of synovitis, as well as with radiographic evidence of joint erosions[52].

Mechanistic studies may also offer valuable insight and understanding into the function of significant proteins. This may lead to development of new drug targets, which could inform both *de novo* drug development, as well as repurposing of pre-existing therapeutic agents.

8.8. Final conclusions

This thesis sought to identify biological factors that might contribute to the variability in circulating TNFi drug levels and treatment response to biologic agents in patients with RA. The popPK studies did not identify any significant model covariates, but did provide proof-of-concept for future personalised dosing studies. The discovery proteomics study identified a number of proteins that were significantly associated with various RA clinical outcome measures. Perhaps the most promising of these proteins are C4B, A1AG1 and IKKA, which had associations in two independent cohorts as well as a genetic basis behind their expression, with associated pQTLs. Findings from this thesis require external validation with replication studies in an independent cohort, but once confirmed, this could pave the way for future biomarker and/or drug target development.

REFERENCES

1. Haraoui B, Pope J. Treatment of early rheumatoid arthritis: concepts in management. *Semin Arthritis Rheum* 2011;40(5):371-88. doi: 10.1016/j.semarthrit.2010.10.004
2. Ling S, Bluett J, Barton A. Prediction of response to methotrexate in rheumatoid arthritis. *Expert Rev Clin Immunol* 2018;14(5):419-29. doi: 10.1080/1744666X.2018.1465409
3. Smolen JS, Landewe R, Bijlsma J, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2016 update. *Ann Rheum Dis* 2017;76(6):960-77. doi: 10.1136/annrheumdis-2016-210715 [published Online First: 2017/03/08]
4. Singh JA, Saag KG, Bridges SL, Jr., et al. 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. *Arthritis Care Res (Hoboken)* 2016;68(1):1-25. doi: 10.1002/acr.22783 [published Online First: 2015/11/08]
5. Prescribing Costs in Hospitals and the Community, England 2017/18 [Available from: https://digital.nhs.uk/data-and-information/publications/statistical/prescribing-costs-in-hospitals-and-the-community/2017-18 accessed 19th November 2018.
6. NHS. NHS cuts medicines costs by three quarters of a billion pounds 2019 [Available from: https://www.england.nhs.uk/2019/08/nhs-cuts-medicines-costs-by-three-quarters-of-a-billion-pounds/ accessed 3rd June 2021.
7. Agency EM. Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: non-clinical and clinical issues 2014 [Available from: https://www.ema.europa.eu/documents/scientific-guideline/guideline-similar-biological-medicinal-products-containing-biotechnology-derived-proteins-active_en-2.pdf accessed 19th November 2018.
8. NICE. Adalimumab, etanercept, infliximab and abatacept for treating moderate rheumatoid arthritis after conventional DMARDs have failed 2021 [updated 14th July 2021. Available from: https://www.nice.org.uk/guidance/ta715 accessed 28th September 2021.
9. Weinblatt ME, Kremer JM, Bankhurst AD, et al. A trial of etanercept, a recombinant tumor necrosis factor receptor:Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate. *N Engl J Med* 1999;340(4):253-9. doi: 10.1056/NEJM199901283400401
10. Finckh A, Simard JF, Gabay C, et al. Evidence for differential acquired drug resistance to anti-tumour necrosis factor agents in rheumatoid arthritis. *Ann Rheum Dis* 2006;65(6):746-52. doi: 10.1136/ard.2005.045062
11. Jani M, Chinoy H, Warren RB, et al. Clinical utility of random anti-tumor necrosis factor drug-level testing and measurement of antidrug antibodies on the long-term treatment response in rheumatoid arthritis. *Arthritis Rheumatol* 2015;67(8):2011-9. doi: 10.1002/art.39169
12. Jani M, Dixon WG, Lunt M, et al. The association of biologic drug-levels with infection risk: results from the British Society for Rheumatology Biologics Register for Rheumatoid Arthritis. *Ann Rheum Dis* 2018;77:A163.
13. Smolen JS, Aletaha D, Barton A, et al. Rheumatoid arthritis. *Nat Rev Dis Primers* 2018;4:18001. doi: 10.1038/nrdp.2018.1 [published Online First: 2018/02/09]
14. Thurlings RM, Wijbrandts CA, Mebius RE, et al. Synovial lymphoid neogenesis does not define a specific clinical rheumatoid arthritis phenotype. *Arthritis Rheum* 2008;58(6):1582-9. doi: 10.1002/art.23505 [published Online First: 2008/06/03]
15. van Baarsen LG, Wijbrandts CA, Timmer TC, et al. Synovial tissue heterogeneity in rheumatoid arthritis in relation to disease activity and biomarkers in peripheral blood. *Arthritis Rheum* 2010;62(6):1602-7. doi: 10.1002/art.27415 [published Online First: 2010/02/24]
16. Kasperkovitz PV, Timmer TC, Smeets TJ, et al. Fibroblast-like synoviocytes derived from patients with rheumatoid arthritis show the imprint of synovial tissue heterogeneity: evidence of a link between an increased myofibroblast-like phenotype and high-inflammation synovitis. *Arthritis Rheum* 2005;52(2):430-41. doi: 10.1002/art.20811 [published Online First: 2005/02/05]

17. Karsdal MA, Bay-Jensen AC, Henriksen K, et al. Rheumatoid arthritis: a case for personalized health care? *Arthritis Care Res (Hoboken)* 2014;66(9):1273-80. doi: 10.1002/acr.22289 [published Online First: 2014/01/29]
18. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum* 1988;31(3):315-24. [published Online First: 1988/03/01]
19. Funovits J, Aletaha D, Bykerk V, et al. The 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for rheumatoid arthritis: methodological report phase I. *Ann Rheum Dis* 2010;69(9):1589-95. doi: 10.1136/ard.2010.130310
20. Smolen JS, Breedveld FC, Eberl G, et al. Validity and reliability of the twenty-eight-joint count for the assessment of rheumatoid arthritis activity. *Arthritis Rheum* 1995;38(1):38-43. [published Online First: 1995/01/01]
21. John H, Kitas G, Toms T, et al. Cardiovascular co-morbidity in early rheumatoid arthritis. *Best Pract Res Clin Rheumatol* 2009;23(1):71-82. doi: 10.1016/j.berh.2008.11.007 [published Online First: 2009/02/24]
22. Koduri G, Norton S, Young A, et al. Interstitial lung disease has a poor prognosis in rheumatoid arthritis: results from an inception cohort. *Rheumatology (Oxford)* 2010;49(8):1483-9. doi: 10.1093/rheumatology/keq035 [published Online First: 2010/03/13]
23. Cojocaru M, Cojocaru IM, Silosi I, et al. Extra-articular manifestations in rheumatoid arthritis. *Maedica (Bucharest)* 2010;5(4):286-91.
24. Turesson C, W.M. OF, Crowson CS, et al. Extra-articular manifestations in rheumatoid arthritis: incidence trends and risk factors over 46 years. *Ann Rheum Dis* 2003;62(8):722-27.
25. Hochberg MC, Johnston SS, John AK. The incidence and prevalence of extra-articular and systemic manifestations in a cohort of newly-diagnosed patients with rheumatoid arthritis between 1999 and 2006. *Curr Med Res Opin* 2008;24(2):469-80.
26. Minichiello E, Semerano L, Boissier MC. Time trends in the incidence, prevalence, and severity of rheumatoid arthritis: A systematic literature review. *Joint Bone Spine* 2016;83(6):625-30. doi: 10.1016/j.jbspin.2016.07.007 [published Online First: 2016/09/13]
27. McInnes IB, Schett G. The pathogenesis of rheumatoid arthritis. *N Engl J Med* 2011;365(23):2205-19. doi: 10.1056/NEJMra1004965 [published Online First: 2011/12/14]
28. Combe B, Landewe R, Daien CI, et al. 2016 update of the EULAR recommendations for the management of early arthritis. *Ann Rheum Dis* 2017;76(6):948-59. doi: 10.1136/annrheumdis-2016-210602 [published Online First: 2016/12/17]
29. NICE. Rheumatoid arthritis in adults: management 2018 [Available from: https://www.nice.org.uk/guidance/ng100/chapter/Recommendations accessed 30 January 2020.
30. van Zanten A, Arends S, Roozendaal C, et al. Presence of anticitrullinated protein antibodies in a large population-based cohort from the Netherlands. *Ann Rheum Dis* 2017;76(7):1184-90. doi: 10.1136/annrheumdis-2016-209991 [published Online First: 2017/01/04]
31. Ingegnoli F, Castelli R, Gualtierotti R. Rheumatoid factors: clinical applications. *Dis Markers* 2013;35(6):727-34. doi: 10.1155/2013/726598 [published Online First: 2013/12/11]
32. Forslind K, Ahlmen M, Eberhardt K, et al. Prediction of radiological outcome in early rheumatoid arthritis in clinical practice: role of antibodies to citrullinated peptides (anti-CCP). *Ann Rheum Dis* 2004;63(9):1090-5. doi: 10.1136/ard.2003.014233
33. Farragher TM, Lunt M, Plant D, et al. Benefit of early treatment in inflammatory polyarthritis patients with anti-cyclic citrullinated peptide antibodies versus those without antibodies. *Arthritis Care Res (Hoboken)* 2010;62(5):664-75. doi: 10.1002/acr.20207 [published Online First: 2010/05/13]
34. Potter C, Hyrich KL, Tracey A, et al. Association of rheumatoid factor and anti-cyclic citrullinated peptide positivity, but not carriage of shared epitope or PTPN22 susceptibility variants, with anti-tumour necrosis factor response in rheumatoid arthritis. *Ann Rheum Dis* 2009;68(1):69-74. doi: 10.1136/ard.2007.084715 [published Online First: 2008/04/01]
35. Ling SF, Nair N, Verstappen SMM, et al. Proteomic analysis to define predictors of treatment response to adalimumab or methotrexate in rheumatoid arthritis patients.

Pharmacogenomics J 2019 doi: 10.1038/s41397-019-0139-4 [published Online First: 2019/12/11]
36. Gwinnutt J, Hyrich K, Lunt M, et al. The association between joint erosions plus autoantibody positivity at initiation of methotrexate or biologic therapy for rheumatoid arthritis and disease activity and disability over one year. *Ann Rheum Dis* 2019;78:A1035.
37. Shi J, Knevel R, Suwannalai P, et al. Autoantibodies recognizing carbamylated proteins are present in sera of patients with rheumatoid arthritis and predict joint damage. *Proc Natl Acad Sci U S A* 2011;108(42):17372-7. doi: 10.1073/pnas.1114465108
38. van Gaalen FA, Visser H, Huizinga TW. A comparison of the diagnostic accuracy and prognostic value of the first and second anti-cyclic citrullinated peptides (CCP1 and CCP2) autoantibody tests for rheumatoid arthritis. *Ann Rheum Dis* 2005;64(10):1510-2. doi: 10.1136/ard.2004.035089 [published Online First: 2005/04/01]
39. van der Heijde DM, van 't Hof MA, van Riel PL, et al. Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. *Ann Rheum Dis* 1990;49(11):916-20.
40. Kirwan JR, Chaput de Saintonge DM, Joyce CR, et al. Clinical judgment in rheumatoid arthritis. I. Rheumatologists' opinions and the development of 'paper patients'. *Ann Rheum Dis* 1983;42(6):644-7.
41. Ritchie DM, Boyle JA, McInnes JM, et al. Clinical studies with an articular index for the assessment of joint tenderness in patients with rheumatoid arthritis. *Q J Med* 1968;37(147):393-406.
42. Prevoo ML, van 't Hof MA, Kuper HH, et al. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis Rheum* 1995;38(1):44-8.
43. Fransen J, Welsing PM, de Keijzer RM, et al. Disease activity scores using C-reactive protein: CRP may replace ESR in the assessment of RA disease activity [abstract]. *Ann Rheum Dis* 2003;62(Suppl 1):151.
44. Hensor EM, Emery P, Bingham SJ, et al. Discrepancies in categorizing rheumatoid arthritis patients by DAS-28(ESR) and DAS-28(CRP): can they be reduced? *Rheumatology (Oxford)* 2010;49(8):1521-9. doi: 10.1093/rheumatology/keq117
45. Siemons L, Vonkeman HE, ten Klooster PM, et al. Interchangeability of 28-joint disease activity scores using the erythrocyte sedimentation rate or the C-reactive protein as inflammatory marker. *Clin Rheumatol* 2014;33(6):783-9. doi: 10.1007/s10067-014-2538-x
46. Madsen OR. Is DAS28-CRP with three and four variables interchangeable in individual patients selected for biological treatment in daily clinical practice? *Clin Rheumatol* 2011;30(12):1577-82. doi: 10.1007/s10067-011-1847-6
47. Fransen J, Creemers MC, Van Riel PL. Remission in rheumatoid arthritis: agreement of the disease activity score (DAS28) with the ARA preliminary remission criteria. *Rheumatology (Oxford)* 2004;43(10):1252-5. doi: 10.1093/rheumatology/keh297
48. Cohen G, Gossec L, Dougados M, et al. Radiological damage in patients with rheumatoid arthritis on sustained remission. *Ann Rheum Dis* 2007;66(3):358-63. doi: 10.1136/ard.2006.057497
49. Viatte S, Plant D, Han B, et al. Association of HLA-DRB1 haplotypes with rheumatoid arthritis severity, mortality, and treatment response. *JAMA* 2015;313(16):1645-56. doi: 10.1001/jama.2015.3435
50. Ling SF, Viatte S, Lunt M, et al. HLA-DRB1 Amino Acid Positions 11/13, 71, and 74 Are Associated With Inflammation Level, Disease Activity, and the Health Assessment Questionnaire Score in Patients With Inflammatory Polyarthritis. *Arthritis Rheumatol* 2016;68(11):2618-28. doi: 10.1002/art.39780
51. Baker JF, Conaghan PG, Smolen JS, et al. Development and validation of modified disease activity scores in rheumatoid arthritis: superior correlation with magnetic resonance imaging-detected synovitis and radiographic progression. *Arthritis Rheumatol* 2014;66(4):794-802. doi: 10.1002/art.38304

52. Hensor EMA, McKeigue P, Buch MH, et al. Validity of a 2-component imaging-derived disease activity score (2C-DAS28) for improved assessment of synovitis in early rheumatoid arthritis [abstract]. *Rheumatology (Oxford)* 2018;57(Suppl 3):key074.194.
53. Smolen JS, Breedveld FC, Schiff MH, et al. A simplified disease activity index for rheumatoid arthritis for use in clinical practice. *Rheumatology (Oxford)* 2003;42(2):244-57.
54. Aletaha D, Nell VP, Stamm T, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. *Arthritis Res Ther* 2005;7(4):R796-806. doi: 10.1186/ar1740
55. van Gestel AM, Haagsma CJ, van Riel PL. Validation of rheumatoid arthritis improvement criteria that include simplified joint counts. *Arthritis Rheum* 1998;41(10):1845-50. doi: 10.1002/1529-0131(199810)41:10<1845::AID-ART17>3.0.CO;2-K
56. van Gestel AM, Prevoo ML, van 't Hof MA, et al. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis. Comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism Criteria. *Arthritis Rheum* 1996;39(1):34-40.
57. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. *Arthritis Rheum* 1995;38(6):727-35.
58. Fransen J, van Riel PL. The Disease Activity Score and the EULAR response criteria. *Clin Exp Rheumatol* 2005;23(5 Suppl 39):S93-9.
59. Sugiyama D, Nishimura K, Tamaki K, et al. Impact of smoking as a risk factor for developing rheumatoid arthritis: a meta-analysis of observational studies. *Ann Rheum Dis* 2010;69(1):70-81. doi: 10.1136/ard.2008.096487 [published Online First: 2009/01/29]
60. Soderlin MK, Petersson IF, Geborek P. The effect of smoking on response and drug survival in rheumatoid arthritis patients treated with their first anti-TNF drug. *Scand J Rheumatol* 2012;41(1):1-9. doi: 10.3109/03009742.2011.599073 [published Online First: 2011/11/29]
61. Kearsley-Fleet L, Davies R, De Cock D, et al. Biologic refractory disease in rheumatoid arthritis: results from the British Society for Rheumatology Biologics Register for Rheumatoid Arthritis. *Ann Rheum Dis* 2018;77(10):1405-12. doi: 10.1136/annrheumdis-2018-213378 [published Online First: 2018/07/08]
62. Matcham F, Davies R, Hotopf M, et al. The relationship between depression and biologic treatment response in rheumatoid arthritis: An analysis of the British Society for Rheumatology Biologics Register. *Rheumatology (Oxford)* 2018;57(5):835-43. doi: 10.1093/rheumatology/kex528 [published Online First: 2018/02/16]
63. Organisation WH. Adherence to long-term therapies: evidence for action 2003 [Available from: https://www.who.int/chp/knowledge/publications/adherence_full_report.pdf?ua=1 accessed 29th July 2021.
64. Bluett J, Morgan C, Thurston L, et al. Impact of inadequate adherence on response to subcutaneously administered anti-tumour necrosis factor drugs: results from the Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate cohort. *Rheumatology (Oxford)* 2015;54(3):494-9. doi: 10.1093/rheumatology/keu358 [published Online First: 2014/09/13]
65. Hope HF, Bluett J, Barton A, et al. Psychological factors predict adherence to methotrexate in rheumatoid arthritis; findings from a systematic review of rates, predictors and associations with patient-reported and clinical outcomes. *RMD Open* 2016;2(1):e000171. doi: 10.1136/rmdopen-2015-000171 [published Online First: 2016/02/06]
66. Jani M, Isaacs JD, Morgan AW, et al. High frequency of antidrug antibodies and association of random drug levels with efficacy in certolizumab pegol-treated patients with rheumatoid arthritis: results from the BRAGGSS cohort. *Ann Rheum Dis* 2017;76(1):208-13. doi: 10.1136/annrheumdis-2015-208849 [published Online First: 2016/06/02]
67. Sergeant JC, Hyrich KL, Anderson J, et al. Prediction of primary non-response to methotrexate therapy using demographic, clinical and psychosocial variables: results from the UK Rheumatoid Arthritis Medication Study (RAMS). *Arthritis Res Ther* 2018;20(1):147. doi: 10.1186/s13075-018-1645-5 [published Online First: 2018/07/15]

68. Vordenbaumen S, Lueking A, Budde P, et al. Sequential high-content profiling of the IgG-autoantibody repertoire reveals novel antigens in rheumatoid arthritis. *Arthritis Res Ther* 2016;18(1):235. doi: 10.1186/s13075-016-1135-6 [published Online First: 2016/10/13]
69. Zhu Y, Wang X, Forouzmand E, et al. Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. *Mol Cell* 2018;69(1):62-74 e4. doi: 10.1016/j.molcel.2017.11.031 [published Online First: 2017/12/26]
70. Shi Y, Mosser DD, Morimoto RI. Molecular chaperones as HSF1-specific transcriptional repressors. *Genes Dev* 1998;12(5):654-66. doi: 10.1101/gad.12.5.654 [published Online First: 1998/04/16]
71. James P. Protein identification in the post-genome era: the rapid rise of proteomics. *Q Rev Biophys* 1997;30(4):279-331.
72. Zhang Y, Fonslow BR, Shan B, et al. Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* 2013;113(4):2343-94. doi: 10.1021/cr3003533
73. Doyle HA, Mamula MJ. Posttranslational modifications of self-antigens. *Ann N Y Acad Sci* 2005;1050:1-9. doi: 10.1196/annals.1313.001
74. Mastrangelo A, Colasanti T, Barbati C, et al. The Role of Posttranslational Protein Modifications in Rheumatological Diseases: Focus on Rheumatoid Arthritis. *J Immunol Res* 2015;2015:712490. doi: 10.1155/2015/712490
75. Hueber W, Kidd BA, Tomooka BH, et al. Antigen microarray profiling of autoantibodies in rheumatoid arthritis. *Arthritis Rheum* 2005;52(9):2645-55. doi: 10.1002/art.21269
76. Aletaha D, Neogi T, Silman AJ, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Ann Rheum Dis* 2010;69(9):1580-8. doi: 10.1136/ard.2010.138461
77. Han X, Jin M, Breuker K, et al. Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science* 2006;314(5796):109-12. doi: 10.1126/science.1128868
78. Tran JC, Zamdborg L, Ahlf DR, et al. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 2011;480(7376):254-8. doi: 10.1038/nature10575
79. SomaLogic. SomaScan(R) Assay v4.1 2021 [updated 2021. Available from: https://cdn2.hubspot.net/hubfs/6686502/SOMAscanWhitePaper.pdf?__hsfp=939966733&__hssc=251652889.3.1625749146734&__hstc=251652889.a552207981861e17c225efb586e12c52.1620839172483.1625681656017.1625749146734.28 accessed 9th August 2021.
80. Gold L, Ayers D, Bertino J, et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* 2010;5(12):e15004. doi: 10.1371/journal.pone.0015004 [published Online First: 2010/12/18]
81. SomaLogic. SomaLogic website [Available from: www.somalogic.com accessed 9th August 2021.
82. Olink. Olink website [Available from: www.olink.com accessed 9th August 2021.
83. Assarsson E, Lundberg M, Holmquist G, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS One* 2014;9(4):e95192. doi: 10.1371/journal.pone.0095192 [published Online First: 2014/04/24]
84. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. *Nature* 2018;558(7708):73-79. doi: 10.1038/s41586-018-0175-2 [published Online First: 2018/06/08]
85. Yates JR, 3rd. Mass spectrometry and the age of the proteome. *J Mass Spectrom* 1998;33(1):1-19. doi: 10.1002/(SICI)1096-9888(199801)33:1<1::AID-JMS624>3.0.CO;2-9
86. IUPAC. Compendium of Chemical Terminology. In: McNaught AD, Wilkinson A, eds. The "Gold Book". 2nd ed. Online version (2019-) created by S.J. Chalk, 1997.
87. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* 2005;4(10):1419-40. doi: 10.1074/mcp.R500012-MCP200
88. Yates JR, 3rd, Washburn MP. Quantitative proteomics. *Anal Chem* 2013;85(19):8881. doi: 10.1021/ac402745w
89. Picotti P, Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods* 2012;9(6):555-66. doi: 10.1038/nmeth.2015

90. Anderson L, Hunter CL. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol Cell Proteomics* 2006;5(4):573-88. doi: 10.1074/mcp.M500331-MCP200 [published Online First: 2005/12/08]
91. Addona TA, Abbatiello SE, Schilling B, et al. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat Biotechnol* 2009;27(7):633-41. doi: 10.1038/nbt.1546
92. Gillet LC, Navarro P, Tate S, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* 2012;11(6):O111 016717. doi: 10.1074/mcp.O111.016717
93. Frederick K, Ciborowski P. SWATH-MS: data acquisition and analysis. In: Ciborowski P, Silberring J, eds. Proteomic Profiling and Analytical Chemistry: Elsevier 2016.
94. Schubert OT, Gillet LC, Collins BC, et al. Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat Protoc* 2015;10(3):426-41. doi: 10.1038/nprot.2015.015 [published Online First: 2015/02/13]
95. Li S, Arnold RJ, Tang H, et al. On the accuracy and limits of peptide fragmentation spectrum prediction. *Anal Chem* 2011;83(3):790-6. doi: 10.1021/ac102272r [published Online First: 2010/12/24]
96. SWATHAtlas [Available from: http://www.swathatlas.org accessed 7th October 2020.
97. Huang Q, Yang L, Luo J, et al. SWATH enables precise label-free quantification on proteome scale. *Proteomics* 2015;15(7):1215-23. doi: 10.1002/pmic.201400270
98. Harrison OJ, Cagampang F, Ohri SK, et al. Candidate plasma biomarkers for predicting ascending aortic aneurysm in bicuspid aortic valve disease. *J Cardiothorac Surg* 2018;13(1):76. doi: 10.1186/s13019-018-0762-1
99. Sims MC, Mayer L, Collins J, et al. Novel manifestations of immune dysregulation and granule defects in gray platelet syndrome. *Blood* 2020 doi: 10.1182/blood.2019004776 [published Online First: 2020/07/22]
100. Albrecht D, Kniemeyer O, Brakhage AA, et al. Missing values in gel-based proteomics. *Proteomics* 2010;10(6):1202-11. doi: 10.1002/pmic.200800576 [published Online First: 2010/01/16]
101. McGurk KA, Dagliati A, Chiasserini D, et al. The use of missing values in proteomic data-independent acquisition mass spectrometry to enable disease activity discrimination. *Bioinformatics* 2020;36(7):2217-23. doi: 10.1093/bioinformatics/btz898 [published Online First: 2019/12/04]
102. Lazar C, Gatto L, Ferro M, et al. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J Proteome Res* 2016;15(4):1116-25. doi: 10.1021/acs.jproteome.5b00981 [published Online First: 2016/02/26]
103. Shao S, Guo T, Aebersold R. Mass spectrometry-based proteomic quest for diabetes biomarkers. *Biochim Biophys Acta* 2015;1854(6):519-27. doi: 10.1016/j.bbapap.2014.12.012
104. Wu C, Tran JC, Zamdborg L, et al. A protease for 'middle-down' proteomics. *Nat Methods* 2012;9(8):822-4. doi: 10.1038/nmeth.2074
105. Cristobal A, Marino F, Post H, et al. Toward an Optimized Workflow for Middle-Down Proteomics. *Anal Chem* 2017;89(6):3318-25. doi: 10.1021/acs.analchem.6b03756
106. Sadygov RG, Cociorva D, Yates JR, 3rd. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* 2004;1(3):195-202. doi: 10.1038/nmeth725
107. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;5(11):976-89. doi: 10.1016/1044-0305(94)80016-2
108. Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 1994;66(24):4390-9.
109. Bafna V, Edwards N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* 2001;17 Suppl 1:S13-21.

110. Perkins DN, Pappin DJ, Creasy DM, et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20(18):3551-67. doi: 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2
111. Bern M, Goldberg D, McDonald WH, et al. Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics* 2004;20 Suppl 1:i49-54. doi: 10.1093/bioinformatics/bth947
112. Tabb DL, MacCoss MJ, Wu CC, et al. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal Chem* 2003;75(10):2470-7.
113. Scherl A, Francois P, Converset V, et al. Nonredundant mass spectrometry: a strategy to integrate mass spectrometry acquisition and analysis. *Proteomics* 2004;4(4):917-27. doi: 10.1002/pmic.200300673
114. Nesvizhskii AI, Aebersold R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov Today* 2004;9(4):173-81. doi: 10.1016/S1359-6446(03)02978-7
115. Sadygov RG, Liu H, Yates JR. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal Chem* 2004;76(6):1664-71. doi: 10.1021/ac035112y
116. Tabb DL, McDonald WH, Yates JR, 3rd. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* 2002;1(1):21-6.
117. Kislinger T, Rahman K, Radulovic D, et al. PRISM, a generic large scale proteomic investigation strategy for mammals. *Mol Cell Proteomics* 2003;2(2):96-106. doi: 10.1074/mcp.M200074-MCP200
118. Ma ZQ, Dasari S, Chambers MC, et al. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* 2009;8(8):3872-81. doi: 10.1021/pr900360j
119. Nesvizhskii AI, Keller A, Kolker E, et al. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 2003;75(17):4646-58.
120. Serang O, MacCoss MJ, Noble WS. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J Proteome Res* 2010;9(10):5346-57. doi: 10.1021/pr100594k
121. Ting YS, Egertson JD, Payne SH, et al. Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Mol Cell Proteomics* 2015;14(9):2301-7. doi: 10.1074/mcp.O114.047035 [published Online First: 2015/07/29]
122. Navarro P, Kuharev J, Gillet LC, et al. A multicenter study benchmarks software tools for label-free proteome quantification. *Nat Biotechnol* 2016;34(11):1130-36. doi: 10.1038/nbt.3685 [published Online First: 2016/11/01]
123. Tsou CC, Avtonomov D, Larsen B, et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* 2015;12(3):258-64, 7 p following 64. doi: 10.1038/nmeth.3255 [published Online First: 2015/01/20]
124. NICE. Infliximab and adalimumab for the treatment of Crohn's disease 2010 [Available from: https://www.nice.org.uk/guidance/ta187 accessed 18th December 2018.
125. NICE. 2015 [Available from: https://www.nice.org.uk/guidance/ta329 accessed 18th December 2018.
126. Sekigawa I, Yanagida M, Iwabuchi K, et al. Protein biomarker analysis by mass spectrometry in patients with rheumatoid arthritis receiving anti-tumor necrosis factor-alpha antibody therapy. *Clin Exp Rheumatol* 2008;26(2):261-7.
127. Serada S, Fujimoto M, Ogata A, et al. iTRAQ-based proteomic identification of leucine-rich alpha-2 glycoprotein as a novel inflammatory biomarker in autoimmune diseases. *Ann Rheum Dis* 2010;69(4):770-4. doi: 10.1136/ard.2009.118919
128. Trocme C, Marotte H, Baillet A, et al. Apolipoprotein A-I and platelet factor 4 are biomarkers for infliximab response in rheumatoid arthritis. *Ann Rheum Dis* 2009;68(8):1328-33. doi: 10.1136/ard.2008.093153

129. Fabre S, Dupuy AM, Dossat N, et al. Protein biochip array technology for cytokine profiling predicts etanercept responsiveness in rheumatoid arthritis. *Clin Exp Immunol* 2008;153(2):188-95. doi: 10.1111/j.1365-2249.2008.03691.x
130. Hueber W, Tomooka BH, Batliwalla F, et al. Blood autoantibody and cytokine profiles predict response to anti-tumor necrosis factor therapy in rheumatoid arthritis. *Arthritis Res Ther* 2009;11(3):R76. doi: 10.1186/ar2706
131. Obry A, Hardouin J, Lequerre T, et al. Identification of 7 Proteins in Sera of RA Patients with Potential to Predict ETA/MTX Treatment Response. *Theranostics* 2015;5(11):1214-24. doi: 10.7150/thno.12403
132. Fabre S, Guisset C, Tatem L, et al. Protein biochip array technology to monitor rituximab in rheumatoid arthritis. *Clin Exp Immunol* 2009;155(3):395-402. doi: 10.1111/j.1365-2249.2008.03804.x
133. Nguyen MVC, Courtier A, Adrait A, et al. Fetuin-A and thyroxin binding globulin predict rituximab response in rheumatoid arthritis patients with insufficient response to anti-TNFalpha. *Clin Rheumatol* 2020;39(9):2553-62. doi: 10.1007/s10067-020-05030-6 [published Online First: 2020/03/27]
134. Murota A, Suzuki K, Kassai Y, et al. Serum proteomic analysis identifies interleukin 16 as a biomarker for clinical response during early treatment of rheumatoid arthritis. *Cytokine* 2016;78:87-93. doi: 10.1016/j.cyto.2015.12.002
135. Bartok B, Firestein GS. Fibroblast-like synoviocytes: key effector cells in rheumatoid arthritis. *Immunol Rev* 2010;233(1):233-55. doi: 10.1111/j.0105-2896.2009.00859.x [published Online First: 2010/03/03]
136. Teitsma XM, Jacobs JWG, Concepcion AN, et al. Explorative analyses of protein biomarkers in patients with early rheumatoid arthritis achieving sustained drug-free remission after treatment with tocilizumab- or methotrexate-based strategies: from transcriptomics to proteomics. *Clin Exp Rheumatol* 2018;36(6):976-83. [published Online First: 2018/05/11]
137. Cuppen B, Fritsch-Stork R, Eekhout I, et al. Proteomics to predict the response to tumour necrosis factor-alpha inhibitors in rheumatoid arthritis using a supervised cluster-analysis based protein score. *Scand J Rheumatol* 2018;47(1):12-21. doi: 10.1080/03009742.2017.1309061
138. Tasaki S, Suzuki K, Kassai Y, et al. Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. *Nat Commun* 2018;9(1):2755. doi: 10.1038/s41467-018-05044-4 [published Online First: 2018/07/18]
139. Farutin V, Prod'homme T, McConnell K, et al. Molecular profiling of rheumatoid arthritis patients reveals an association between innate and adaptive cell populations and response to anti-tumor necrosis factor. *Arthritis Res Ther* 2019;21(1):216. doi: 10.1186/s13075-019-1999-3 [published Online First: 2019/10/28]
140. Mellors T, Withers JB, Ameli A, et al. Clinical validation of a blood-based predictive test for stratification of response to tumor necrosis factor inhibitor therapies in rheumatoid arthritis patients. *Network and Systems Medicine* 2020;3(1):91-104. doi: http://doi.org/10.1089/nsm.2020.0007 [published Online First: 14 Jul 2020]
141. Cohen S, Wells AF, Curtis JR, et al. A Molecular Signature Response Classifier to Predict Inadequate Response to Tumor Necrosis Factor-alpha Inhibitors: The NETWORK-004 Prospective Observational Study. *Rheumatol Ther* 2021 doi: 10.1007/s40744-021-00330-y [published Online First: 2021/06/21]
142. Tao W, Concepcion AN, Vianen M, et al. Multiomics and Machine Learning Accurately Predict Clinical Response to Adalimumab and Etanercept Therapy in Patients With Rheumatoid Arthritis. *Arthritis Rheumatol* 2021;73(2):212-22. doi: 10.1002/art.41516 [published Online First: 2020/09/11]
143. Luque-Tevar M, Perez-Sanchez C, Patino-Trives AM, et al. Integrative Clinical, Molecular, and Computational Analysis Identify Novel Biomarkers and Differential Profiles of Anti-TNF Response in Rheumatoid Arthritis. *Front Immunol* 2021;12:631662. doi: 10.3389/fimmu.2021.631662 [published Online First: 2021/04/10]

144. Schwarz UI, Ritchie MD, Bradford Y, et al. Genetic determinants of response to warfarin during initial anticoagulation. *N Engl J Med* 2008;358(10):999-1008. doi: 10.1056/NEJMoa0708078 [published Online First: 2008/03/07]
145. Upchurch KS, Kay J. Evolution of treatment for rheumatoid arthritis. *Rheumatology (Oxford)* 2012;51 Suppl 6:vi28-36. doi: 10.1093/rheumatology/kes278 [published Online First: 2012/12/19]
146. NICE. Adalimumab, etanercept, infliximab, certolizumab pegol, golimumab, tocilizumab and abatacept for rheumatoid arthritis not previously treated with DMARDs or after conventional DMARDs only have failed 2016 [Available from: https://www.nice.org.uk/guidance/ta375 accessed 28th September 2021.
147. NICE. Tofacinitinib for moderate to severe rheumatoid arthritis 2017 [Available from: https://www.nice.org.uk/guidance/ta480 accessed 28th September 2021.
148. NICE. Baricitinib for moderate to severe rheumatoid arthritis 2017 [Available from: https://www.nice.org.uk/guidance/ta466 accessed 28th September 2021.
149. NICE. Filgotinib for treating moderate to severe rheumatoid arthritis 2021 [Available from: https://www.nice.org.uk/guidance/ta676 accessed 28th September 2021.
150. Cohen M, Omair MA, Keystone EC. Monoclonal antibodies in rheumatoid arthritis. *Int J Clin Rheumatol* 2013;8:541-56.
151. Ternant D, Bejan-Angoulvant T, Passot C, et al. Clinical Pharmacokinetics and Pharmacodynamics of Monoclonal Antibodies Approved to Treat Rheumatoid Arthritis. *Clin Pharmacokinet* 2015;54(11):1107-23. doi: 10.1007/s40262-015-0296-9 [published Online First: 2015/07/01]
152. Junghans RP. Finally! The Brambell receptor (FcRB). Mediator of transmission of immunity and protection from catabolism for IgG. *Immunol Res* 1997;16(1):29-57. doi: 10.1007/BF02786322 [published Online First: 1997/02/01]
153. Winstanley P, Walley T. Medical Pharmacology. Second edition ed: Churchill Livingstone 2002.
154. Tabrizi M, Bornstein GG, Suria H. Biodistribution mechanisms of therapeutic monoclonal antibodies in health and disease. *AAPS J* 2010;12(1):33-43. doi: 10.1208/s12248-009-9157-5 [published Online First: 2009/11/20]
155. Dirks NL, Meibohm B. Population pharmacokinetics of therapeutic monoclonal antibodies. *Clin Pharmacokinet* 2010;49(10):633-59. doi: 10.2165/11535960-000000000-00000 [published Online First: 2010/09/08]
156. FDA. Cimzia (certolizumab pegol) label information 2008 [Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2008/125160s000lbl.pdf accessed 3rd December 2020.
157. Hemperly A, Vande Casteele N. Clinical Pharmacokinetics and Pharmacodynamics of Infliximab in the Treatment of Inflammatory Bowel Disease. *Clin Pharmacokinet* 2018;57(8):929-42. doi: 10.1007/s40262-017-0627-0 [published Online First: 2018/01/14]
158. Xu Z, Vu T, Lee H, et al. Population pharmacokinetics of golimumab, an anti-tumor necrosis factor-alpha human monoclonal antibody, in patients with psoriatic arthritis. *J Clin Pharmacol* 2009;49(9):1056-70. doi: 10.1177/0091270009339192 [published Online First: 2009/07/21]
159. Fronton L, Pilari S, Huisinga W. Monoclonal antibody disposition: a simplified PBPK model and its implications for the derivation and interpretation of classical compartment models. *J Pharmacokinet Pharmacodyn* 2014;41(2):87-107. doi: 10.1007/s10928-014-9349-1 [published Online First: 2014/02/05]
160. Waldmann TA, Strober W. Metabolism of immunoglobulins. *Prog Allergy* 1969;13:1-110. doi: 10.1159/000385919 [published Online First: 1969/01/01]
161. Tabrizi MA, Tseng CM, Roskos LK. Elimination mechanisms of therapeutic monoclonal antibodies. *Drug Discov Today* 2006;11(1-2):81-8. doi: 10.1016/S1359-6446(05)03638-X [published Online First: 2006/02/16]
162. Morell A, Terry WD, Waldmann TA. Metabolic properties of IgG subclasses in man. *J Clin Invest* 1970;49(4):673-80. doi: 10.1172/JCI106279 [published Online First: 1970/04/01]

163. An G. Concept of Pharmacologic Target-Mediated Drug Disposition in Large-Molecule and Small-Molecule Compounds. *J Clin Pharmacol* 2020;60(2):149-63. doi: 10.1002/jcph.1545 [published Online First: 2019/12/04]
164. Gibiansky L, Gibiansky E. Target-mediated drug disposition model: approximations, identifiability of model parameters and applications to the population pharmacokinetic-pharmacodynamic modeling of biologics. *Expert Opin Drug Metab Toxicol* 2009;5(7):803-12. doi: 10.1517/17425250902992901 [published Online First: 2009/06/10]
165. Gibiansky L, Gibiansky E, Kakkar T, et al. Approximations of the target-mediated drug disposition model and identifiability of model parameters. *J Pharmacokinet Pharmacodyn* 2008;35(5):573-91. doi: 10.1007/s10928-008-9102-8 [published Online First: 2008/11/14]
166. Mould DR, Upton RN. Basic concepts in population modeling, simulation, and model-based drug development. *CPT Pharmacometrics Syst Pharmacol* 2012;1:e6. doi: 10.1038/psp.2012.4
167. Box GEP, Draper NR. Empirical Model-building and Response Surfaces. New York: John Wiley & Sons, Inc. 1986.
168. Atkinson AJ, Jr., Lalonde RL. Introduction of quantitative methods in pharmacology and clinical pharmacology: a historical overview. *Clin Pharmacol Ther* 2007;82(1):3-6. doi: 10.1038/sj.clpt.6100248 [published Online First: 2007/06/16]
169. Cobelli C, Foster D, Toffolo G. Tracer Kinetics in Biomedical Research: From Data to Model. New York: Kluwer Academic/Plenum Publishers 2000.
170. Wagner JG. Fundamentals of Clinical Pharmacokinetics. Hamilton: Drug Intelligence Publications Inc. 1975.
171. Sheiner LB, Rosenberg B, Melmon KL. Modelling of individual pharmacokinetics for computer-aided drug dosage. *Comput Biomed Res* 1972;5(5):411-59. doi: 10.1016/0010-4809(72)90051-1 [published Online First: 1972/10/01]
172. Lee H, Kimko HC, Rogge M, et al. Population pharmacokinetic and pharmacodynamic modeling of etanercept using logistic regression analysis. *Clin Pharmacol Ther* 2003;73(4):348-65. doi: 10.1016/s0009-9236(02)17635-1 [published Online First: 2003/04/24]
173. Zhou H, Mayer PR, Wajdula J, et al. Unaltered etanercept pharmacokinetics with concurrent methotrexate in patients with rheumatoid arthritis. *J Clin Pharmacol* 2004;44(11):1235-43. doi: 10.1177/0091270004268049 [published Online First: 2004/10/22]
174. Zhou SY, Shu C, Korth-Bradley J, et al. Integrated population pharmacokinetics of etanercept in healthy subjects and in patients with rheumatoid arthritis and ankylosing spondylitis. *J Clin Pharmacol* 2011;51(6):864-75. doi: 10.1177/0091270010375961
175. Shennak M, Al-Jaouni R, Kshirasagar S, et al. An Open-Label, Randomized, Single-Dose, Crossover, Comparative Pharmacokinetics Study of YLB113 and the Etanercept Reference Product in Healthy Adult Male Subjects. *Eur J Drug Metab Pharmacokinet* 2020;45(4):467-75. doi: 10.1007/s13318-020-00613-9 [published Online First: 2020/03/17]
176. Hsu LF, Huang JD. Evaluation of etanercept dose reduction in patients with rheumatoid arthritis using pharmacokinetic/pharmacodynamic modeling and simulation. *Int J Clin Pharmacol Ther* 2014;52(9):776-86. doi: 10.5414/CP202131 [published Online First: 2014/06/03]
177. Research CfDEa, research CfBEa. Application number: 125057/0. Clinical pharmacology and biopharmaceutics review(s) 2002 [Available from: https://www.accessdata.fda.gov/drugsatfda_docs/nda/2002/BLA_125057_S000_HUMIRA_BIOPHARMR.PDF accessed 19th July 2021.
178. Weisman MH, Moreland LW, Furst DE, et al. Efficacy, pharmacokinetic, and safety assessment of adalimumab, a fully human anti-tumor necrosis factor-alpha monoclonal antibody, in adults with rheumatoid arthritis receiving concomitant methotrexate: a pilot study. *Clin Ther* 2003;25(6):1700-21. doi: 10.1016/s0149-2918(03)80164-9 [published Online First: 2003/07/16]
179. Ducourau E, Ternant D, Lequerre T, et al. Towards an individualised target concentration of adalimumab in rheumatoid arthritis. *Ann Rheum Dis* 2014;73(7):1428-9. doi: 10.1136/annrheumdis-2013-204971 [published Online First: 2014/02/20]

180. Ternant D, Ducourau E, Fuzibet P, et al. Pharmacokinetics and concentration-effect relationship of adalimumab in rheumatoid arthritis. *Br J Clin Pharmacol* 2015;79(2):286-97. doi: 10.1111/bcp.12509
181. Lacroix BD, Lovern MR, Stockis A, et al. A pharmacodynamic Markov mixed-effects model for determining the effect of exposure to certolizumab pegol on the ACR20 score in patients with rheumatoid arthritis. *Clin Pharmacol Ther* 2009;86(4):387-95. doi: 10.1038/clpt.2009.136 [published Online First: 2009/07/25]
182. Lacroix BD, Karlsson MO, Friberg LE. Simultaneous Exposure-Response Modeling of ACR20, ACR50, and ACR70 Improvement Scores in Rheumatoid Arthritis Patients Treated With Certolizumab Pegol. *CPT Pharmacometrics Syst Pharmacol* 2014;3:e143. doi: 10.1038/psp.2014.41 [published Online First: 2014/10/30]
183. Kimura K, Takayanagi R, Yokoyama H, et al. Theory-based analysis of anti-inflammatory effect of infliximab on Crohn's disease and rheumatoid arthritis. *Rheumatol Int* 2012;32(1):145-50. doi: 10.1007/s00296-010-1553-8 [published Online First: 2010/08/04]
184. Ternant D, Ducourau E, Perdriger A, et al. Relationship between inflammation and infliximab pharmacokinetics in rheumatoid arthritis. *Br J Clin Pharmacol* 2014;78(1):118-28. doi: 10.1111/bcp.12313 [published Online First: 2013/12/21]
185. Palaparthy R, Rehman MI, von Richter O, et al. Population pharmacokinetics of PF-06438179/GP1111 (an infliximab biosimilar) and reference infliximab in patients with moderately to severely active rheumatoid arthritis. *Expert Opin Biol Ther* 2019;19(10):1065-74. doi: 10.1080/14712598.2019.1635583 [published Online First: 2019/07/10]
186. Roy A, Mould DR, Wang XF, et al. Modeling and simulation of abatacept exposure and interleukin-6 response in support of recommended doses for rheumatoid arthritis. *J Clin Pharmacol* 2007;47(11):1408-20. doi: 10.1177/0091270007307573 [published Online First: 2007/10/27]
187. Li X, Roy A, Murthy B. Population Pharmacokinetics and Exposure-Response Relationship of Intravenous and Subcutaneous Abatacept in Patients With Rheumatoid Arthritis. *J Clin Pharmacol* 2019;59(2):245-57. doi: 10.1002/jcph.1308 [published Online First: 2018/09/20]
188. Hu C, Xu Z, Zhang Y, et al. Population approach for exposure-response modeling of golimumab in patients with rheumatoid arthritis. *J Clin Pharmacol* 2011;51(5):639-48. doi: 10.1177/0091270010372520 [published Online First: 2010/07/14]
189. Hu C, Xu Z, Mendelsohn AM, et al. Latent variable indirect response modeling of categorical endpoints representing change from baseline. *J Pharmacokinet Pharmacodyn* 2013;40(1):81-91. doi: 10.1007/s10928-012-9288-7 [published Online First: 2013/01/01]
190. Zhou H, Jang H, Fleischmann RM, et al. Pharmacokinetics and safety of golimumab, a fully human anti-TNF-alpha monoclonal antibody, in subjects with rheumatoid arthritis. *J Clin Pharmacol* 2007;47(3):383-96. doi: 10.1177/0091270006298188
191. Ogungbenro K, Aarons L. How many subjects are necessary for population pharmacokinetic experiments? Confidence interval approach. *Eur J Clin Pharmacol* 2008;64(7):705-13. doi: 10.1007/s00228-008-0493-7 [published Online First: 2008/05/17]
192. Gueorguieva I, Ogungbenro K, Graham G, et al. A program for individual and population optimal design for univariate and multivariate response pharmacokinetic-pharmacodynamic models. *Comput Methods Programs Biomed* 2007;86(1):51-61. doi: 10.1016/j.cmpb.2007.01.004 [published Online First: 2007/02/13]
193. Yim DS, Zhou H, Buckwalter M, et al. Population pharmacokinetic analysis and simulation of the time-concentration profile of etanercept in pediatric patients with juvenile rheumatoid arthritis. *J Clin Pharmacol* 2005;45(3):246-56. doi: 10.1177/0091270004271945 [published Online First: 2005/02/11]
194. Delyon B, Lavielle M, Moulines E. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics* 1999;27(1):94-128.
195. The R Project for Statistical Computing [Available from: https://www.r-project.org/ accessed 24th June 2019.
196. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth ed. New York: Springer 2002.
197. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag 2016.

198. Wang W, Hallow KM, James DA. A Tutorial on RxODE: Simulating Differential Equation Pharmacometric Models in R. *CPT Pharmacometrics Syst Pharmacol* 2016;5(1):3-10. doi: 10.1002/psp4.12052 [published Online First: 2016/02/05]
199. Borchers HW. pracma: Practical Numerical Math Functions [Available from: https://CRAN.R-project.org/package=pracma accessed 17th January 2022.
200. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012;28(1):112-8. doi: 10.1093/bioinformatics/btr597 [published Online First: 2011/11/01]
201. UniProt [Available from: https://www.uniprot.org.help/about.
202. Kyoto Encyclopedia of Genes and Genomes.
203. Deutsch EW. Mass spectrometer output file format mzML. *Methods Mol Biol* 2010;604:319-31. doi: 10.1007/978-1-60761-444-9_22 [published Online First: 2009/12/17]
204. Rost HL, Rosenberger G, Navarro P, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* 2014;32(3):219-23. doi: 10.1038/nbt.2841 [published Online First: 2014/04/15]
205. Liu Y, Buil A, Collins BC, et al. Quantitative variability of 342 plasma proteins in a human twin population. *Mol Syst Biol* 2015;11(1):786. doi: 10.15252/msb.20145728 [published Online First: 2015/02/06]
206. Teleman J, Rost HL, Rosenberger G, et al. DIANA--algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* 2015;31(4):555-62. doi: 10.1093/bioinformatics/btu686 [published Online First: 2014/10/29]
207. MS Proteomics Tools [Available from: https://github.com/msproteomicstools/msproteomicstools accessed 9th September 2021.
208. Bioconductor [Available from: https://www.bioconductor.org/ accessed 9th September 2021.
209. Choi M, Chang CY, Clough T, et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* 2014;30(17):2524-6. doi: 10.1093/bioinformatics/btu305 [published Online First: 2014/05/06]
210. Blattmann P, Heusel M, Aebersold R. SWATH2stats: An R/Bioconductor Package to Process and Convert Quantitative SWATH-MS Proteomics Data for Downstream Analysis Tools. *PLoS One* 2016;11(4):e0153160. doi: 10.1371/journal.pone.0153160 [published Online First: 2016/04/08]
211. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *Journal of Open Source Software* 2019;4:1686-91. doi: https://doi.org/0.21105/joss.01686
212. Kuhn M. The caret Package 2019 [Available from: https://topepo.github.io/caret/ accessed 9th September 2021.
213. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 2016;32(18):2847-9. doi: 10.1093/bioinformatics/btw313 [published Online First: 2016/05/22]
214. Feng L, Moritz S, Nowak G, et al. imputeR: A general multivariate imputation framework. R package version 2.1. 2018 [Available from: https://cran.r-project.org/web/packages/imputeR/index.html accessed 9th September 2021.
215. Franzin A, Sambo F, Di Camillo B. bnstruct: an R package for Bayesian Network structure learning in the presence of missing data. *Bioinformatics* 2017;33(8):1250-52. doi: 10.1093/bioinformatics/btw807 [published Online First: 2016/12/23]
216. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 2011;45:1-67. doi: https://doi.org/10.18637/jss.v045.i03
217. Leek JT, Johnson WE, Parker HS, et al. sva: Surrogate variable analysis. R package version 3.40.0. 2021 [Available from: https://bioconductor.org/packages/release/bioc/html/sva.html accessed 9th September 2021.
218. Koncevicius K. matrixTests: Fast statistical hypothesis tests on rows and columns of matrices 2020 [Available from: https://CRAN.R-project.org/package=matrixTests accessed 10th September 2021.

219. Pouw MF, Krieckaert CL, Nurmohamed MT, et al. Key findings towards optimising adalimumab treatment: the concentration-effect curve. *Ann Rheum Dis* 2015;74(3):513-8. doi: 10.1136/annrheumdis-2013-204172 [published Online First: 2013/12/12]
220. Jamnitski A, Krieckaert CL, Nurmohamed MT, et al. Patients non-responding to etanercept obtain lower etanercept concentrations compared with responding patients. *Ann Rheum Dis* 2012;71(1):88-91. doi: 10.1136/annrheumdis-2011-200184 [published Online First: 2011/09/15]
221. Bates D, Machler M, Bolker B, et al. FItting linear mixed-effects models using lme4. *Journal of Statistical Software* 2015;67:1-48. doi: https://doi.org/10.18637/jss.v067.i01
222. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* 2017;82:1-26. doi: https://doi.org/10.18637/jss.v082.i13
223. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* 1995;57:289-300.
224. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43(7):e47. doi: 10.1093/nar/gkv007 [published Online First: 2015/01/22]
225. Siriseriwan W. smotefamily: A collection of oversampling techniques for class imbalance problem based on SMOTE 2019 [Available from: https://CRAN.R-project.org/package=smotefamily accessed 9th September 2021.
226. Bischl B, Lang M, Kotthoff L, et al. mlr: Machine learning in R. *Journal of Machine Learning Research* 2016;7:1-5.
227. Greenwood CJ, Youssef GJ, Letcher P, et al. A comparison of penalised regression methods for informing the selection of predictive markers. *PLoS One* 2020;15(11):e0242730. doi: 10.1371/journal.pone.0242730 [published Online First: 2020/11/21]
228. Defazio A, Campbell H. classifierplots: Generates a visalization of classifier performance as a grid of diagnostic plots 2020 [Available from: https://CRAN.R-project.org/package=classifierplots accessed 9th September 2021.
229. Casalicchio G, Bischl B. RBPcurve: The Residual-Based Predictiveness Curve 2017 [accessed 9th September 2021.
230. Lele SR, Keim JL, Solymos P. ResourceSelection: Resource selection(probability) functions for use-availability data 2019 [Available from: https://CRAN.R-project.org/package=ResourceSelection accessed 9th September 2021.
231. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;47(D1):D607-D13. doi: 10.1093/nar/gky1131 [published Online First: 2018/11/27]
232. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;49(D1):D605-D12. doi: 10.1093/nar/gkaa1074 [published Online First: 2020/11/26]
233. Albarino CG, Romanowski V. Phenol extraction revisited: a rapid method for the isolation and preservation of human genomic DNA from whole blood. *Mol Cell Probes* 1994;8(5):423-7. doi: 10.1006/mcpr.1994.1060 [published Online First: 1994/10/01]
234. Purcell S. PLINK v.1.9 [Available from: http://pngu.mgh.harvard.edu/purcell/plink/ accessed 10th September 2021.
235. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81(3):559-75. doi: 10.1086/519795 [published Online First: 2007/08/19]
236. The Haplotype Reference Consortium [Available from: http://www.haplotype-reference-consortium.org/ accessed 10th September 2021.

237. HapMap3 [Available from: https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html accessed 9th April 2022.
238. Michigan Imputation Server [Available from: https://imputationserver.sph.umich.edu/index.html#! accessed 10th September 2021.
239. Loh PR. Eagle v2.4.1 User Manual 2018 [Available from: https://alkesgroup.broadinstitute.org/Eagle/ accessed 9th April 2022.
240. Minimac4 2018 [Available from: https://genome.sph.umich.edu/wiki/Minimac4 accessed 10th September 2021.
241. The Haplotype Reference Consortiu, [Available from: http://www.haplotype-reference-consortium.org/ accessed 9th April 2022.
242. Ochoa A. genio: Genetics Input/Output Functions [updated 26th July 2021. Available from: https://CRAN.R-project.org/package=genio accessed 20th January 2022.
243. Rainer J, Gatto L, Weichenberger CX. ensembldb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics* 2019;35(17):3151-53. doi: 10.1093/bioinformatics/btz031 [published Online First: 2019/01/29]
244. Ensembl [Available from: https://www.ensembl.org/index.html accessed 10th September 2021.
245. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 2012;28(10):1353-8. doi: 10.1093/bioinformatics/bts163 [published Online First: 2012/04/12]
246. dbSNP [Available from: https://www.ncbi.nlm.nih.gov/snp/ accessed 30th January 2022.
247. GTEx Portal [Available from: https://gtexportal.org/home/ accessed 22nd January 2022.
248. Massey J, Plant D, Hyrich K, et al. Genome-wide association study of response to tumour necrosis factor inhibitor therapy in rheumatoid arthritis. *Pharmacogenomics J* 2018;18(5):657-64. doi: 10.1038/s41397-018-0040-6 [published Online First: 2018/09/01]
249. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* 2019;8(7):giz082. doi: https://doi.org/10.1093/gigascience/giz082
250. Korth-Bradley JM, Rubin AS, Hanna RK, et al. The pharmacokinetics of etanercept in healthy volunteers. *Ann Pharmacother* 2000;34(2):161-4. doi: 10.1345/aph.19126 [published Online First: 2000/02/17]
251. FDA. Humira clnical pharmacology biopharmeutics review 2002 [Available from: https://www.accessdata.fda.gov/drugsatfda_docs/nda/2002/BLA_125057_S000_HUMIRA_BIOPHARMR.PDF accessed 2nd December 2021.
252. DeGre M, Mellbye OJ, Clarke-Jenssen O. Immune interferon in serum and synovial fluid in rheumatoid arthritis and related disorders. *Ann Rheum Dis* 1983;42(6):672-6.
253. Biemond P, Swaak AJ, Koster JF. Protective factors against oxygen free radicals and hydrogen peroxide in rheumatoid arthritis synovial fluid. *Arthritis Rheum* 1984;27(7):760-5.
254. Gysen P, Malaise M, Gaspar S, et al. Measurement of proteoglycans, elastase, collagenase and protein in synovial fluid in inflammatory and degenerative arthropathies. *Clin Rheumatol* 1985;4(1):39-50.
255. Malyak M, Swaney RE, Arend WP. Levels of synovial fluid interleukin-1 receptor antagonist in rheumatoid arthritis and other arthropathies. Potential contribution from synovial fluid neutrophils. *Arthritis Rheum* 1993;36(6):781-9.
256. Okano K, Tsukazaki T, Ohtsuru A, et al. Parathyroid hormone-related peptide in synovial fluid and disease activity of rheumatoid arthritis. *Br J Rheumatol* 1996;35(11):1056-62.
257. Schaffler A, Ehling A, Neumann E, et al. Adipocytokines in synovial fluid. *JAMA* 2003;290(13):1709-10. doi: 10.1001/jama.290.13.1709-c
258. Kim CW, Cho EH, Lee YJ, et al. Disease-specific proteins from rheumatoid arthritis patients. *J Korean Med Sci* 2006;21(3):478-84. doi: 10.3346/jkms.2006.21.3.478
259. Tabushi Y, Nakanishi T, Takeuchi T, et al. Detection of citrullinated proteins in synovial fluids derived from patients with rheumatoid arthritis by proteomics-based analysis. *Ann Clin Biochem* 2008;45(Pt 4):413-7. doi: 10.1258/acb.2007.007205

260. Katano M, Okamoto K, Arito M, et al. Implication of granulocyte-macrophage colony-stimulating factor induced neutrophil gelatinase-associated lipocalin in pathogenesis of rheumatoid arthritis revealed by proteome analysis. *Arthritis Res Ther* 2009;11(1):R3.
261. Baillet A, Trocme C, Berthier S, et al. Synovial fluid proteomic fingerprint: S100A8, S100A9 and S100A12 proteins discriminate rheumatoid arthritis from other inflammatory joint diseases. *Rheumatology (Oxford)* 2010;49(4):671-82. doi: 10.1093/rheumatology/kep452
262. Mateos J, Lourido L, Fernandez-Puente P, et al. Differential protein profiling of synovial fluid from rheumatoid arthritis and osteoarthritis patients using LC-MALDI TOF/TOF. *J Proteomics* 2012;75(10):2869-78. doi: 10.1016/j.jprot.2011.12.042
263. Noh R, Park SG, Ju JH, et al. Comparative proteomic analyses of synovial fluids and serums from rheumatoid arthritis patients. *J Microbiol Biotechnol* 2014;24(1):119-26.
264. Yang XY, Zheng KD, Lin K, et al. Energy Metabolism Disorder as a Contributing Factor of Rheumatoid Arthritis: A Comparative Proteomic and Metabolomic Study. *PLoS One* 2015;10(7):e0132695. doi: 10.1371/journal.pone.0132695
265. Meng X, Ezzati P, Smolik I, et al. Characterization of Autoantigens Targeted by Anti-Citrullinated Protein Antibodies In Vivo: Prominent Role for Epitopes Derived from Histone 4 Proteins. *PLoS One* 2016;11(10):e0165501. doi: 10.1371/journal.pone.0165501
266. Firestein GS, Berger AE, Tracey DE, et al. IL-1 receptor antagonist protein production and gene expression in rheumatoid arthritis and osteoarthritis synovium. *J Immunol* 1992;149(3):1054-62.
267. Yamasaki S, Kawakami A, Nakashima T, et al. Importance of NF-kappaB in rheumatoid synovial tissues: in situ NF-kappaB expression and in vitro study using cultured synovial cells. *Ann Rheum Dis* 2001;60(7):678-84.
268. De Rycke L, Nicholas AP, Cantaert T, et al. Synovial intracellular citrullinated proteins colocalizing with peptidyl arginine deiminase as pathophysiologically relevant antigenic determinants of rheumatoid arthritis-specific humoral autoimmunity. *Arthritis Rheum* 2005;52(8):2323-30. doi: 10.1002/art.21220
269. Chang X, Cui Y, Zong M, et al. Identification of proteins with increased expression in rheumatoid arthritis synovial tissues. *J Rheumatol* 2009;36(5):872-80. doi: 10.3899/jrheum.080939
270. Wang JG, Xu WD, Zhai WT, et al. Disorders in angiogenesis and redox pathways are main factors contributing to the progression of rheumatoid arthritis: a comparative proteomics study. *Arthritis Rheum* 2012;64(4):993-1004. doi: 10.1002/art.33425 [published Online First: 2011/10/19]
271. Yan X, Zhao Y, Pan J, et al. Vitamin D-binding protein (group-specific component) has decreased expression in rheumatoid arthritis. *Clin Exp Rheumatol* 2012;30(4):525-33.
272. Chang X, Zhao Y, Wang Y, et al. Screening citrullinated proteins in synovial tissues of rheumatoid arthritis using 2-dimensional western blotting. *J Rheumatol* 2013;40(3):219-27. doi: 10.3899/jrheum.120751
273. Doran MC, Goodstone NJ, Hobbs RN, et al. Cellular immunity to cartilage link protein in patients with inflammatory arthritis and non-arthritic controls. *Ann Rheum Dis* 1995;54(6):466-70.
274. Swedlund HA, Hunder GG, Gleich GJ. Alpha 1-antitrypsin in serum and synovial fluid in rheumatoid arthritis. *Ann Rheum Dis* 1974;33(2):162-4.
275. Baskol G, Demir H, Baskol M, et al. Investigation of protein oxidation and lipid peroxidation in patients with rheumatoid arthritis. *Cell Biochem Funct* 2006;24(4):307-11. doi: 10.1002/cbf.1257 [published Online First: 2005/09/06]
276. Grazio S, Razdorov G, Erjavec I, et al. Differential expression of proteins with heparin affinity in patients with rheumatoid and psoriatic arthritis: a preliminary study. *Clin Exp Rheumatol* 2013;31(5):665-71.
277. Yang L, Zou QH, Zhang Y, et al. Proteomic analysis of plasma from rheumatoid arthritis patients with mild cognitive impairment. *Clin Rheumatol* 2018;37(7):1773-82. doi: 10.1007/s10067-017-3974-1

278. Schulz M, Dotzlaw H, Mikkat S, et al. Proteomic analysis of peripheral blood mononuclear cells: selective protein processing observed in patients with rheumatoid arthritis. *J Proteome Res* 2007;6(9):3752-9. doi: 10.1021/pr070285f
279. Lu MC, Lai NS, Yu HC, et al. Anti-citrullinated protein antibodies bind surface-expressed citrullinated Grp78 on monocyte/macrophages and stimulate tumor necrosis factor alpha production. *Arthritis Rheum* 2010;62(5):1213-23. doi: 10.1002/art.27386
280. Darrah E, Kim A, Zhang X, et al. Proteolysis by Granzyme B Enhances Presentation of Autoantigenic Peptidylarginine Deiminase 4 Epitopes in Rheumatoid Arthritis. *J Proteome Res* 2017;16(1):355-65. doi: 10.1021/acs.jproteome.6b00617
281. Olszewski WL, Pazdur J, Kubasiewicz E, et al. Lymph draining from foot joints in rheumatoid arthritis provides insight into local cytokine and chemokine production and transport to lymph nodes. *Arthritis Rheum* 2001;44(3):541-9. doi: 10.1002/1529-0131(200103)44:3<541::AID-ANR102>3.0.CO;2-6
282. Doroshevskaya AY, Kondratovskii PM, Dubikov AI, et al. Apoptosis regulator proteins: basis for the development of innovation strategies for the treatment of rheumatoid arthritis in patients of different age. *Bull Exp Biol Med* 2014;156(3):377-80. doi: 10.1007/s10517-014-2353-z
283. Wagatsuma M, Kimura M, Suzuki R, et al. Ezrin, radixin and moesin are possible auto-immune antigens in rheumatoid arthritis. *Mol Immunol* 1996;33(15):1171-6.
284. Chandra PE, Sokolove J, Hipp BG, et al. Novel multiplex technology for diagnostic characterization of rheumatoid arthritis. *Arthritis Res Ther* 2011;13(3):R102. doi: 10.1186/ar3383
285. Urbaniak B, Nowicki P, Sikorska D, et al. The feature selection approach for evaluation of potential rheumatoid arthritis markers using MALDI-TOF datasets. *Anal Biochem* 2017;525:29-37. doi: 10.1016/j.ab.2017.02.016
286. Seok A, Lee HJ, Lee S, et al. Identification and Validation of SAA4 as a Rheumatoid Arthritis Prescreening Marker by Liquid Chromatography Tandem-mass Spectrometry. *Molecules* 2017;22(5) doi: 10.3390/molecules22050805
287. Kim D, Mun S, Lee J, et al. Proteomics analysis reveals differential pattern of widespread protein expression and novel role of histidine-rich glycoprotein and lipopolysaccharide-binding protein in rheumatoid arthritis. *Int J Biol Macromol* 2018;109:704-10. doi: 10.1016/j.ijbiomac.2017.12.075
288. Giusti L, Baldini C, Ciregia F, et al. Is GRP78/BiP a potential salivary biomarker in patients with rheumatoid arthritis? *Proteomics Clin Appl* 2010;4(3):315-24. doi: 10.1002/prca.200900082
289. Siebert S, Porter D, Paterson C, et al. Urinary proteomics can define distinct diagnostic inflammatory arthritis subgroups. *Sci Rep* 2017;7:40473. doi: 10.1038/srep40473
290. Yang X, Qiu P, Chen B, et al. KIAA1199 as a potential diagnostic biomarker of rheumatoid arthritis related to angiogenesis. *Arthritis Res Ther* 2015;17:140. doi: 10.1186/s13075-015-0637-y
291. Hueber W, Tomooka BH, Zhao X, et al. Proteomic analysis of secreted proteins in early rheumatoid arthritis: anti-citrulline autoreactivity is associated with up regulation of proinflammatory cytokines. *Ann Rheum Dis* 2007;66(6):712-9. doi: 10.1136/ard.2006.054924
292. Cheng Y, Chen Y, Sun X, et al. Identification of potential serum biomarkers for rheumatoid arthritis by high-resolution quantitative proteomic analysis. *Inflammation* 2014;37(5):1459-67. doi: 10.1007/s10753-014-9871-8
293. Kang MJ, Park YJ, You S, et al. Urinary proteome profile predictive of disease activity in rheumatoid arthritis. *J Proteome Res* 2014;13(11):5206-17. doi: 10.1021/pr500467d
294. Park YJ, Yoo SA, Hwang D, et al. Identification of novel urinary biomarkers for assessing disease activity and prognosis of rheumatoid arthritis. *Exp Mol Med* 2016;48:e211. doi: 10.1038/emm.2015.120
295. Krasny L, Bland P, Kogata N, et al. SWATH mass spectrometry as a tool for quantitative profiling of the matrisome. *J Proteomics* 2018;189:11-22. doi: 10.1016/j.jprot.2018.02.026 [published Online First: 2018/03/05]

296. Luo R, Zhao H. Protein quantitation using iTRAQ: Review on the sources of variations and analysis of nonrandom missingness. *Stat Interface* 2012;5(1):99-107. doi: 10.4310/sii.2012.v5.n1.a9 [published Online First: 2012/01/01]
297. Jaffe AE, Tao R, Norris AL, et al. qSVA framework for RNA quality correction in differential expression analysis. *Proc Natl Acad Sci U S A* 2017;114(27):7130-35. doi: 10.1073/pnas.1617384114 [published Online First: 2017/06/22]
298. Segurado OG, Sasso EH. Vectra DA for the objective measurement of disease activity in patients with rheumatoid arthritis. *Clin Exp Rheumatol* 2014;32(5 Suppl 85):S-29-34. [published Online First: 2014/11/05]
299. UniProtKB. P0C0L5 (CO4B_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/P0C0L5 accessed 3rd January 2022.
300. Rigby WF, Wu YL, Zan M, et al. Increased frequency of complement C4B deficiency in rheumatoid arthritis. *Arthritis Rheum* 2012;64(5):1338-44. doi: 10.1002/art.33472 [published Online First: 2011/11/15]
301. Holers VM, Borodovsky A, Scheinman RI, et al. Key Components of the Complement Lectin Pathway Are Not Only Required for the Development of Inflammatory Arthritis but Also Regulate the Transcription of Factor D. *Front Immunol* 2020;11:201. doi: 10.3389/fimmu.2020.00201 [published Online First: 2020/03/11]
302. UniProtKB. P02763 (A1AG1_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/P02763 accessed 4th January 2022.
303. Ben-Hadj-Mohamed M, Khelil S, Ben Dbibis M, et al. Hepatic Proteins and Inflammatory Markers in Rheumatoid Arthritis Patients. *Iran J Public Health* 2017;46(8):1071-78. [published Online First: 2017/09/13]
304. Saroha A, Biswas S, Chatterjee BP, et al. Altered glycosylation and expression of plasma alpha-1-acid glycoprotein and haptoglobin in rheumatoid arthritis. *J Chromatogr B Analyt Technol Biomed Life Sci* 2011;879(20):1839-43. doi: 10.1016/j.jchromb.2011.04.024 [published Online First: 2011/05/24]
305. Ryden I, Pahlsson P, Lundblad A, et al. Fucosylation of alpha1-acid glycoprotein (orosomucoid) compared with traditional biochemical markers of inflammation in recent onset rheumatoid arthritis. *Clin Chim Acta* 2002;317(1-2):221-9. doi: 10.1016/s0009-8981(01)00803-8 [published Online First: 2002/01/30]
306. Haston JL, FitzGerald O, Kane D, et al. The influence of alpha1-acid glycoprotein on collagenase-3 activity in early rheumatoid arthritis. *Biomed Chromatogr* 2003;17(6):361-4. doi: 10.1002/bmc.251 [published Online First: 2003/09/19]
307. Smith KD, Pollacchi A, Field M, et al. The heterogeneity of the glycosylation of alpha-1-acid glycoprotein between the sera and synovial fluid in rheumatoid arthritis. *Biomed Chromatogr* 2002;16(4):261-6. doi: 10.1002/bmc.158 [published Online First: 2002/04/05]
308. Fischer A, Abdollahi-Roodsaz S, Bohm C, et al. The involvement of Toll-like receptor 9 in the pathogenesis of erosive autoimmune arthritis. *J Cell Mol Med* 2018;22(9):4399-409. doi: 10.1111/jcmm.13735 [published Online First: 2018/07/12]
309. Shembade N, Pujari R, Harhaj NS, et al. The kinase IKKalpha inhibits activation of the transcription factor NF-kappaB by phosphorylating the regulatory molecule TAX1BP1. *Nat Immunol* 2011;12(9):834-43. doi: 10.1038/ni.2066 [published Online First: 2011/07/19]
310. Razani B, Zarnegar B, Ytterberg AJ, et al. Negative feedback in noncanonical NF-kappaB signaling modulates NIK stability through IKKalpha-mediated phosphorylation. *Sci Signal* 2010;3(123):ra41. doi: 10.1126/scisignal.2000778 [published Online First: 2010/05/27]
311. Makarov SS. NF-kappa B in rheumatoid arthritis: a pivotal regulator of inflammation, hyperplasia, and tissue destruction. *Arthritis Res* 2001;3(4):200-6. doi: 10.1186/ar300 [published Online First: 2001/07/05]
312. Potter C, Cordell HJ, Barton A, et al. Association between anti-tumour necrosis factor treatment response and genetic variants within the TLR and NF{kappa}B signalling pathways. *Ann Rheum Dis* 2010;69(7):1315-20. doi: 10.1136/ard.2009.117309 [published Online First: 2010/05/08]

313. Ferreiro-Iglesias A, Montes A, Perez-Pampin E, et al. Replication of PTPRC as genetic biomarker of response to TNF inhibitors in patients with rheumatoid arthritis. *Pharmacogenomics J* 2016;16(2):137-40. doi: 10.1038/tpj.2015.29 [published Online First: 2015/04/22]
314. Yang J, Zaitlen NA, Goddard ME, et al. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* 2014;46(2):100-6. doi: 10.1038/ng.2876 [published Online First: 2014/01/30]
315. Cai B, Giridharan SSP, Zhang J, et al. Differential roles of C-terminal Eps15 homology domain proteins as vesiculators and tubulators of recycling endosomes. *J Biol Chem* 2013;288(42):30172-80. doi: 10.1074/jbc.M113.488627 [published Online First: 2013/09/11]
316. Freund A, Zhong FL, Venteicher AS, et al. Proteostatic control of telomerase function through TRiC-mediated folding of TCAB1. *Cell* 2014;159(6):1389-403. doi: 10.1016/j.cell.2014.10.059 [published Online First: 2014/12/04]
317. Zipfel PF, Edey M, Heinen S, et al. Deletion of complement factor H-related genes CFHR1 and CFHR3 is associated with atypical hemolytic uremic syndrome. *PLoS Genet* 2007;3(3):e41. doi: 10.1371/journal.pgen.0030041 [published Online First: 2007/03/21]
318. UniProtKB. Q02985 (FHR3_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/Q02985 accessed 9th January 2022.
319. Zoidis E, Ghirlanda-Keller C, Schmid C. Stimulation of glucose transport in osteoblastic cells by parathyroid hormone and insulin-like growth factor I. *Mol Cell Biochem* 2011;348(1-2):33-42. doi: 10.1007/s11010-010-0634-z [published Online First: 2010/11/16]
320. Baum R, Gravallese EM. Impact of inflammation on the osteoblast in rheumatic diseases. *Curr Osteoporos Rep* 2014;12(1):9-16. doi: 10.1007/s11914-013-0183-y [published Online First: 2013/12/24]
321. Harashima A, Guettouche T, Barber GN. Phosphorylation of the NFAR proteins by the dsRNA-dependent protein kinase PKR constitutes a novel mechanism of translational regulation and cellular defense. *Genes Dev* 2010;24(23):2640-53. doi: 10.1101/gad.1965010 [published Online First: 2010/12/03]
322. Dinchuk JE, Focht RJ, Kelley JA, et al. Absence of post-translational aspartyl beta-hydroxylation of epidermal growth factor domains in mice leads to developmental defects and an increased incidence of intestinal neoplasia. *J Biol Chem* 2002;277(15):12970-7. doi: 10.1074/jbc.M110389200 [published Online First: 2002/01/05]
323. Srikanth S, Jew M, Kim KD, et al. Junctate is a Ca2+-sensing structural component of Orai1 and stromal interaction molecule 1 (STIM1). *Proc Natl Acad Sci U S A* 2012;109(22):8682-7. doi: 10.1073/pnas.1200667109 [published Online First: 2012/05/16]
324. Bloom BR, Bennett B. Mechanism of a reaction in vitro associated with delayed-type hypersensitivity. *Science* 1966;153(3731):80-2. doi: 10.1126/science.153.3731.80 [published Online First: 1966/07/01]
325. UniProtKB. P0DJI8 (SAA1_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/P0DJI8 accessed 9th January 2022.
326. UniProtKB. P0DJI9 (SAA2_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/P0DJI9 accessed 9th January 2022.
327. Kanayama HO, Tamura T, Ugai S, et al. Demonstration that a human 26S proteolytic complex consists of a proteasome and multiple associated protein components and hydrolyzes ATP and ubiquitin-ligated proteins by closely linked mechanisms. *Eur J Biochem* 1992;206(2):567-78. doi: 10.1111/j.1432-1033.1992.tb16961.x [published Online First: 1992/06/01]
328. Qin X, Hao Z, Tian Q, et al. Cocrystal structures of glycyl-tRNA synthetase in complex with tRNA suggest multiple conformational states in glycylation. *J Biol Chem* 2014;289(29):20359-69. doi: 10.1074/jbc.M114.557249 [published Online First: 2014/06/06]
329. Guo RT, Chong YE, Guo M, et al. Crystal structures and biochemical analyses suggest a unique mechanism and role for human glycyl-tRNA synthetase in Ap4A homeostasis. *J Biol Chem* 2009;284(42):28968-76. doi: 10.1074/jbc.M109.030692 [published Online First: 2009/08/28]

330. UniProtKB. P06727 (APOA4_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/P06727 accessed 10th January 2022.
331. Rajbhandary R, Khezri A, Panush RS. Rheumatoid cachexia: what is it and why is it important? *J Rheumatol* 2011;38(3):406-8. doi: 10.3899/jrheum.101036 [published Online First: 2011/03/03]
332. Tralau T, Meyer-Hoffert U, Schroder JM, et al. Human leukocyte elastase and cathepsin G are specific inhibitors of C5a-dependent neutrophil enzyme release and chemotaxis. *Exp Dermatol* 2004;13(5):316-25. doi: 10.1111/j.0906-6705.2004.00145.x [published Online First: 2004/05/14]
333. Wright HL, Moots RJ, Edwards SW. The multifactorial role of neutrophils in rheumatoid arthritis. *Nat Rev Rheumatol* 2014;10(10):593-601. doi: 10.1038/nrrheum.2014.80 [published Online First: 2014/06/11]
334. Raingeaud J, Whitmarsh AJ, Barrett T, et al. MKK3- and MKK6-regulated gene expression is mediated by the p38 mitogen-activated protein kinase signal transduction pathway. *Mol Cell Biol* 1996;16(3):1247-55. doi: 10.1128/MCB.16.3.1247 [published Online First: 1996/03/01]
335. UniProtKB. Q00610 (CLH1_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/Q00610 accessed 10th January 2022.
336. UniProtKB. Cellular component - Clathrin-coated pit [Available from: https://www.uniprot.org/locations/69 accessed 10th January 2022.
337. Jerala R. Structural biology of the LPS recognition. *Int J Med Microbiol* 2007;297(5):353-63. doi: 10.1016/j.ijmm.2007.04.001 [published Online First: 2007/05/08]
338. Goicoechea de Jorge E, Caesar JJ, Malik TH, et al. Dimerization of complement factor H-related proteins modulates complement activation in vivo. *Proc Natl Acad Sci U S A* 2013;110(12):4685-90. doi: 10.1073/pnas.1219260110 [published Online First: 2013/03/15]
339. Peterson PA. Studies on the interaction between prealbumin, retinol-binding protein, and vitamin A. *J Biol Chem* 1971;246(1):44-9. [published Online First: 1971/01/10]
340. UniProtKB. P49908 (SEPP1_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/P49908 accessed 10th January 2022.
341. Wang J, Chun HJ, Wong W, et al. Caspase-10 is an initiator caspase in death receptor signaling. *Proc Natl Acad Sci U S A* 2001;98(24):13884-8. doi: 10.1073/pnas.241358198 [published Online First: 2001/11/22]
342. UniProtKB. P29144 (TPP2_HUMAN) [updated 29th Septembe 2021. Available from: https://www.uniprot.org/uniprot/P29144 accessed 10th January 2022.
343. Starheim KK, Arnesen T, Gromyko D, et al. Identification of the human N(alpha)-acetyltransferase complex B (hNatB): a complex important for cell-cycle progression. *Biochem J* 2008;415(2):325-31. doi: 10.1042/BJ20080658 [published Online First: 2008/06/24]
344. UniProtKB. Q04917 (1433F_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/Q04917 accessed 10th January 2022.
345. UniProtKB. P02786 (TFR1_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/P02786 accessed 10th January 2022.
346. UniProtKB. Q9Y446 (PKP3_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/Q9Y446 accessed 10th January 2022.
347. Khan NA, Spencer HJ, Nikiphorou E, et al. Intercentre variance in patient reported outcomes is lower than objective rheumatoid arthritis activity measures: a cross-sectional study. *Rheumatology (Oxford)* 2017;56(8):1395-400. doi: 10.1093/rheumatology/kex076 [published Online First: 2017/06/03]
348. UniProtKB. P20774 (MIME_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/P20774 accessed 19th January 2022.
349. Arnold SM, Fessler LI, Fessler JH, et al. Two homologues encoding human UDP-glucose:glycoprotein glucosyltransferase differ in mRNA expression and enzymatic activity.

Biochemistry 2000;39(9):2149-63. doi: 10.1021/bi9916473 [published Online First: 2000/03/01]
350. Hers I, Vincent EE, Tavare JM. Akt signalling in health and disease. *Cell Signal* 2011;23(10):1515-27. doi: 10.1016/j.cellsig.2011.05.004 [published Online First: 2011/05/31]
351. Yang J, Goetz D, Li JY, et al. An iron delivery pathway mediated by a lipocalin. *Mol Cell* 2002;10(5):1045-56. doi: 10.1016/s1097-2765(02)00710-4 [published Online First: 2002/11/28]
352. Bao G, Clifton M, Hoette TM, et al. Iron traffics in circulation bound to a siderocalin (Ngal)-catechol complex. *Nat Chem Biol* 2010;6(8):602-9. doi: 10.1038/nchembio.402 [published Online First: 2010/06/29]
353. Lee YH, Campbell HD, Stallcup MR. Developmentally essential protein flightless I is a nuclear receptor coactivator with actin binding activity. *Mol Cell Biol* 2004;24(5):2103-17. doi: 10.1128/MCB.24.5.2103-2117.2004 [published Online First: 2004/02/18]
354. Romacho T, Villalobos LA, Cercas E, et al. Visfatin as a novel mediator released by inflamed human endothelial cells. *PLoS One* 2013;8(10):e78283. doi: 10.1371/journal.pone.0078283 [published Online First: 2013/10/17]
355. Clements CM, McNally RS, Conti BJ, et al. DJ-1, a cancer- and Parkinson's disease-associated protein, stabilizes the antioxidant transcriptional master regulator Nrf2. *Proc Natl Acad Sci U S A* 2006;103(41):15091-6. doi: 10.1073/pnas.0607260103 [published Online First: 2006/10/04]
356. Richarme G, Mihoub M, Dairou J, et al. Parkinsonism-associated protein DJ-1/Park7 is a major protein deglycase that repairs methylglyoxal- and glyoxal-glycated cysteine, arginine, and lysine residues. *J Biol Chem* 2015;290(3):1885-97. doi: 10.1074/jbc.M114.597815 [published Online First: 2014/11/25]
357. Advedissian T, Deshayes F, Poirier F, et al. The Parkinsonism-associated protein DJ-1/Park7 prevents glycation damage in human keratinocyte. *Biochem Biophys Res Commun* 2016;473(1):87-91. doi: 10.1016/j.bbrc.2016.03.056 [published Online First: 2016/03/21]
358. Matsuda N, Kimura M, Queliconi BB, et al. Parkinson's disease-related DJ-1 functions in thiol quality control against aldehyde attack in vitro. *Sci Rep* 2017;7(1):12816. doi: 10.1038/s41598-017-13146-0 [published Online First: 2017/10/11]
359. Cao Z, Bhella D, Lindsay JG. Reconstitution of the mitochondrial PrxIII antioxidant defence pathway: general properties and factors affecting PrxIII activity and oligomeric state. *J Mol Biol* 2007;372(4):1022-33. doi: 10.1016/j.jmb.2007.07.018 [published Online First: 2007/08/21]
360. Huber PA, Redwood CS, Avent ND, et al. Identification of functioning regulatory sites and a new myosin binding site in the C-terminal 288 amino acids of caldesmon expressed from a human clone. *J Muscle Res Cell Motil* 1993;14(4):385-91. doi: 10.1007/BF00121289 [published Online First: 1993/08/01]
361. UniProtKB. Q14012 (KCC1A_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/Q14012 accessed 19th January 2022.
362. Cunningham O, Gore MG, Mantle TJ. Initial-rate kinetics of the flavin reductase reaction catalysed by human biliverdin-IXbeta reductase (BVR-B). *Biochem J* 2000;345 Pt 2:393-9. [published Online First: 2000/01/06]
363. Shnitsar V, Li J, Li X, et al. A substrate access tunnel in the cytosolic domain is not an essential feature of the solute carrier 4 (SLC4) family of bicarbonate transporters. *J Biol Chem* 2013;288(47):33848-60. doi: 10.1074/jbc.M113.511865 [published Online First: 2013/10/15]
364. Suriano AR, Sanford AN, Kim N, et al. GCF2/LRRFIP1 represses tumor necrosis factor alpha expression. *Mol Cell Biol* 2005;25(20):9073-81. doi: 10.1128/MCB.25.20.9073-9081.2005 [published Online First: 2005/10/04]
365. Dai P, Jeong SY, Yu Y, et al. Modulation of TLR signaling by multiple MyD88-interacting partners including leucine-rich repeat Fli-I-interacting proteins. *J Immunol* 2009;182(6):3450-60. doi: 10.4049/jimmunol.0802260 [published Online First: 2009/03/07]

366. Chuang NN, Huang CC. Interaction of integrin beta1 with cytokeratin 1 in neuroblastoma NMB7 cells. *Biochem Soc Trans* 2007;35(Pt 5):1292-4. doi: 10.1042/BST0351292 [published Online First: 2007/10/25]
367. Miyazono K, Olofsson A, Colosetti P, et al. A role of the latent TGF-beta 1-binding protein in the assembly and secretion of TGF-beta 1. *EMBO J* 1991;10(5):1091-101. [published Online First: 1991/05/01]
368. Stockis J, Colau D, Coulie PG, et al. Membrane protein GARP is a receptor for latent TGF-beta on the surface of activated human Treg. *Eur J Immunol* 2009;39(12):3315-22. doi: 10.1002/eji.200939684 [published Online First: 2009/09/15]
369. Kikuchi M, Doi E, Tsujimoto I, et al. Functional analysis of human P5, a protein disulfide isomerase homologue. *J Biochem* 2002;132(3):451-5. doi: 10.1093/oxfordjournals.jbchem.a003242 [published Online First: 2002/09/03]
370. Tuteja N, Tuteja R, Ochem A, et al. Human DNA helicase II: a novel DNA unwinding enzyme identified as the Ku autoantigen. *EMBO J* 1994;13(20):4991-5001. [published Online First: 1994/10/17]
371. Reeves WH. Antibodies to the p70/p80 (Ku) antigens in systemic lupus erythematosus. *Rheum Dis Clin North Am* 1992;18(2):391-414. [published Online First: 1992/05/01]
372. Tsang WY, Spektor A, Luciano DJ, et al. CP110 cooperates with two calcium-binding proteins to regulate cytokinesis and genome stability. *Mol Biol Cell* 2006;17(8):3423-34. doi: 10.1091/mbc.e06-04-0371 [published Online First: 2006/06/09]
373. UniProtKB. P12110 (CO6A2_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/P12110 accessed 20th January 2022.
374. Song F, Zhang X, Ren XB, et al. Cyclophilin A (CyPA) induces chemotaxis independent of its peptidylprolyl cis-trans isomerase activity: direct binding between CyPA and the ectodomain of CD147. *J Biol Chem* 2011;286(10):8197-203. doi: 10.1074/jbc.C110.181347 [published Online First: 2011/01/20]
375. Wei Y, Jinchuan Y, Yi L, et al. Antiapoptotic and proapoptotic signaling of cyclophilin A in endothelial cells. *Inflammation* 2013;36(3):567-72. doi: 10.1007/s10753-012-9578-7 [published Online First: 2012/11/28]
376. UniProtKB. Q15404 (RSU1_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/Q15404 accessed 20th January 2022.
377. Downward J. Targeting RAS signalling pathways in cancer therapy. *Nat Rev Cancer* 2003;3(1):11-22. doi: 10.1038/nrc969 [published Online First: 2003/01/02]
378. Hugle M, Omoumi P, van Laar JM, et al. Applied machine learning and artificial intelligence in rheumatology. *Rheumatol Adv Pract* 2020;4(1):rkaa005. doi: 10.1093/rap/rkaa005 [published Online First: 2020/04/17]
379. MacQuarrie JL, Stafford AR, Yau JW, et al. Histidine-rich glycoprotein binds factor XIIa with high affinity and inhibits contact-initiated coagulation. *Blood* 2011;117(15):4134-41. doi: 10.1182/blood-2010-07-290551 [published Online First: 2011/02/10]
380. Tan F, Weerasinghe DK, Skidgel RA, et al. The deduced protein sequence of the human carboxypeptidase N high molecular weight subunit reveals the presence of leucine-rich tandem repeats. *J Biol Chem* 1990;265(1):13-9. [published Online First: 1990/01/05]
381. UniProtKB. P0C0L4 (CO4A_HUMAN) [updated 29th September 2021. Available from: https://www.uniprot.org/uniprot/P0C0L4 accessed 23rd January 2022.
382. Gorski JP, Hugli TE, Muller-Eberhard HJ. C4a: the third anaphylatoxin of the human complement system. *Proc Natl Acad Sci U S A* 1979;76(10):5299-302. doi: 10.1073/pnas.76.10.5299 [published Online First: 1979/10/01]
383. Keilhack H, Muller M, Bohmer SA, et al. Negative regulation of Ros receptor tyrosine kinase signaling. An epithelial function of the SH2 domain protein tyrosine phosphatase SHP-1. *J Cell Biol* 2001;152(2):325-34. doi: 10.1083/jcb.152.2.325 [published Online First: 2001/03/27]
384. Kremer LS, Danhauser K, Herebian D, et al. NAXE Mutations Disrupt the Cellular NAD(P)HX Repair System and Cause a Lethal Neurometabolic Disorder of Early Childhood. *Am J Hum*

Genet 2016;99(4):894-902. doi: 10.1016/j.ajhg.2016.07.018 [published Online First: 2016/09/13]
385. Fang L, Choi SH, Baek JS, et al. Control of angiogenesis by AIBP-mediated cholesterol efflux. *Nature* 2013;498(7452):118-22. doi: 10.1038/nature12166 [published Online First: 2013/05/31]
386. Luo S, Clarke SLN, Ramanan AV, et al. Platelet Glycoprotein Ib alpha-Chain as a Putative Therapeutic Target for Juvenile Idiopathic Arthritis: A Mendelian Randomization Study. *Arthritis Rheumatol* 2021;73(4):693-701. doi: 10.1002/art.41561 [published Online First: 2020/10/21]
387. He P, Deng FY, Wang BH, et al. Epigenetically-regulated RPN2 gene influences lymphocyte activation and is involved in pathogenesis of rheumatoid arthritis. *Gene* 2022;810:146059. doi: 10.1016/j.gene.2021.146059 [published Online First: 2021/11/07]
388. Breedveld FC, Jones HE, Peifer K, et al. A Pilot Dose-Finding Study of Etanercept in Rheumatoid Arthritis. *Clin Transl Sci* 2018;11(1):38-45. doi: 10.1111/cts.12502 [published Online First: 2017/09/12]
389. Moots RJ, Xavier RM, Mok CC, et al. The impact of anti-drug antibodies on drug concentrations and clinical outcomes in rheumatoid arthritis patients treated with adalimumab, etanercept, or infliximab: Results from a multinational, real-world clinical practice, non-interventional study. *PLoS One* 2017;12(4):e0175207. doi: 10.1371/journal.pone.0175207 [published Online First: 2017/04/28]
390. Chen DY, Chen YM, Tsai WC, et al. Significant associations of antidrug antibody levels with serum drug trough levels and therapeutic response of adalimumab and etanercept treatment in rheumatoid arthritis. *Ann Rheum Dis* 2015;74(3):e16. doi: 10.1136/annrheumdis-2013-203893
391. Sanmarti R, Inciarte-Mundo J, Estrada-Alarcon P, et al. Towards optimal cut-off trough levels of adalimumab and etanercept for a good therapeutic response in rheumatoid arthritis. Results of the INMUNOREMAR study. *Ann Rheum Dis* 2015;74(8):e42. doi: 10.1136/annrheumdis-2015-207530 [published Online First: 2015/03/26]
392. Yang Y, Chung EK, Wu YL, et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am J Hum Genet* 2007;80(6):1037-54. doi: 10.1086/518257 [published Online First: 2007/05/16]
393. Dente L, Pizza MG, Metspalu A, et al. Structure and expression of the genes coding for human alpha 1-acid glycoprotein. *EMBO J* 1987;6(8):2289-96. [published Online First: 1987/08/01]
394. Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 2014;506(7488):376-81. doi: 10.1038/nature12873 [published Online First: 2014/01/07]
395. Tran H, Schilling M, Wirbelauer C, et al. Facilitation of mRNA deadenylation and decay by the exosome-bound, DExH protein RHAU. *Mol Cell* 2004;13(1):101-11. doi: 10.1016/s1097-2765(03)00481-7 [published Online First: 2004/01/21]
396. Rhodes B, Furnrohr BG, Vyse TJ. C-reactive protein in rheumatology: biology and genetics. *Nat Rev Rheumatol* 2011;7(5):282-9. doi: 10.1038/nrrheum.2011.37 [published Online First: 2011/04/07]
397. Cordingley L, Prajapati R, Plant D, et al. Impact of psychological factors on subjective disease activity assessments in patients with severe rheumatoid arthritis. *Arthritis Care Res (Hoboken)* 2014;66(6):861-8. doi: 10.1002/acr.22249 [published Online First: 2013/12/18]

APPENDIX ONE: ADDITIONAL ETHICAL APPROVALS FOR THE BRAGGSS-PD SUB-STUDY

Summary of documents contained in Appendix One:

- Favourable ethical opinion for BRAGGSS Substantial Amendment 17a (BRAGGSS-PD sub-study).
- BRAGGSS-PD patient information leaflet v.1.
- BRAGGSS PD informed consent form v.1.
- Favourable ethical opinion for BRAGGSS Substantial Amendment 17b (combined BRAGGSS prospective arm and BRAGGSS-PD patient information leaflet and consent form).
- Combined BRAGGSS prospective arm and BRAGGSS-PD patient information leaflet v.1.
- Combined BRAGGSS prospective arm and BRAGGSS-PD informed consent form v.1.

NHS
Health Research Authority

North West - Greater Manchester South Research Ethics Committee

3rd Floor, Barlow House
4 Minshull Street
Manchester
M1 3DZ

Please note: This is the favourable opinion of the REC only and does not allow the amendment to be implementedat NHS sites in England until the outcome of the HRA assessment has been confirmed.

12 November

2018 Professor

Anne Barton
Arthritis Research UK
Epidemiology UnitSchool of
Medicine
The University of
Manchester
Oxford Road
Manchester
M13 9PT

Dear Professor Barton

Study title:	**Investigation of genes influencing response to therapy with Etanercept in patients with rheumatoid arthritis and related arthropathies.**
REC reference:	**04/Q1403/37**
Amendment number:	**Amendment 17a**
Amendment date:	**09 October 2018**
IRAS project ID:	**31668**

The above amendment was reviewed by the Sub-Committee in correspondence.

Ethical opinion

The members of the Committee taking part in the review gave a favourable ethical opinion of the amendment on the basis described in the notice of

amendment form and supportingdocumentation.

The Committee agreed that where home visits were to be made a lone worker policy shouldbe confirmed.

The researcher confirmed that a lone worker policy was in place and this was provided.

The Committee was satisfied.

Approved documents

The documents reviewed and approved at the meeting were:

Document	*Version*	*Date*
Notice of Substantial Amendment (non-CTIMP) [Minimal dataset]	Amendment 17a	09 October 2018
Other [Lone Worker Policy]		
Participant information sheet (PIS) [Patient information leaflet]	2	09 October 2018
Research protocol or project proposal	2	09 October 2018

Membership of the Committee

The members of the Committee who took part in the review are listed on the attachedsheet.

Working with NHS Care Organisations

Sponsors should ensure that they notify the R&D office for the relevant NHS care organisation of this amendment in line with the terms detailed in the categorisation emailissued by the lead nation for the study.

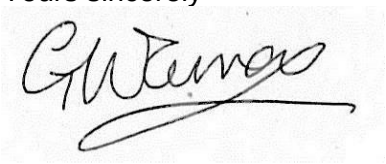
Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures forResearch Ethics Committees in the UK.

We are pleased to welcome researchers and R & D staff at our Research Ethics Committeemembers' training days – see details at http://www.hra.nhs.uk/hra-training/

04/Q1403/37: Please quote this number on all correspondence

Yours sincerely



pp
Professor
Sobhan
VinjamuriChair

PATIENT INFORMATION LEAFLET

Personalising Dosing of Biologics in Rheumatoid Arthritis to Maximise Cost-Benefit

(BRAGGSS-PD)

A Sub Study of the BRAGGSS Study;
Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate

A Sub Study of the BRAGGSS Study;
Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate

You are being invited to participate in the BRAGGSS-PD sub-study. Before you decide whether to take part or not, it is important for you to understand why this research is being done and what it will involve. Please take time to read the following information and ask for anything that is unclear to be explained.

- Part 1 tells you the purpose of the study and what will happen to you if you take part.
- Part 2 gives you more detailed information about the conduct of the study and other useful study information.

Part 1

What is the purpose of the study?

We would like to find out whether clinical and/or psychological factors combined with serological, immune or genetic factors either in blood or joint fluid influence arthritis or treatment response. We have previously found that drug levels in the blood correlate with how well a treatment works, so we also want to explore what influences the drug levels. We aim to determine whether changing the time between each drug dose could potentially

be used in the future to adjust drug levels for each person (personalising the dosing intervals).

Why have I been chosen?

You have been chosen to participate as you are about to start Adalimumab, Etanercept or Certolizumab biologic treatment, or a biosimilar derived from these medications.

Do I have to take part?

You do not have to take part. If you do decide to take part, you can keep this information leaflet and will be asked to sign a consent form. You are still free to stop taking part in the study at any time without giving a reason. Your participation will not interfere with the normal healthcare and treatment that you receive.

What will happen to me if I take part?

	Before you start biologic treatment	**After you've started biologic treatment**					
		1hour	1 week*	2 weeks	4 weeks	6 weeks	3 months
Questionnaire completed by you	X						X
2 extra blood samples	X	X	X	X			
1 extra blood sample					X	X	X
Synovial fluid aspiration *(some participants only)*	X						X

* The one week blood sample is only for participants taking Etanercept

- If you agree to take part in the study, the study doctor will ask you to sign a consent form and arrange for a blood test to be carried out. This blood test would be performed as part of the treatment that you would receive anyway, whether or not you choose to participate in this study, but if you do agree to take part, an additional sample of blood will be taken. You should not need an extra needle for this sample; the extra blood would just be added on to your routine sampling. However, this may require an extra visit if your routine bloods have already been taken.

- The study doctor will ask you questions about your health and you will be asked to complete some questionnaires about your feelings towards your rheumatoid arthritis and your treatment with a biologic therapy.
- If you have swollen joints, your study doctor will ask to perform a joint aspiration. Further information about this procedure is provided on page 4. However, you can still opt not to provide us with synovial fluid if you don't want to. If you would like to provide a synovial fluid sample, the aspiration can either take place on the same day or can be arranged at another point before you start biologic treatment; your study doctor will discuss this with you.
- After processing the sample you donate, we might realise that it doesn't provide enough information to study, or perhaps the blood tubes may be broken or lost during delivery. If this situation arises, we would like permission to be able to contact you to replace the lost/insufficient sample.
- The study doctor will be in touch with you to find out the date you are expected to start your biologic therapy and arrange for the rest of the questionnaires and samples to take place, which are detailed in the table above.

In total, your participation will involve the following in a period of just over 3 months:

(i) Agreement to answer questions about your health, feelings towards both your illness and the biologic treatment at 2 different time points.

(ii) Agreement to provide blood samples at 6 or 7 different time points (depending on which medication you are starting). These would be stored and used to investigate genetic and other factors (including drug levels) that may influence your arthritis and your response to treatment.

(iii) Agreement to link information already collected from your specialist to the results of the genetic studies.

(iv) If you have swollen joints, you will be asked to provide synovial fluid samples twice during the study. If you decide you don't want to have fluid taken from the joint, you can still take part in the rest of the study.

(v) There will be additional hospital visits at the 2 week, 4 week and 6 week timepoints. Travel and parking will be reimbursed for these visits.

(vi) If you are taking Etanercept, there will also be an additional hospital visit 1 week after you start your biologic treatment. Travel and parking will be reimbursed for this visit.

What is serology?

Serology is the study of serum. Serum is the fluid obtained when blood is separated into its solid and liquid components after it has been allowed to clot. It contains antibodies (as well as other proteins), which are made by the immune system to defend against infection. Antibodies are thought to play a role in rheumatoid arthritis.

What is genetics?

Genetics is the study of genes. DNA (deoxyribonucleic acid) is a molecule contained within nearly all our body's cells and it contains genes within it. It is our genes that help to determine certain characteristics, such as hair colour and gender, as well as the likelihood that we will develop certain diseases. Genes differ between people and the purpose of this study is to investigate whether differences in genes affects how people respond to treatment with biologic therapy.

Once a gene is identified, we need to work out how it can make individuals susceptible to developing certain diseases. Although DNA is present in all cells, the genes it codes for are not 'switched on' (expressed) in all tissues. As part of this study, we'll be measuring whether a gene is expressed. We think that genes that make people more susceptible to developing rheumatoid arthritis may be expressed in blood cells, because these are important in the immune system. Knowing whether a gene is expressed in blood cells will help us understand how it might make people more susceptible to developing diseases.

What are white blood cells?

White blood cells make up a large part of our blood. They are one of the main components that respond to being treated with biologic treatments in diseases like rheumatoid arthritis. These cells can be extracted from a blood sample. Looking at them can help us to understand whether their components, which vary from person to person, affect response to biologic treatment.

What is joint (synovial) fluid?

Every joint in the human body contains synovial fluid. The synovial membrane secretes this fluid into the joint cavity. It lubricates the joints and allows it to move easily. The synovial membrane is also the main place where inflammation occurs in joint diseases such as arthritis. Therefore, if you have swollen joints, you will have more synovial fluid than normal.. As part of this study, we are hoping to perform joint aspirations (to extract this fluid) on patients, so we can find out more about your treatment response.

Joint aspiration is a procedure to remove fluid from the space around a joint using a needle and syringe. This is usually done under a local anaesthetic to relieve swelling and/or to obtain fluid for analysis to diagnose a joint disorder or problem. Joint aspiration is most often done on the knee joint and will only be performed if you have swollen joints.

If you don't want to provide us with synovial fluid, that's ok. You can still take part in the rest of the study.

What are the possible benefits of taking part?

We cannot promise that the results of the study will help you, but the information we get might help to improve the treatment of people with rheumatoid arthritis in the future.

Will the research influence the treatment I receive?

The research will not alter the treatment you receive. Your specialist will start and stop treatments as determined by how you are doing.

Unfortunately, we won't be able to provide you with any of the results of this study as you are undergoing your treatment, but we hope to provide you with information about the outcome of the study once it has ended.

Will my taking part in the study be kept confidential?

Yes. All the information about your participation in this study will be kept confidential. The details are included in part 2.

Who is organising and co-ordinating the study?

The study is being co-ordinated by the Arthritis Research UK, Centre for Musculoskeletal Research at the University of Manchester and the lead researcher, Professor Anne Barton, can be contacted for further details (Anne.Barton@manchester.ac.uk). The study doctor is Stephanie Ling (Email: Stephanie.Ling@manchester.ac.uk) and the Study Coordinators are Sarah Ashton and James Anderson (Tel: 0161 276 0539, Email: Sarah.Ashton-2@manchester.ac.uk James.Anderson@manchester.ac.uk).

This completes Part 1 of the information sheet.

If the information in Part 1 has interested you and you are considering participation, please continue to read the additional information in Part 2 before making any decision.

Part 2

What will happen if I don't want to carry on with the study?

You can choose to withdraw from this study at any time, with no impact on your treatment or follow up with your Consultant and rheumatology team. If you do withdraw from the study, we would, with your permission, like to use any samples already donated and the data collected up until your withdrawal.

What will happen to my research data?

The University of Manchester is the sponsor for this study based in the United Kingdom. We will be using information from you and your medical records in order to undertake this study and will act as the data controller for this study. This means that we are responsible for looking after your information and using it properly. The University of Manchester will keep identifiable information about you for 5 years after the study has finished. Other study information will then be completely anonymised and kept for future research.
Paper copies of any study documents will be retained in locked filing cabinets, which will be in turn kept in locked rooms in the Centre for Musculoskeletal Research at the University of Manchester. Information will be entered into the computer and stored electronically on University of Manchester databases which will be password protected and stored in restricted areas.
Your rights to access, change or move your information are limited, as we need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, we will keep the information about you that we have already obtained. To safeguard your rights, we will use the minimum personally-identifiable information possible.
You can find out more about how we use your information by contacting;

- The Chief investigator, Professor Anne Barton, (Anne.Barton@manchester.ac.uk)
- The Study Doctor is Stephanie Ling (Stephanie.Ling@manchester.ac.uk)

- The Study Coordinators are Sarah Ashton and James Anderson (Tel: 0161 276 0539, Sarah.Ashton-2@manchester.ac.uk, James.Anderson@manchester.ac.uk)

The research staff at your hospital will collect information from your medical records for this research study in accordance with our instructions. They will provide your health information to the University of Manchester team along with your name, NHS number and contact details so that the team in Manchester can contact you about the research study at the follow up timepoints. All of the information received at the University of Manchester is regarded as a special category of information.

The research team at your hospital will pass these details to the study coordinator along with the information collected from you and your medical records. The only people at the University of Manchester who will have access to information that identifies you will be people who need to contact you at the follow-up timepoints. The people who analyse the information will not be able to identify you and will not be able to find out your name, NHS number or contact details.

Information provided from the questionnaires you complete, and the clinical information provided by your study doctor, will be combined with the data we get from the blood tests and your synovial fluid. The individuals performing the genetic studies will have no access to any personal identifiable information about you, apart from your year of birth, your gender and the information collected during the research about your response to treatment with a biologic therapy, if you participate.

You will be assigned a unique identification number and the results of the genetic tests will be stored on secure databases within the Arthritis Research UK, Centre for Musculoskeletal Research. Professor Anne Barton will act as custodian of the samples and, therefore, be responsible for protecting this information.

Individuals from the University of Manchester and regulatory organisations, such as the Health Research Authority, may need to look at the data collected for this study to make sure the project is being carried out as planned. This may involve looking at identifiable data, such as your name, but all individuals involved in auditing and monitoring the study, will have a strict duty of confidentiality to you as a research participant.

If you agree to take part in this research study, the information about your health and care may be provided to researchers running other research studies in this organisation and in other organisations. These organisations may be universities, NHS organisations or companies involved in health and care research in this country or abroad. Your information will only be used by organisations and researchers to conduct research in accordance with the UK Policy Framework for Health and Social Care Research. Information that could identify you will not be shared with any other organisation.

What will happen to the samples I give?

- The samples will be considered a gift to the Arthritis Research UK, Centre for Musculoskeletal Research, and stored under the custodianship of the Chief Investigator, Professor Anne Barton.
- The samples you provide will be processed and stored at the Centre for Musculoskeletal Research or at an approved offsite Biobank facility.

- DNA will be extracted from the blood samples.
- The serum sample will be used to look for antibodies and proteins that may be present which may influence the likelihood of someone getting arthritis and influence their treatment response.
- Cells will be extracted from the blood samples to assess whether immune cell patterns predict treatment response.
- DNA, RNA, cells and serum will be made available to other researchers who are undertaking research in this field. This will include researchers in other recognised Institutions, both in the UK and abroad. All of this data will be pseudo-anonymised; no identifiable information will be made available to these researchers.
- The nature of this research is to examine patterns of genes and proteins in large numbers of individuals and no results on your own data will be fed back to you.

What will happen to the results of the research study?

Results of these studies will be published in scientific journals and presented at national and international rheumatology / genetics meetings.

Who is funding the research?

The BRAGGSS study is being funded from a core programme grant from the Arthritis Research UK, Centre for Musculoskeletal Research at The University of Manchester. Additional funding for the Personalising Dosing arm of BRAGGSS has been secured from NIHR Manchester Biomedical Research Centre via a competitive process.

Who has reviewed the study?

Before any research goes ahead it has to be checked by a Research Ethics Committee. They make sure that the research is fair. Your project has been checked by the Greater Manchester South Research Ethics Committee and has received the green light to go ahead.

Thank you for taking time to read this information leaflet, which you should keep for future reference.

Consent Form

Personalising Dosing of Biologics in Rheumatoid Arthritis to Maximise Cost-Benefit (BRAGGSS-PD)

A Sub Study of the BRAGGSS Study;
Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate

Name of Researcher: Professor Anne Barton / Dr. Stephanie Ling

Please initial box

1) I confirm that I have read and understand the information sheet version 1, dated 30/04/2018 for the above study and have had the opportunity to consider the information and ask questions. ☐

2) I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason and without my medical care or legal rights being affected. ☐

3) If I choose to withdraw from the study, I agree that my samples and data collected up to that point will be used in the research. ☐

4) I agree to complete the questionnaires and other survey forms about my health and about my feelings towards my illness and therapy. ☐

5) I agree to provide blood samples from a vein in my arm. These will be used to extract DNA, RNA, plasma, proteins, serum, cells and drug levels. ☐

6) I understand that all of the samples collected from me will be gifted to the Centre for Musculoskeletal Research and that these will be made available to researchers who are undertaking research. Such researchers will receive no identifiable information about me apart from my year of birth, my gender, and the information collected during the research about response to treatment with biologic/targeted therapy drugs if I participate. ☐

7) I agree to have my samples processed and stored at the Centre for Musculoskeletal Research or at an approved offsite research facility for the period of the study, ☐

8) I agree to have my samples stored and used to investigate genetic, clinical, serological and psychological factors involved in arthritis and response to treatment.

9) I agree that my samples (and data created from analysis of my samples) may be provided to other bona-fide researchers working in the field for research purposes (this may include researchers both in this country and abroad). ☐

10) I understand that the nature of the research is to examine patterns of genes and cell types in large numbers of individuals and no results on my own genes will be fed back to me. Similarly, it will not be possible for researchers to inform me of my own disease progression or treatment response. ☐

11) I agree that my study doctor may provide the researchers with information from my Health Records that is relevant to this study. ☐

12) I understand that relevant sections of my data collected during the study, may be looked at by individuals from the University of Manchester, from regulatory authorities or from the NHS Trust, where it is relevant to my taking part in this research. I give permission for these individuals to have access to my data. ☐

13) I agree to information, from which I can be identified, being held securely by the research team at the Arthritis Research UK, Centre for Musculoskeletal Research. ☐

14) I agree to take part in the above study. ☐

15) Synovial Fluid Samples **(if applicable)** ☐

I agree to donate synovial fluid samples to this study at the following timepoints; before I start biologic therapy, and 3 months after I have started my biologic therapy. These samples will be gifted the Centre for Musculoskeletal Research who will store and use these samples in the same way as my blood samples (see points 6, 7, 8, 9 and 10 above)

Name of patient	Date	Signature of patient
Name of person taking consent	Date	Signature

1 copy for the participant, 1 copy for the study file (original), and 1 copy sent to co-ordinating site

Health Research Authority

North West - Greater Manchester South Research Ethics Committee

3rd Floor, Barlow House
4 Minshull Street
Manchester
M1 3DZ

Tel: 0207 104 8063

Please note: This is the favourable opinion of the REC only and does not allow the amendment to be implementedat NHS sites in England until the outcome of the HRA assessment has been confirmed.

04 December 2020

Professor Anne Barton
Arthritis Research UK Epidemiology Unit
School of Medicine
The University of Manchester
Oxford Road
Manchester
M13 9PT

Dear Professor Barton

Study title:	**Investigation of genes influencing response to therapy with Etanercept in patients with rheumatoid arthritis and related arthropathies.**
REC reference:	**04/Q1403/37**
Amendment number:	**SA17b**
Amendment date:	**11 November 2020**
IRAS project ID:	**31668**

The above amendment was reviewed 30 November 2020 by the Sub-Committee in correspondence.

Ethical opinion

The members of the Committee taking part in the review gave a favourable ethical opinion of the amendment on the basis described in the notice of amendment form and supporting documentation.

The Sub Committee raised no ethical issues with the amendment.

Approved documents

The documents reviewed and approved at the meeting were:

Document	*Version*	*Date*	
Completed Amendment Tool [Substantial Amendment 17b amendment tool]	1.2	11 November 2020	
Covering letter on headed paper [SA17b Cover Letter]	1	11 November 2020	
Participant consent form [Prospective and PD combined consent form]	1	11 November 2020	
Participant information sheet (PIS) [Prospective and PD combined information leaflet]	1	11 November 2020	

Membership of the Committee

The members of the Committee who took part in the review are listed on the attached sheet.

Working with NHS Care Organisations

Sponsors should ensure that they notify the R&D office for the relevant NHS care organisation of this amendment in line with the terms detailed in the categorisation email issued by the lead nation for the study.

Amendments related to COVID-19

We will update your research summary for the above study on the research summaries section of our website. During this public health emergency, it is vital that everyone can promptly identify all relevant research related to COVID-19 that is taking place globally. If you have not already done so, please register your study on a public registry as soon as possible and provide the HRA with the registration detail, which will be posted alongside other information relating to your project.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

HRA Learning

We are pleased to welcome researchers and research staff to our HRA Learning Eventsand online learning opportunities– see details at: https://www.hra.nhs.uk/planning-and- improving-research/learning/

IRAS Project ID - 31668:	Please quote this number on all correspondence

Yours sincerely

pp

Professor
Sobhan
VinjamuriChair

E-mail: gmsouth.rec@hra.nhs.uk

Patient Information Sheet

Prospective Arm

Combined with

Personalising Dosing of Biologics in Rheumatoid Arthritis to Maximise Cost-Benefit

(BRAGGSS-PD)

STUDY TITLE: Can clinical, serological, genetic and psychological factors be used to predict response to biologic treatment in rheumatoid arthritis?

You are being invited to participate in the BRAGGSS Prospective Arm research study. Before you decide, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information and ask for anything that is unclear to be explained.

About the research

Who will conduct the research?

Prof. Anne Barton
Centre for Musculoskeletal Research
Division of Musculoskeletal and Dermatological Research
School of Biological Sciences
University of Manchester

Dr. Stephanie Ling
Centre for Musculoskeletal Research
Division of Musculoskeletal and Dermatological Research
School of Biological Sciences
University of Manchester

What is the purpose of the study?

When a person is treated with a biologic/targeted therapy drug for rheumatoid arthritis, information is routinely collected to assess whether these treatments have a greater risk of serious side effects and long term health problems than established treatments. This study is an extension of this to find out whether clinical, serological, genetic and psychological factors influence arthritis or response to treatment.

We also would like to find out whether clinical and/or psychological factors combined with serological, immune or genetic factors either in blood or joint fluid influence arthritis or

treatment response. We have previously found that drug levels in the blood correlate with how well a treatment works, so we also want to explore what influences the drug levels. We aim to determine whether changing the time between each drug dose could potentially be used in the future to adjust drug levels for each person (personalising the dosing intervals).

Who is sponsoring the study?
The University of Manchester is the sponsor for this study based in the United Kingdom. We will be using information from you and your medical records in order to undertake this study and will act as the data controller for this study. This means that we are responsible for looking after your information and using it properly.

Who is funding the research?
The study is being funded from a core programme grant from Versus Arthritis.

Who has reviewed the study?
Before any research goes ahead it has to be checked by a Research Ethics Committee. They make sure that the research is fair.

Will the outcomes of the research be published?
At the end of the trial the results will be analysed and published in recognised medical journals and/or presented at scientific meetings. You will not be identified in any publication about the study.

In addition, a summary of the findings will be made available on request to study participants (Please let your research nurse know if you are interested). You will be able to review a summary of these research findings, along with summaries of other research studies from the Centre for Musculoskeletal Research on our website; https://www.musculoskeletal.manchester.ac.uk/.

What would my involvement be?
What will I be asked to do if I take part?
Your participation will involve the following:
(i) Agreement to answer questions about your health, feelings towards both your illness and the biologic treatment at 4 different time points.
(ii) Agreement to provide DNA, RNA, serum and cells at up to 9 different time points, as shown in the table on the next page. This would be obtained from blood samples, stored and used to investigate genetic factors that may influence arthritis and response to treatment.
(iii) You may be asked to donate additional blood (20ml = roughly 1 table spoon) for the purpose of cell analysis. This request will be dependent on the treatment you are about to receive.
(iv) If you have swollen joints, your study doctor will ask to perform a joint aspiration. Further information about this procedure is provided later in the information leaflet. However, you can still opt not to provide us with synovial fluid if you don't want to. If you would like to provide a synovial fluid sample, the aspiration can either take place on the same day or can be arranged at another point before you start biologic treatment; your study doctor will discuss this with you.
(v) Agreement to link information already collected from your specialist to the results of the genetic studies.

(vi) There will be additional hospital visits at the 2 week, 4 week, 6 week and 12 week time points. Travel and parking will be reimbursed for these visits. Alternatively, some visits could take place at your home. Please discuss this with your research doctor.
(vii) If you are taking Etanercept, there will also be an additional hospital visit 6 days after you start your biologic treatment. Travel and parking will be reimbursed for this visit.
(viii) The 6 and 12 month blood kits will be sent to you at your home address by the team at The University of Manchester. There won't be any additional hospital visits for these. You can take the blood kit to your next scheduled hospital appointment or GP visit so the samples can be taken there.

What do I have to do?

- If you agree to take part in the study, a member of the research team will ask you to sign a consent form and arrange for a blood test to be carried out. This blood test would be performed as part of the treatment that you would receive anyway, regardless of whether or not you choose to participate in this study, but if you do agree to take part, an additional sample of blood will be taken. You should not need an extra needle for this sample; the extra blood would just be added on to your routine sampling. However, this may require an extra visit if your routine bloods have already been taken.
- A member of the research team will ask you questions about your health and you will be asked to complete some questionnaires about your feelings towards your rheumatoid arthritis and your treatment with a biologic therapy.
- The study doctor will be in touch with you to find out the date you are expected to start your biologic therapy and arrange for the rest of the questionnaires and samples to take place, which are detailed in the table above.
- At regular intervals over the course of one year (3 months, 6 months and 12 months after starting treatment), you will be asked similar questions regarding your health and feelings about your disease and treatment. Further blood samples will be taken at these times.
- It may occur that on processing the sample you donated does not provide sufficient yield to analyse, or on occasion blood tubes may be broken/lost during delivery. If this situation arises, we would like permission to be able to contact you to replace the lost/insufficient sample.

This table might help to explain what taking part in the study involves;

	Before you start biologic treatment	**After you've started biologic treatment**							
		1hour	**6 days***	**2 weeks**	**4 weeks**	**6 weeks**	**3 months**	**6 months**	**12 months**
Questionnaire completed by you	X						X	X	X
2 tubes of blood								X''	X''
3 tubes of blood		X	X	X	X	X			
6 tubes of blood	X						X		
Synovial fluid aspiration *(some participants only)*	X						X		

* The day 6 blood sample is only for participants taking Etanercept
'' 6 and 12 month blood kits will be sent to you at home

Why have I been chosen to take part?
You have been chosen to participate as you are about to start Adalimumab, Etanercept or Certolizumab biologic treatment, or a biosimilar derived from these medications.

What will happen to the samples I give?
- The sample(s) will be considered a gift to the Centre for Musculoskeletal Research, and stored under the custodianship of the Chief Investigator, Professor Anne Barton.
- The sample(s) you provide will be processed and stored at the Centre for Musculoskeletal Research or an approved offsite Biobank facility.
- DNA will be extracted from the blood samples
- The serum and plasma samples will be used to look for antibodies that may be present that may predispose to disease and influence treatment response.
- Cells will be extracted from the blood samples to assess whether immune cell patterns predict treatment response.
- DNA, RNA, cells, plasma and serum will be made available to other researchers who are undertaking research in this field. This will include researchers in other recognised Institutions, both in this country and abroad.

What is genetics?
DNA (deoxyribonucleic acid) is a molecule contained within nearly all our body's cells and it contains genes within it. It is our genes that help to determine certain characteristics, such as hair colour and gender, as well as the likelihood that we will develop certain diseases. Genes vary between people and the purpose of this study is to investigate

whether variation in genes affects how people respond to treatment with biologic therapy. Genetics is the study of genes.

Once a gene is identified, we need to work out how it predisposes to disease. Although DNA is present in all cells, the genes it codes for are not 'switched on' (expressed) in all tissues. It is possible to test whether a gene is expressed by measuring RNA (ribonucleic acid). We think that genes that predispose to rheumatoid arthritis may be expressed in blood cells, because these are important in the immune system. Knowing whether a gene is expressed in blood cells will help us understand how it might predispose to disease.

What are cells?
White blood cells make up a large part of our blood. They are one of the main components that respond to being treated with biologic treatments in complex diseases like rheumatoid arthritis. Cells can be extracted from a blood sample and analysed to assess whether the protein components, which vary from person to person, affect response to treatment.
What is serology?
Serum is the fluid obtained when blood is separated into its solid and liquid components after it has been allowed to clot. It contains antibodies, which are made by the immune system to defend against infection. Antibodies are thought to play a role in rheumatoid arthritis. Serology is the study of serum.

What is joint (synovial) fluid?
Every joint in the human body contains synovial fluid. The synovial membrane secretes this fluid into the joint cavity. It lubricates the joints and allows it to move easily. The synovial membrane is also the main place where inflammation occurs in joint diseases such as arthritis. Therefore, if you have swollen joints, you will have more synovial fluid than normal.. As part of this study, we are hoping to perform joint aspirations (to extract this fluid) on patients, so we can find out more about your treatment response.

Joint aspiration is a procedure to remove fluid from the space around a joint using a needle and syringe. This is usually done under a local anaesthetic to relieve swelling and/or to obtain fluid for analysis to diagnose a joint disorder or problem. Joint aspiration is most often done on the knee joint and will only be performed if you have swollen joints.

If you don't want to provide us with synovial fluid, that's ok. You can still take part in the rest of the study.

Do I have to take part?
No. It is up to you to decide whether to join this study. Participation is entirely voluntary. We will describe the study and go through this information sheet. If you agree to take part, we will then ask you to sign a consent form. You are free to withdraw at any time, without giving a reason, and your medical care will not be affected.

What are the possible benefits of taking part?
We cannot promise that the results of the study will help you but the information we get might help to improve the future treatment of people with rheumatoid arthritis.

Will the research influence the treatment I receive?
The research does not alter the treatment you receive. Your specialist will start and stop treatments as determined by your clinical condition.

What will happen if I don't want to carry on with the study?
You can choose to withdraw from this study at any time, with no impact on your treatment or follow up with your Consultant. If you decide to stop taking part, no more tests will be performed as part of the research study.

If you decide to withdraw from the trial, the researcher will seek your permission to retain the information and samples that have been obtained from the start of the trial to the point of withdrawal.

- You can choose to withdraw yourself, your data and blood samples completely. In this case, your data will not be included in our analysis.
- You can choose to withdraw yourself but still allow for your information and blood samples collected until the point of withdrawal to the used. In this case, any information already provided or results from tests already performed on you or your blood samples will continue to be used in the trial.

In the unlikely event of a loss of capacity, the research team would retain your blood samples and data already collected and continue to use them in connection with the purposes for which consent is being sought. This could also include further research after the current project has ended subject to ethical approval.

Data Protection and Confidentiality
What information will you collect about me?
In order to undertake the research project we will need to collect the following personal information/data about you at the University of Manchester:

- Name
- Contact Details
- Date of Birth
- Medical History

This information will be sent to The University of Manchester by your research team at your hospital.

The University of Manchester are collecting and storing this personal information in accordance with the General Data Protection Regulation (GDPR) and Data Protection Act 2018 which legislate to protect your personal information. The legal basis upon which we are using your personal information is "public interest task" and "for research purposes" if sensitive information is collected. For more information about the way we process your personal information and comply with data protection law please see our Privacy Notice for Research Participants;
http://documents.manchester.ac.uk/display.aspx?DocID=37095.

The University of Manchester, as Data Controller for this project takes responsibility for the protection of the personal information that The National Repository study is collecting about you. In order to comply with the legal obligations to protect your personal data the University has safeguards in place such as policies and procedures. All researchers are appropriately trained and your data will be looked after in the following way:
Your consent form will be retained for 10 years at the University of Manchester in either electronic or paper format. In paper format the forms will be stored in a secure way at the University of Manchester, following University regulations. In electronic format the data

will be stored securely within the University of Manchester with access limited to study personnel

What are my rights in relation to the information you will collect about me?
You have a number of rights under data protection law regarding your personal information. For example you can request a copy of the information we hold about you. This is known as a Subject Access Request. If you would like to know more about your different rights, please consult our privacy notice for research and if you wish to contact us about your data protection rights, please email dataprotection@manchester.ac.uk or write to The Information Governance Office, Christie Building, University of Manchester, Oxford Road, M13 9PL. at the University and we will guide you through the process of exercising your rights.

If you would like to know more about your different rights or the way we use your personal information to ensure we follow the law, please consult our Privacy Notice for Research; http://documents.manchester.ac.uk/display.aspx?DocID=37095.
Under what legal basis are you collecting this information?
We are collecting and storing this personal identifiable information in accordance with data protection law which protect your rights. These state that we must have a legal basis (specific reason) for collecting your data. For this study, the specific reason is that it is "a public interest task" and "a process necessary for research purposes".
Will my taking part be kept confidential and my personal identifiable information be protected?

Your participation in the study will be kept confidential to the study team and those with access to your personal information as listed above.

Clinical information about you will be combined with serological and genetic data. The individuals performing the genetic studies will not combine it with any personal identifiable information about you apart from your year of birth, your gender and the health information collected provided by your hospital research nurse. They will not have access to your name.

The University of Manchester is the sponsor for this study based in the United Kingdom. We will be using information from you and your medical records in order to undertake this study and will act as the data controller for this study. This means that we are responsible for looking after your information and using it properly. The University of Manchester will keep identifiable information about you for 10 years after the study has finished.

Your NHS Trust will collect information from you and your medical records for this research study in accordance with our instructions.

Your NHS Trust will use your name, gender, date of birth and contact details to contact you about the research study, and make sure that relevant information about the study is recorded for your care, and to oversee the quality of the study. This is regarded as a special category of information. Individuals from The University of Manchester and regulatory organisations may look at your medical and research records to check the accuracy of the research study. Your NHS Trust team will pass information about your gender and date of birth to The University of Manchester along with the information collected from you and your medical records. The only people in The University of Manchester who will have access

to information that identifies you will be people who need to contact you or your healthcare team regarding your participation in the study or audit the data collection process. The people who analyse the information will not be able to identify you and will not be able to find out your name, NHS number, or contact details.

When you agree to take part in a research study, the information about your health and care may be provided to researchers running other research studies in this organisation. The future research should not be incompatible with this research project and will concern relevant treatments. These organisations may be universities, NHS organisations or companies involved in health and care research in this country or abroad. Your information will only be used by organisations and researchers to conduct research in accordance with the UK Policy Framework for Health and Social Care Research (https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/uk-policy-framework-health-social-care-research/).

This information will not identify you and will not be combined with other information in a way that could identify you. The information will only be used for the purpose of health and care research, and cannot be used to contact you regarding any other matter or to affect your care. It will not be used to make decisions about future services available to you.

What if I have a complaint?
If you wish to make a formal complaint to someone independent of the research team or if you are not satisfied with the response you have gained from the researchers in the first instance then please contact
The Research Ethics Manager, Research Office, Christie Building, The University of Manchester, Oxford Road, Manchester, M13 9PL, by emailing: research.complaints@manchester.ac.uk or by telephoning 0161 275 2674.
If you wish to contact us about your data protection rights, please email dataprotection@manchester.ac.uk or write to The Information Governance Office, Christie Building, The University of Manchester, Oxford Road, M13 9PL at the University and we will guide you through the process of exercising your rights.
You also have a right to complain to the Information Commissioner's Office (https://ico.org.uk/make-a-complaint/), Tel 0303 123 1113.

Contact Details
The study is being co-ordinated by the Arthritis Research UK, Centre for Musculoskeletal Research at the University of Manchester and the lead researcher, Professor Anne Barton, can be contacted for further details (Tel: 0161 275 1638, Fax: 0161 275 5043, Email: Anne.Barton@manchester.ac.uk). The study doctor is Stephanie Ling (Tel: 07392314928, Email: Stephanie.Ling@manchester.ac.uk). Alternatively, please contact the Study Coordinator (Tel: 01613060539/ 07785692979, Email: sarah.ashton-2@manchester.ac.uk).

Thank you for taking time to read the information sheet, which you should keep for future reference.

Combined Consent Form

Can clinical, serological, genetic and psychological factors be used to predict response to biologic treatment in rheumatoid arthritis? (BRAGGSS Prospective Arm)

Personalising Dosing of Biologics in Rheumatoid Arthritis to Maximise Cost-Benefit

(BRAGGSS-PD)

The BRAGGSS Study;

Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate

Name of Researcher: Professor Anne Barton / Dr. Stephanie Ling

	Activities	Initials
1	I confirm that I have read the attached information sheet (**Version 1, Date11/11/2020**) for the above study and have had the opportunity to consider the information and ask questions and had these answered satisfactorily.	
2	I understand that my participation in the study is voluntary and that I am free to withdraw at any time without giving a reason and without detriment to myself. I understand that it will not be possible to remove my data from the project once it has been anonymised and forms part of the data set. I agree to take part on this basis.	
3	If I choose to withdraw from the study, I agree that my samples and data collected up to that point will be used in the research.	
4	I agree to have a blood sample taken for the research purpose as explained to me. These will be used to extract DNA, RNA, plasma, proteins, serum, cells and drug levels. I understand that the research using my sample will be genetic, clinical, serological and psychological research examining the factors involved in arthritis and response to treatment.	
5	I understand that the sponsors of this study may make my blood sample/DNA available to other researchers for future research and that this may include researchers working abroad. I give permission for these individuals to have access to my sample, but not any personal identifying information about me. I offer my blood sample as a gift.	
6	I agree to be contacted by the study team with a request for a further blood sample should my sample(s) be damaged in the post or a low yield of DNA extracted for analysis purposes	

7	I agree to that my samples processed and stored at the Centre for Musculoskeletal Research or at an approved offsite research facility for the period of the study.	
8	I understand that the nature of the research is to examine patterns of genes and cell types in large numbers of individuals and no results on my own genes will be fed back to me. Similarly, it will not be possible for researchers to inform me of my own disease progression or treatment response	
9	I agree that my study doctor may provide the researchers with information from my Health Records that is relevant to this study.	
10	I agree to information, from which I can be identified, being held securely by the research team at the Arthritis Research UK, Centre for Musculoskeletal Research.	
11	I understand that data collected during the study may be looked at by individuals from The University of Manchester or regulatory authorities, where it is relevant to my taking part in this research. I give permission for these individuals to have access to my data.	
12	I agree that any anonymised data collected may be shared with other researchers both at the University of Manchester and at other institutions.	
13	I agree to complete the questionnaires and other survey forms about my health and about my feelings towards my illness and therapy.	
14	I agree that any data collected may be published in anonymous form in academic books, reports or journals.	
15	I agree that the researchers may contact me in future about other research projects (optional).	
16	I agree to take part in this study.	

17) Synovial Fluid Samples **(if applicable)** I agree to donate synovial fluid samples to this study at the following timepoints; before I start biologic therapy, and 3 months after I have started my biologic therapy. These samples will be gifted the Centre for Musculoskeletal Research who will store and use these samples in the same way as my blood samples	

Name of patient	Date	Signature of patient
Name of person taking consent	Date	Signature

1 copy for the participant, 1 copy for the study file (original), and 1 copy sent to co-ordinating site

Data Protection

The personal information we collect and use to conduct this research will be processed in accordance with data protection law as explained in the Participant Information Sheet and the Privacy Notice for Research Participants http://documents.manchester.ac.uk/display.aspx?DocID=37095.

APPENDIX TWO: LONE WORKER RISK ASSESSMENT, POLICY AND TRAINING CERTIFICATION

Summary of documents contained in Appendix Two:

- Lone worker risk assessment for BRAGGSS-PD.
- Lone worker policy for BRAGGSS-PD.
- Certification of lone worker training.

General Risk Assessment Form

Date: (1)	**Assessed by:** (2)	**Checked / Validated* by:** (3)	**Location:** (4)	**Assessment ref no** (5)	**Review date:** (6)

Task / premises: (7)

Risk assessment for home visits as part of the PD arm of the BRAGGSS study

Activity (8)	**Hazard** (9)	**Who might be harmed and how** (10)	**Existing measures to control risk** (11)	**Risk rating** (12)	**Result** (13)
Home visits	Have any risk been identified by any other agency?	Clinician	• For home visits background information on the family is gathered beforehand, a specific risk assessment conducted where necessary. • Where higher risk identified visits not to be conducted alone • Staff own experience and training in recognising signs of aggression and avoiding / de-escalating this. • NHS Conflict Resolution training up-to-date • Clinician will have met participant in advance of the home visit		

Date: (1)	**Assessed by:** (2)	**Checked / Validated* by:** (3)	**Location:** (4)	**Assessment ref no** (5)	**Review date:** (6)
Task / premises: (7) Risk assessment for home visits as part of the PD arm of the BRAGGSS study					

Activity (8)	**Hazard** (9)	**Who might be harmed and how** (10)	**Existing measures to control risk** (11)	**Risk rating** (12)	**Result** (13)
	Are the entrances /exits to the property easily accessible?	Clinician	• Clinician to gather the information relating to this participant from the appropriate healthcare personnel/ the participant themselves to enable them to make a suitable and sufficient assessment of the risks • Clinician to research the property in advance via Google Maps and make visual assessment before entering the property		
	Are there any dangers/hazards associated with the property?	Clinician	• Clinician to gather the information relating to this participant from the appropriate healthcare personnel/ the participant themselves to enable them to make a suitable and sufficient assessment of the risks		

Date: (1)	**Assessed by:** (2)	**Checked / Validated* by:** (3)	**Location:** (4)	**Assessment ref no** (5)	**Review date:** (6)

Task / premises: (7)

Risk assessment for home visits as part of the PD arm of the BRAGGSS study

Activity (8)	**Hazard** (9)	**Who might be harmed and how** (10)	**Existing measures to control risk** (11)	**Risk rating** (12)	**Result** (13)
	Are there pets in the household, are they threatening?	Clinician	• Clinician to gather the information relating to this participant from the appropriate healthcare personnel/ the participant themselves to enable them to make a suitable and sufficient assessment of the risks		
	Transportation of needles, blood and other materials required for the home visit	Clinician	• Travel sized sharps bins to be acquired • Lockable tool box acquired • Kit to be checked by clinician before each visit • Blood samples to be transported in approved padded postal boxes to minimise any damage that could occur to them. • All tubes to be appropriately labelled in advance of the visit		

	Clinician working in unknown environment without colleague support	Clinician	• Agreed schedule –times and location of visits to be known. • Online learning module undertaken • Response procedure in event of overdue contact. • Contact point available in office (including an agreed emergency code phrase) at the start and end of each visit • Ensure colleagues know the mode of transport the clinician is taking and (if relevant) the car details. • Reduce time spent working alone so far as is reasonably practicable. • All staff to be familiar with lone working procedures. • Regular supervision and arrangements for debrief / feedback from clinician • Clinician to complete lone worker training • Contact point to follow Steph via Whatsapp Location Tracker • A locked document to be kept on the shared drive containing the clinicians photo and emergency contact details		

BLANK PAGE

Lone worker policy for BRAGGSS-PD

Ensuring minimal risk to the researcher
Risks can be assessed and identified at the hospital informed consent appointment so Dr. Ling can visit participant's homes safely.

Maintaining contact with a designated person
Researchers will have a work mobile phone. The number will be given to participants (to enable them to cancel/rearrange visits). This work phone will also enable the researcher to contact their CfMR Buddy prior and after each visit. The designated person can also contact the researcher at any point during the visit if they need to for whatever reason.

Safety checks
Researchers will complete electronic safety checks which the designated person will have for each visit the researcher undertakes. This will include the name and contact details of participants, time/location of appointment and expected time of departure (see below).

Date of visit

Appointment time

Expected time of departure

Participant's name

Participant's address

Participants telephone number (Home or Mobile)

Safety procedure
If the researcher has not called the designated person by the expected time of departure from the visit, then the designated person must ring the researcher. If the researcher is safe but the visit is still running the designated person must continue to ring every 10 minutes until the researcher has left the appointment and confirmed they are safe.

If the researcher does not answer, then the designated person must leave a message instructing them to call back within 10 minutes. If the researcher does not call back within 10 minutes, then the designated person must attempt to contact the participant. If the participant doesn't answer then the designated person must contact the emergency services and inform them of the researchers last known whereabouts.

If at any time during the appointment the researcher does not feel safe, they should explain that they need to fetch some paperwork from the car and exit the participants home to phone the designated person to discuss what action to take.

In the event of wanting to report an emergency without raising the alarm of the participants, the researchers should ring the designated person and say 'tell Sid I'm going to be late'. The designated person should then immediately ring the police on 999 and ask for assistance at the participant's address

.

iHasco
training made simple

CPD
REGISTERED PRESENTER
The CPD Certification Service

CERTIFICATE OF ACHIEVEMENT

STEPHANIE LING

The University Of Manchester

LONE WORKING OUT OF THE WORKPLACE

100%

COMPLETED ON: 18/10/2018

Provided by: iHasco
CPD Presenter: #1557
Certificate: #101757/1657618/3150004

ROSPA
Approved E-learning course

APPENDIX THREE: UPDATED BIOCOSHH FOR BRAGGSS FOR INCLUSION OF PATIENTS WITH RESOLVED HEPATITIS B AND C

Safety Services

APPLICATION TO HANDLE BIOLOGICAL MATERIALS & ASSESSMENT OF RISK

(This form is NOT to be used for Genetically Modified Organisms, for which a separate form is used). All applications to handle human materials must also apply for Ethics approval).

Application received:	*(Local BSA use)*	**Project No:**	*(Local BSA use)*

This form **must** be completed when:

- using a biological agent listed on the "Approved List of biological agents" including HG1 micro-organisms
- using any Specified Animal Pathogen listed on SAPO schedule 1
- using biological material which is likely, knowingly or suspected to be contaminated with a biological agent from the Approved List or the SAPO list.
- If working with pathogens, identify any that are listed on ATCSA schedule 5 or COSHH schedule 3 part V, and bring this to the attention of your local BSA/ the University Biological Safety Advisor

NB Your local BSA may also require the completion of this form where the infectious status of the biological material requires detailed consideration (eg tissue of unknown infectivity status, cell cultures and clinical samples).

Sources of assistance and guidance to help you complete this form are available on the Safety Services Biological Materials webpage.

1. School/Division/ **Research Group:**	Centre for Musculoskeletal Research Division of Musculoskeletal and Dermatological Sciences School of Biological Sciences		
2. Principal Investigator: *(title, forename, surname)* ***Employer, if not University:***	Professor Anne Barton	**3. Position:**	Centre Lead, Centre for Musculoskeletal Research
4. Other Investigators: *(title, forename, surname)*	Mohammed Shafi Ahmed Ruairí McErlean	**5. Positions:** *(e.g. academic*	Research technicians

For each person named, state the employer, if not University staff or student	Megan Sutcliffe Stephanie Ling Nisha Nair	*staff, technician, research student, research associate, etc.)*	PhD student Clinical research fellow Research associate
6. Project Title:	Working with human bodily fluids (including DNA and RNA extraction from whole blood) – hepatitis B and C extension for the Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate (BRAGGSS) project		
7. Principal areas where the work will be done: include building, floor & room no's, type of room e.g. cold room, centrifuge room, research lab	1st floor laboratory, AV Hill Building CIGMR blood laboratory, 3rd floor, Stopford Building		
8. Containment level of area(s)	1 (2) 3		
9. Hazard Group of agent(s)	1 2 (3)		
10. Containment level required	1 (2) 3		

11. Brief Summary of Project: *Please write ¼ - ½ page, defining all abbreviations used.*

A variety of human-derived biological sample types are collected, processed, stored and analysed as part of the Centre for Musculoskeletal Research's research portfolio, including blood. Donors are from low-risk populations of patients who have been screened for blood-borne viruses prior to commencing powerful immunosuppressive medication. This includes both Caucasian patients, as well as those from Black and Minority Ethnic groups.

Blood is sourced from hospital patients and is collected using one of two methods:

1. Most commonly, patients are screened for inclusion in the study by clinic nurses and blood drawn alongside routine blood samples. Samples are transported to the laboratory in IATA-approved packaging by post.
2. Patients recruited to the BRAGGSS-Personalised Dosing (PD) sub-study are screened for inclusion by hospital rheumatology teams. Patients are visited at home for phlebotomy by the study doctor in order to ensure that blood is taken when they

receive device training at home by an external provider for their first dose of medication. Samples are transported to the laboratory in IATA-approved packaging by car.

DNA and/or RNA are extracted from the blood and/or buffy coat samples to study genotypes, gene expression and response to medication in patients with rheumatoid arthritis.

Serum, plasma and peripheral blood mononuclear cells are isolated from blood samples and stored for use in a variety of immunoassays internally only.

12. Summary of experimental procedures:

Blood samples are fractionated to allow for storage of serum plasma, cells and buffy coat samples.

DNA extractions are carried out by manual phenol chloroform method from blood, or using the automated Maxwell® 16 system for buffy coat.

Below highlights the potential areas of biological risk from aerosols/sharps/spills in these procedures:

- Removal of blood tube lid: carried out inside class II cabinet for blood fractionation and high-throughput DNA extraction; in addition, safety glasses used for RNA extraction.
- Use of rollers for mixing blood tubes with reagents: blood tube placed inside a 50ml falcon tube before rolling.
- Centrifugation: centrifuge buckets sealed with clip-on/screw-on lids.
- Subsequent pouring off of supernatant: carried out inside a class II cabinet or with safety glasses into a sealable “blood bucket” containing 10% Distel (laboratory disinfectant grade) for disinfection overnight.
- Serum/plasma aliquoting: carried out inside a class II cabinet.
- No use of sharps or glass.
- Small risk of skin puncture from pipette tips.

13. Nature of biological agent (for cell lines state species of origin and how authenticity has been determined):

<u>Human blood for RNA extraction</u>

≤3ml whole blood per sample

≤50 samples extracted per day

<u>Human blood for serum/plasma/cell separation</u>

≤10ml whole blood per sample

≤100 samples processed per day

<u>Human blood for DNA extraction</u>

≤10ml whole blood per sample

≤100 samples extracted per batch (each batch takes 1.5 days)

Human buffy coat for DNA extraction

≤2ml buffy coat per sample

≤16 samples extracted per batch (each batch takes 1 hour)

Maximum 7 batches per day = 112 samples

Samples selected from patients at hospital rheumatology clinics. Nurses/specialist clinical pharmacists screen patients for suitability and blood samples are collected at the same time as routine blood samples where possible. Otherwise, blood samples collected by study doctor on home visit. Patients are screened for blood-borne viruses prior to commencing treatment that is being studied in BRAGGSS-PD, so the study doctor will know in advance if samples are contaminated. Standard laboratory protocol is to treat all samples as if they are contaminated i.e. medium-risk. No patients will be recruited with active HBV/HCV – only with evidence of previous infection e.g. hepatitis B core antibody (anti-HBc) positive + hepatitis B surface antigen (HBsAg) negative.

Potential hazard to humans and/or animals

14. Pathogenicity

Whole blood has the risk of carrying blood-borne viruses; in particular, hepatitis B virus (HBV), hepatitis C virus (HCV) and the human immunodeficiency virus (HIV). These can cause severe disease and death.

However, both HBV and HCV, whilst in Hazard Group 3 (HG3) according to the Health and Safety Executive (HSE) Approved List of Biological Agents, are also stipulated in the same document to present a limited risk of infection for workers because they are not normally infectious by the airborne route. Containment Level 3 (CL3) measures are not necessarily required, but all other aspects of the work should reflect the high standards expected at CL3.

A large retrospective study of organ donors carried out by DE Feo *et al* demonstrated no risk of HBV transmission via heart and kidney transplants from donors who had tested positive for hepatitis B core antibody (anti-HBc), although recipients of livers who had not previously been vaccinated against HBV or had previous infection were at more risk of subsequently developing HBV. Given that a needlestick injury provides an infinitesimal fraction of the dose of a solid-organ transplant, the risk of developing HBV from a needlestick inoculation injury from an anti-HBc positive sample is near to none, particularly if laboratory staff are immunised against HBV, as is mandatory.

Reference: De Feo TM *et al*. Risk of transmission of hepatitis B virus from anti-HBC positive cadaveric organ donors: a collaborative study. Transplant Proc 2005;37(2):1238-9.

15. Epidemiology

0.1% of the UK have HIV (diagnosed and undiagnosed people aged 15-59 years).

0.1-0.5% carry HBV.

0.5-1.0% have chronic HCV.

Reference: Advisory Committee on Dangerous Pathogens. Protection against blood-borne infections in the workplace: HIV and Hepatitis. HSE (1995).

16. Infectious dose

From needle stick injuries, HBV is estimated to have a transmission rate up to 100 times greater than HIV (which is around 0.3%), and 10 times that of HCV. HBV transmission depends on the immune response of the individual.

17. Routes of transmission

Skin puncture by blood-contaminated sharp objects.

Skin lesion e.g. eczema, skin break.

Splashing of mucous membranes i.e. eyes, nose, mouth.

18. Medical data

Health Surveillance

Staff working with human blood should have hepatitis B vaccination through the University Occupational Health Service.

First aid

Puncture wounds should be washed (not scrubbed) with soap and water for several minutes. Pressure above the wound to induce bleeding from the contaminated injury should also be performed.

Mucous membrane exposure

Copious irrigation with tap water, sterile saline or sterile water for several minutes. If necessary, the person who has been potentially inoculated should report to the nearest Emergency Department for treatment, whilst reporting of the incident to Occupational Health should also occur.

Prophylaxis

Post-exposure prophylaxis for HIV is recommended for healthcare workers occupationally exposed to HIV. For HCV, healthcare workers with known exposure should be monitored for seroconversion and referred for medical follow-up if seroconversion occurs. For HBV, workers should be immunised against the virus. The incident should be logged with Occupational Health and follow-up screening should occur.

19. Environmental stability

HIV

Viral load reduces after storage at 4C. Heating blood to 56-60C reduces infectivity to below detectable levels. May remain infectious in dried blood for months.

HBC and HCV

Stable at 37C for 1 hour and at 56C for 30 minutes, but not when heated above 60C. May remain infectious in dried blood for up to 3 weeks.

20. Possible involvement of non-laboratory personnel (e.g. cleaners, security, UG students, visitors)

Unlikely if clean-up and waste disposal procedures are followed. Cleaners are made aware that blood is used in the laboratory.

21. Special containment procedures

Appendix 8 of the HSE document "Management and operation of microbiological containment laboratories" states that the main physical control measures that may not be required are:

- The laboratory does not need to be maintained at negative air pressure because the agents are not transmissible by the airborne route.
- The laboratory does not need to have exhaust air extracted using HEPA filtration. Any work that could give rise to an aerosol of infectious material must be carried out in a microbiological safety cabinet (MSC) or equivalent containment.
- The laboratory does not need to be sealable to permit fumigation because these agents are easily broken down and cannot survive in the environment.

Dispensing with these physical containment measures means that work can take place in a CL2 laboratory, but the other procedural/management measures normally required at CL3 must still be in place:

- It is important to separate work with infected samples from the routine work that may also be carried out in the laboratory to control potential exposure.
- Production of aerosols or droplets should also be considered. However, this particularly project does not include:
    - Working with samples in the infectious and/or transmissive stage of HBV/HCV; patients with HIV will not be recruited.
    - Tissue culture.
    - Passaging HBV/HCV into an intermediate host.
    - Potential means of transmission of a parasite from host to host (including humans).

Use of sharps is not required in any protocol.

Blunt-nosed scissors are used to open sample bags.

Screw or clip-on lids are used on centrifuge buckets to contain tube leaks/breaks.

Blood Vacutainer tubes and buffy coat cryovials are plastic.

Special care is taken when opening Vacutainer tubes of blood to reduce aerosol formation, which is carried out inside a MSC.

Manual phenol chloroform extraction: class II MSC is used for the majority of the protocol for potentially hazardous samples.

RNA extractions: carried out on a drip tray with safety glasses.

Maxwell extractions: involve pipetting buffy coat into reagent cartridges and then peeling off foil seal before being placed inside the machine; carried out on lab bench with safety glasses.

Blood samples are taken from populations not known to be high-risk. However, if a sample is sent to use from a patient known to be carrying a blood-borne viral infection, then the following special containment procedures apply:

Upon receipt, infected blood samples will be logged in then isolated, marked as infected using orange infectious hazard tape and stored in a dedicated “infectious samples” box at -80C. The box will have clear marking on the sides and lid to warn that it contains infected samples. The freezer will also carry a sign stating that some contents may be infected.

For samples that have already been processed and stored prior to discovery of an infection, the following procedures will apply:

All stored samples from the patient, except DNA and RNA, will be identified and marked as infected on our Laboratory Information Management System (LIMS). Blood and buffy coat samples will be relocated to the infected samples box as above. Serum, plasma and cell samples which have been stored in 2C barcoded Matrix tubes will be transferred onto a new Matrix track. This will not involve opening the tubes or any transfer of the biological material itself, just the container tube. The new Matrix rack will then be wrapped with orange, infectious samples hazard tape so that the plate lid cannot be removed without first removing the tape. This will then be returned to -80C storage. The destination freezer will carry a sign stating that some contents may be infected.

22. Are the containment measures (a) in good working order and

(b) on recorded inspection and maintenance programmes?

a) Yes
b) Yes

23. Are the work area, floors and benching suitable and free from defects?

Yes

24. Animal work: Where will this be performed?

N/A

25. Protective clothing and equipment

Howie-style laboratory coats worn at all times. Routine use of nitrile gloves to prevent direct contact with materials. Ecoshield gloves from Stopford Stores conform to EN374-2-2003 standard with an acceptable quality level (AQL) of 0.65.

Safety glasses worn when transferring blood products outside of a class II MSC.

26. Storage & transport arrangements

Vacutainer blood tubes are initially logged into the LIMS (system hosted in the CIGMR blood lab) before being frozen. Plasma and buffy coat are also extracted from EDTA blood tubes into Matrix tubes and cryovials at this stage. Vacutainer blood tubes and buffy coat cryovials are stored in -80C and -40C freezers in secondary containers, in freezer rooms located inside the AV Hill building 1st floor laboratory space. Transport to/from the laboratory is in a secondary container, on dry ice if already frozen.

27. Disinfection & disposal procedures

Blood waste is disinfected 1:1 with 10% Distel overnight, in metal blood waste buckets, before being poured down the sink.

Blood-contaminated consumables waste is disinfected 1:1 with 10% Distel overnight in a beaker, then placed in a blood waste carton; this is sealed and put in a yellow biohazard waste bag before disposal by incineration. All other consumables waste is placed in yellow biohazard waste bags for incineration.

Maxwell extractions only: dilute further to 2.75L per run in a waste beaker/bucket before pouring down the sink (lysis buffer harmful).

28. Immunisation & health surveillance

All laboratory staff and postgraduate students vaccinated against HBV as part of the health surveillance from Occupational Health. All staff have fitness to work certificates from Occupational Health.

29. Environmental monitoring

Not necessary.

30. Emergency procedures

Spills should be mopped up with blue roll and placed in a blood waste carton for incineration. The area should be disinfected with 10% Distel. For large spills, spill kit absorbent sheets can be used.

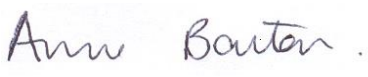
Contaminated clothing should be autoclaved before being laundered.

31. Will the areas be shared by other workers not directly involved in the work? If so who?

Separate areas of the laboratory are used for -80C freezer storage and cell work by staff and postgraduate students in our research group. Some of their work will involve sharing centrifuges.

Other research groups share the laboratory, but use different bays, with their own freezers, centrifuges and waste disposal.
32. If the answer to 31 is yes, how will they be informed of the hazards and risks associated with this work?
All laboratory staff will be shown the risk assessments for the laboratory and be advised of the nature of the hazards/procedures carried out. All staff working regularly in the laboratory (even not directly with blood) are offered HBV vaccinations.

APPROVALS and SIGNATURES

33. I certify that I and all co-workers will (a) sign the reverse of this form to indicate that they are familiar with the contents of this Risk Assessment, (b) will attend appropriate safety courses, (c) carry out the work in accordance with the COSHH regulations 2002 as amended Approved Code of Practice, 5th Edition 2005 and the ACDP guidance document on "The Approved List of biological agents and other relevant legislation and (d) obtain ethical approval where required. **(State University/NRES Ethic Approval number.....)** **Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate (BRAGGSS): 04/Q1403/37** I will submit an updated form if I plan to extend the work outside the areas of risk covered by the present application ***Name of Principal Investigator****(printed):* *Professor Anne Barton*	*Signature* 	*Date* *10/03/2021*
34. I agree with the risk assessment for this project. ***Name of Local BSA*** *(printed):*	*Signature*	*Date*
35. For HG1 and routine HG2 (e.g. clinical samples), with control measures in place –	*Signature*	*Date*

I have agreed to allow the work to proceed in accordance with this assessment. ***Name of Local BSA*** *(printed):*	*If this section is signed, no further signatures or approvals are required.*	
OR		
For all other HG2 work, all HG3 work and forms with any unresolved queries I agree to allow the work to proceed in accordance with this assessment ***Name of Local GM/Bio Safety Committee Chair*** *on behalf of the Local GM/Bio Safety Committee (printed):*	*Signature* *If this section is signed, the Local GM/Bio Committee will communicate whether work can commence, or if notification to the HSE is required (via the UBSA)*	*Date*
Other signatures to be obtained at the discretion of the University BSA, depending on the risk, legal requirements and other relevant factors, eg ATCSA Schedule 5 matters**:**		
36. I agree with this assessment and application ***University BSA***	*Signature*	*Date*
37. I agree with this assessment and application ***Chair of University GM and Biohazards Safety Advisory Group***	*Signature*	*Date*
Office Use		

APPENDIX FOUR: GOOD CLINICAL PRACTICE CERTIFICATION FOR THE DURATION OF THIS PHD

National Institute for Health Research

CERTIFICATE of ACHIEVEMENT

This is to certify that

Stephanie Ling

has completed the course

Introduction to Good Clinical Practice eLearning (Secondary Care)

April 24, 2018

A practical guide to ethical and scientific quality standards in clinical research

Including EU Directives, Medicines for Human Use (Clinical Trials) Regulations & the Department of Health UK Policy Framework for Health & Social Care Research, as applied to the conduct of Clinical Trials & other studies conducted in the NHS

Modules completed:

Introduction to Research and the GCP standards
Preparing to deliver your study
Identifying and recruiting participants: eligibility and informed consent
Ongoing study delivery and data collection
Safety Reporting
Study closure

This course is worth 4 CPD credits

CPD The CPD Certification Service

Delivering research to make patients, and the NHS, better

NIHR | National Institute for Health Research

Issued by:
Clinical Research Network
Coordinating Centre

CERTIFICATE OF ACHIEVEMENT

Stephanie Ling

has completed the course:

Good Clinical Practice (GCP) Refresher eLearning

April 16, 2021

Modules Completed

- Good Clinical Practice (GCP) Refresher: Revisiting Key Concepts
- GCP Refresher Hot Topics
- Good Clinical Practice (GCP) Refresher: Reflecting on your own practice and experience

This course is worth 3 CPD Credits

CPD® The CPD Certification Service

Version: Nov 2020

APPENDIX FIVE: LIST OF PROTEINS INCLUDED IN THE BESPOKE RHEUMATOID ARTHRITIS PROTEIN LIBRARY FOR SWATH-MS EXTRACTION

UniProtKB ID	Protein name	Gene name	Reference
Q04446	1,4-alpha-glucan-branching enzyme	GBE1	Wang 2012
P31946	14-3-3 protein beta/alpha	YWHAB	Giusti 2010
Q04917	14-3-3 protein eta	YWHAH	Giusti 2010
P61981	14-3-3 protein gamma	YWHAG	Giusti 2010
P31947	14-3-3 protein sigma	SFN	Giusti 2010
P63104	14-3-3 protein zeta/delta	YWHAZ	Giusti 2010
P32754	4-hydroxyphenylpyruvate dioxygenase	HPD	Serada 2010
P62841	40S ribosomal protein S15	RSP15	Wang 2012
M0R210	40S ribosomal protein S16	RPS16	Noh 2014
P62081	40S ribosomal protein S7	RPS7	Wang 2012
P08865	40S ribosomal protein SA	RPSA	Noh 2014
P21589	5'-nucleotidase	NT5E	Wang 2012
P52209	6-phosphogluconate dehydrogenase, decarboxylating	PGD	Wang 2012, Yang 2015
P10809	60 kDa heat shock protein, mitochondrial	HSPD1	Schulz 2007
Q02952	A-kinase anchor protein 12	AKAP12	Wang 2012
Q9Y2D5	A-kinase anchor protein 2	AKAP2	Wang 2012
Q5JQC9	A-kinase anchor protein 4	AKAP4	Noh 2014
Q96CW1	AP-2 complex subunit mu-1	AP2M1	Wang 2012
Q01813	ATP-dependent 6-phosphofructokinase, platelet type	PFKP	Wang 2012, Yang 2015
Q92499	ATP-dependent RNA helicase DDX1	DDX1	Wang 2012
P60709	Actin, cytoplasmic 1	ACTB	Schulz 2007, Katano 2009
P63261	Actin, cytoplasmic 2	ACTG1	Schulz 2007, Chang 2009
P07108	Acyl-CoA-binding protein	DB1	Wang 2012
Q15848	Adiponectin	ADIPOQ	Schaffler 2003
P02763	Alpha-1-acid glycoprotein 1	ORM1	Park 2016, Kang 2014

UniProtKB ID	Protein name	Gene name	Reference
P19652	Alpha-1-acid glycoprotein 2	ORM2	Park 2016, Kang 2014
P01011	Alpha-1-antichymotrypsin	SERPINA3	Serada 2010
P01009	Alpha-1-antitrypsin	SERPINA1	Swedlund 1974, Gysen, 1985, Chang 2013
P01023	Alpha-2-macroglobulin	A2M	Cheng 2014
O43707	Alpha-actinin-4	ACTN4	Yang 2018
P35611	Alpha-adducin	ADD1	Wang 2012
P06733	Alpha-enolase	ENO1	Noh 2014
P15144	Aminopeptidase N	ANPEP	Wang 2012
P01019	Angiotensinogen	AGT	Urbaniak 2017
P09525	Annexin A4	ANXA4	Wang 2012
P02647	Apolipoprotein A-I	APOA1	Noh 2014
P02652	Apolipoprotein A-II	APOA2	Yang 2018
P06727	Apolipoprotein A-IV	APOA4	Cheng 2014
P04114	Apolipoprotein B-100	APOB	Mateos 2012
P02654	Apolipoprotein C-I	APOC1	Kim 2018
P02656	Apolipoprotein C-III	APOC3	Cheng 2014
P55056	Apolipoprotein C-IV	APOC4	Cheng 2014
P05090	Apolipoprotein D	APOD	Cheng 2014
P02649	Apolipoprotein E	APOE	Mateos 2012
Q13790	Apolipoprotein F	APOF	Serada 2010
O14791	Apolipoprotein L1	APOL1	Serada 2010
O95445	Apolipoprotein M	APOM	Serada 2010
Q12797	Aspartyl/asparaginyl beta-hydroxylase	ASPH	Wang 2012
Q15121	Astrocytic phosphoprotein PEA-15	PEA15	Wang 2012
P20160	Azurocidin	AZU1	Mateos 2012
P20749	B-cell lymphoma 3 protein	BCL3	KEGG pathway
O95429	BAG family molecular chaperone regulator 4	BAG4	KEGG pathway
Q8N8U9	BMP-binding endothelial regulator protein	BMPER	Katano 2009
Q13490	Baculoviral IAP repeat-containing protein 2	BIRC2	KEGG pathway
Q13489	Baculoviral IAP repeat-containing protein 3	BIRC3	KEGG pathway

UniProtKB ID	Protein name	Gene name	Reference
Q9Y2J2	Band 4.1-like protein 3	EPB41L3	Wang 2012
Q9BXH1	Bcl-2-binding component 3	BBC3	Doroshevskaya 2014
P21810	Biglycan	BGN	Hueber 2009
P54132	Bloom syndrome protein	BLM	Serada 2010
Q92583	C-C motif chemokine 17	CCL17	Cuppen 2017
Q99731	C-C motif chemokine 19	CCL19	Cuppen 2017
P13500	C-C motif chemokine 2	CCL2	Hueber 2007
P78556	C-C motif chemokine 20	CCL20	KEGG pathway
O00626	C-C motif chemokine 22	CCL22	Cuppen 2017
P10147	C-C motif chemokine 3	CCL3	Olszewski 2001
P13501	C-C motif chemokine 5	CCL5	KEGG pathway
P02778	C-X-C motif chemokine 10	CXCL10	KEGG pathway
P02778	C-X-C motif chemokine 10	CXCL10	Hueber 2007
P19875	C-X-C motif chemokine 2	CXCL2	KEGG pathway
P19876	C-X-C motif chemokine 3	CXCL3	KEGG pathway
P42830	C-X-C motif chemokine 5	CXCL5	KEGG pathway
P02741	C-reactive protein	CRP	Cheng 2014
P02741	C-reactive protein	CRP	Kim 2018, Seok 2017
O15519	CASP8 and FADD-like apoptosis regulator	CFLAR	KEGG pathway
Q13166	CATR tumorigenic conversion 1 protein	CATR1	Urbaniak 2017
P17676	CCAAT/enhancer-binding protein beta	CEBPB	KEGG pathway
P29279	CCN family member 2	CCN2	Wang 2012
P14209	CD99 antigen	CD99	Yang 2018
Q9BPX6	Calcium uptake protein 1, mitchondrial	MICU1	Noh 2014
Q05682	Caldesmon	CALD1	Wang 2012
P20810	Calpastatin	CAST	Hueber 2009
Q99439	Calponin 2	CNN2	Wang 2012
O43852	Calumenin	CALU	Wang 2012
P00915	Carbonic anhydrase 1	CA1	Yang 2018
P22792	Carboxypeptidase N subunit 2	CPN2	Obry 2015

UniProtKB ID	Protein name	Gene name	Reference
Q9NQ79	Cartilage acidic protein 1	CRTAC1	Grazio 2013
P49747	Cartilage oligomeric matrix protein	COMP	Chandra 2011
Q92851	Caspase-10	CASP10	KEGG pathway
P42574	Caspase-3	CASP3	KEGG pathway
P55210	Caspase-7	CASP7	KEGG pathway
Q14790	Caspase-8	CASP8	KEGG pathway
P04040	Catalase	CAT	Biemond 1984
P07858	Cathepsin B	CTSB	Liao 2004
Q9UBR2	Cathepsin Z	CTSZ	Serada 2010
Q8WUJ3	Cell migration-inducing and hyaluronan-binding protein	CEMIP	Yang 2015
Q8WUJ3	Cell migration-inducing and hyaluronan-binding protein	CEMIP	Wang 2012
P04637	Cellular tumor antigen p53	TP53	Doroshevskaya 2014
P00450	Ceruloplasmin	CP	Biemond 1984
P36222	Chitinase-3-like protein 1	CHI3L1	Hueber 2005
P06276	Cholinesterase	BCHE	Kim 2018, Seok 2017
O75390	Citrate synthase, mitochondrial	CS	Wang 2012, Yang 2015
P10909	Clusterin	CLU	Cheng 2014
P00740	Coagulation factor IX	F9	Grazio 2013
P00451	Coagulation factor VIII	F8	Serada 2010
P00742	Coagulation factor X	F10	Kim 2018
P02461	Collagen alpha-1(III) chain	COL3A1	Siebert 2017
P27658	Collagen alpha-1(VIII) chain	COL8A1	Siebert 2017
P02452	Collagen alpha1(I) chain	COL1A1	Wang 2012
P20908	Collagen alpha1(V) chain	COL5A1	Wang 2012
P08123	Collagen alpha2(I) chain	COL1A2	Wang 2012
P00736	Complement C1r subcomponent	C1R	Obry 2015
P01024	Complement C3	C3	Noh 2014
P0C0L4	Complement C4-A	C4A	Mateos 2012, Urbaniak 2017
P10643	Complement component C7	C7	Obry 2015
P07357	Complement component C8 alpha chain	C8A	Mateos 2012

UniProtKB ID	Protein name	Gene name	Reference
P07358	Complement component C8 beta chain	C8B	Grazio 2013
P07360	Complement component C8 gamma chain	C8G	Mateos 2012
P02748	Complement component C9	C9	Serada 2010
Q03591	Complement factor H-related protein 1	CFHR1	Cheng 2014
Q9BXR6	Complement factor H-related protein 5	CFHR5	Grazio 2013
P05156	Complement factor I	CFI	Grazio 2013
Q9ULV4	Coronin-1C	CORO1C	Sekigawa 2008
P15336	Cyclic AMP-dependent transcription factor ATF-2	ATF2	KEGG pathway
P18848	Cyclic AMP-dependent transcription factor ATF-4	ATF4	KEGG pathway
Q99941	Cyclic AMP-dependent transcription factor ATF-6 beta	ATF6B	KEGG pathway
P16220	Cyclic AMP-responsive element-binding protein 1	CREB1	KEGG pathway
O43889	Cyclic AMP-responsive element-binding protein 3	CREB3	KEGG pathway
Q96BA8	Cyclic AMP-responsive element-binding protein 3-like protein 1	CREB3L1	KEGG pathway
Q70SY1	Cyclic AMP-responsive element-binding protein 3-like protein 2	CREB3L2	KEGG pathway
Q68CJ9	Cyclic AMP-responsive element-binding protein 3-like protein 3	CREB3L3	KEGG pathway
Q8TEY5	Cyclic AMP-responsive element-binding protein 3-like protein 4	CREB3L4	KEGG pathway
Q02930	Cyclic AMP-responsive element-binding protein 5	CREB5	KEGG pathway
P04080	Cystatin B	CSTB	Wang 2012
P21291	Cysteine and glycine-rich protein 1	CSRP1	Sekigawa 2008
Q16678	Cytochrome P450 1B1	CYP1B1	Wang 2012
P10176	Cytochrome c oxidase subunit 8A	COX8A	Urbaniak 2017
P21399	Cytoplasmic aconitate hydratase	ACO1	Wang 2012, Yang 2015
Q9Y6G9	Cytoplasmic dynein 1 light intermediate chain 1	DYNC1LI1	Noh 2014
O43639	Cytoplasmic protein NCK2	NCK2	Sekigawa 2008
P78527	DNA-dependent protein kinase catalytic subunit	PRKDC	Wang 2012
Q14574	Desmocollin-3	DSC3	Sekigawa 2008
P60981	Destrin	DSTN	Wang 2012
P36957	Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex, mitochondrial	DLST	Wang 2012, Yang 2015
Q5VWQ8	Disabled homolog 2-interacting protein	DAB2IP	KEGG pathway

UniProtKB ID	Protein name	Gene name	Reference
O14672	Disintegrin and metalloproteinase domain-containing protein 10	ADAM10	Wang 2012
Q02750	Dual specificity mitogen-activated protein kinase kinase 1	MAP2K1	KEGG pathway
P46734	Dual specificity mitogen-activated protein kinase kinase 3	MAP2K3	KEGG pathway
P45985	Dual specificity mitogen-activated protein kinase kinase 4	MAP2K4	KEGG pathway
P52564	Dual specificity mitogen-activated protein kinase kinase 6	MAP2K6	KEGG pathway
O14733	Dual specificity mitogen-activated protein kinase kinase 7	MAP2K7	KEGG pathway
O00429	Dynamin-1-like protein	DNM1L	KEGG pathway
Q8TD57	Dynein heavy chain 3, axonemal	DNAH3	Chang 2013
P16581	E-selectin	SELE	KEGG pathway
Q96J02	E3 ubiquitin-protein ligase Itchy homolog	ITCH	KEGG pathway
Q00987	E3 ubiquitin-protein ligase Mdm2	MDM2	Doroshevskaya 2014
Q96C19	EF-hand domain-containing protein D2	EFHD2	Schulz 2007
P11021	Endoplasmic reticulum chaperone BiP	HSPA5	Schulz 2007, Lu 2010
Q9BS26	Endoplasmic reticulum resident protein 44	ERP44	Wang 2012
P05305	Endothelin-1	EDN1	KEGG pathway
P51671	Eotaxin	CCL11	Hueber 2007
P56537	Eukaryotic translation initiation factor 6	EIF6	Wang 2012
P15311	Ezrin	EZR	Wagatsuma 1996
Q13158	FAS-associated death domain protein	FADD	KEGG pathway
Q96M96	FYVE, RhoGEF and PH domain-containing protein 4	FGD4	Katano 2009
Q96AE4	Far upstream element binding protein 1	FUBP1	Wang 2012
Q01469	Fatty acid-binding protein 5	FABP5	Giusti 2010
P02792	Ferritin light chain (ferritin L subunit)	FTL	Chang 2009
P02671	Fibrinogen alpha chain	FGA	Tabushi 2008, Yang 2018
P02675	Fibrinogen beta chain	FGB	Chang 2013, Yang 2018
P02679	Fibrinogen gamma chain	FGG	Yang 2018
Q08830	Fibrinogen-like protein 1	FGL1	Yang 2018
P09038	Fibroblast growth factor 2	FGF2	Hueber 2009
Q9BYJ0	Fibroblast growth factor-binding protein 2	FGFBP2	Yang 2018
Q06828	Fibromodulin	FMOD	Hueber 2009

UniProtKB ID	Protein name	Gene name	Reference
P02751	Fibronectin	FN1	Wang 2012, Tabushi 2008, Kim 2006
Q15485	Ficolin-2	FCN2	Cheng 2014
O75636	Ficolin-3	FCN3	Mateos 2012
P20930	Filaggrin	FLG	Hueber 2005
P21333	Filamin-A	FLNA	Sekigawa 2008
P78423	Fractalkine	CX3CL1	KEGG pathway
P05062	Fructose-bisphosphate aldolase B	ALDOB	Serada 2010
Q92820	Gamma-glutamyl hydrolase	GGH	Wang 2012
P06396	Gelsolin	GSN	Park 2016, Kang 2014
P11413	Glucose-6-phosphate 1-dehydrogenase	G6PD	Yang 2015
O94808	Glutamine--fructose-6-phosphate aminotransferase [isomerizing] 2	GFPT2	Wang 2012, Yang 2015
P22352	Glutathione peroxidase 3	GPX3	Cheng 2014
Q96SL4	Glutathione peroxidase 7	GPX7	Wang 2012
P04406	Glyceraldehyde-3-phosphate dehydrogenase	GAPDH	Grazio 2013
P41250	Glycine--tRNA ligase	GARS	Wang 2012
P04141	Granulocyte-macrophage colony-stimulating factor	CSF2	Olszewski 2001
Q14451	Growth factor receptor-bound protein 7	GRB7	Kim 2006
P09341	Growth-regulated alpha protein	CXCL1	KEGG pathway
O14775	Guanine nucleotide-binding protein subunit beta-5	GNB5	Noh 2014
P04441	H-2 class II histocompatibility antigen gamma chain	CD74	Cuppen 2017
Q8N7B1	HORMA domain-containing protein 2	HORMAD2	Noh 2014
P00738	Haptoglobin	HP	Noh 2014
C7G492	Heat shock protein 90kDa alpha (cytosolic), class A member 1	HSP90AA1	Noh 2014
P04792	Heat shock protein beta-1	HSPB1	Meng 2016
P69905	Hemoglobin subunit alpha	HBA1	Seok 2017
P68871	Haemoglobin subunit beta	HBB	Serada 2010
P61978	Heterogeneous nuclear ribonucleoprotein K	HNRNPK	Schulz 2007
P52926	High mobility group protein HMGI-C	HMGA2	Wang 2012
P04196	Histidine-rich glycoprotein	HRG	Kim 2018
P07305	Histone H1.0	H1F0	Wang 2012

UniProtKB ID	Protein name	Gene name	Reference
P62807	Histone H2B type 1-E	HIST1H2BE	Chandra 2011
O60814	Histone H2B type 1-K	HIST1H2BK	Siebert 2017
Q16778	Histone H2B type 2-E	HIST2H2BE	Hueber 2009
P62805	Histone H4	HIST1H4A	Meng 2016, Mateos 2012
P39880	Homeobox protein cut-like 1	CUX1	Siebert 2017
P10915	Hyaluronan and proteoglycan link protein 1	HAPLN1	Doran 1995
Q9Y4L1	Hypoxia up-regulated protein 1	HYOU1	Wang 2012
P01591	Immunoglobulin J chain	JCHAIN	Grazio 2013
P01871	Immunoglobulin heavy constant mu	IGHM	Kim 2006
P01766	Immunoglobulin heavy variable 3-13	IGHV3-13	Grazio 2013
P01764	Immunoglobulin heavy variable 3-23	IGHV3-23	Grazio 2013
P01763	Immunoglobulin heavy variable 3-48	IGHV3-48	Grazio 2013
P01767	Immunoglobulin heavy variable 3-53	IGHV3-53	Grazio 2013
P01834	Immunoglobulin kappa constant	IGKC	Chang 2009
P0DOY2	Immunoglobulin lambda constant 2	IGLC2	Seok 2017
B9A064	Immunoglobulin lambda-like polypeptide 5	IGLL5	Seok 2017
Q14974	Importin subunit beta-1	KPNB1	Wang 2012
O15111	Inhibitor of nuclear factor kappa-B kinase subunit alpha	CHUK	KEGG pathway
O14920	Inhibitor of nuclear factor kappa-B kinase subunit beta	IKBKB	KEGG pathway
Q15181	Inorganic pyrophosphatase	PPA1	Wang 2012
P05019	Insulin-like growth factor I	IGF1	Serada 2010
A2RTY6	Inter-alpha (Globulin) inhibitor H2	ITIH2	Cheng 2014
P19827	Inter-alpha-trypsin inhibitor heavy chain H1	ITIH1	Grazio 2013
Q06033	Inter-alpha-trypsin inhibitor heavy chain H3	ITIH3	Obry 2015, Serada 2010
P05362	Intercellular adhesion molecule 1	ICAM1	KEGG pathway
P01574	Interferon beta	IFNB1	KEGG pathway
P01579	Interferon gamma	IFNG	Degre 1983
P10914	Interferon regulatory factor 1	IRF1	KEGG pathway
Q8N0X8	Interleukin 12 p40		Hueber 2007
P01583	Interleukin-1 alpha	IL1A	Hueber 2009

UniProtKB ID	Protein name	Gene name	Reference
P01583	Interleukin-1 alpha	IL1A	Hueber 2007
P01584	Interleukin-1 beta	IL1B	Olszewski 2001
Q9NPH3	Interleukin-1 receptor accessory protein	IL1RAP	Serada 2010
P18510	Interleukin-1 receptor antagonist protein	IL1RN	Malyak 1993, Firestein 1992, Olszewski 2001
P14778	Interleukin-1 receptor type 1	IL1R1	Cuppen 2017
P29459	Interleukin-12 subunit alpha	IL12A	Hueber 2009
P29460	Interleukin-12 subunit beta	IL12B	Hueber 2009
P35225	Interleukin-13	IL13	Hueber 2007
P40933	Interleukin-15	IL15	Hueber 2009
Q13478	Interleukin-18 receptor 1	IL18R1	KEGG pathway
P05112	Interleukin-4	IL4	Cuppen 2017
P05231	Interleukin-6	IL6	Olszewski 2001
P13232	Interleukin-7	IL7	Cuppen 2017
P10145	Interleukin-8	CXCL8	Olszewski 2001
P03956	Interstitial collagenase	MMP1	Wang 2012, Mateos 2012, Gysen 1985
A5A6M9	Keratin, hair, acidic, 2	KRTHA2	Noh 2014
P13645	Keratin, type I cytoskeletal 10	KRT10	Noh 2014
P08779	Keratin, type I cytoskeletal 16	KRT16	Noh 2014
Q2M2I5	Keratin, type I cytoskeletal 24	KRT24	Serada 2010
P35527	Keratin, type I cytoskeletal 9	KRT 9	Noh 2014
Q9NSB4	Keratin, type II cuticular Hb2	KRT82	Serada 2010
Q9NSB2	Keratin, type II cuticular Hb4	KRT84	Chang 2013
P04264	Keratin, type II cytoskeletal 1	KRT1	Noh 2014
P35908	Keratin, type II cytoskeletal 2 epidermal	KRT2	Noh 2014
P13647	Keratin, type II cytoskeletal 5	KRT5	Serada 2010
O95678	Keratin, type II cytoskeletal 75	KRT20	Serada 2010
Q86UP2	Kinectin	KTN1	Wang 2012
Q16719	Kynureninase [human]	KYNU	Wang 2012
P14151	L-selectin	SELL	Serada 2010
Q14847	LIM and SH3 domain protein 1	LASP1	Wang 2012

UniProtKB ID	Protein name	Gene name	Reference
P02788	Lactotransferrin	LTF	Meng 2016
Q04760	Lactoylglutathione lyase	GLO1	Wang 2012
Q14766	Latent-transforming growth factor beta-binding protein 1	LTBP1	Sekigawa 2008
P02750	Leucine-rich alpha-2-glycoprotein	LRG1	Seok 2017
P15018	Leukemia inhibitory factor	LIF	KEGG pathway
P30740	Leukocyte elastase inhibitor	SERPINB1	Mateos 2012
P09960	Leukotriene A-4 hydrolase	LTA4H	Serada 2010
Q86UK5	Limbin	EVC2	Serada 2010
P18428	Lipopolysaccharide-binding protein	LBP	Kim 2018
P05451	Lithostathine-1-alpha	REG1A	Sekigawa 2008
O60488	Long-chain-fatty acid--CoA ligase 4	ACSL4	Wang 2012, Yang 2015
P51884	Lumican	LUM	Chang 2013
P01374	Lymphotoxin-alpha	LTA	KEGG pathway
Q14108	Lysosome membrane protein 2	SCARB2	Wang 2012
P09603	Macrophage colony-stimulating factor 1	CSF1	KEGG pathway
P40925	Malate dehydrogenase, cytoplasmic	MDH1	Yang 2015
P48740	Mannan-binding lectin serine protease 1	MASP1	Serada 2010
P50281	Matrix metalloproteinase-14	MMP14	KEGG pathway
P14780	Matrix metalloproteinase-9	MMP9	KEGG pathway
Q96QZ7	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 1	MAGI1	Katano 2009
Q687X5	Metalloreductase STEAP4	STEAP4	Noh 2014
Q9UPN3	Microtubule-actin cross-linking factor 1, isoforms 1/2/3/5	MACF1	Noh 2014
P46821	Microtubule-associated protein 1B	MAP1B	Wang 2012
Q15691	Microtubule-associated protein RP/EB family member 1	MAPRE1	Sekigawa 2008
P20774	Mimecan	OGN	Hueber 2009, Cuppen 2017
P28482	Mitogen-activated protein kinase 1	MAPK1	KEGG pathway
P53779	Mitogen-activated protein kinase 10	MAPK10	KEGG pathway
Q15759	Mitogen-activated protein kinase 11	MAPK11	KEGG pathway
P53778	Mitogen-activated protein kinase 12	MAPK12	KEGG pathway

UniProtKB ID	Protein name	Gene name	Reference
O15264	Mitogen-activated protein kinase 13	MAPK13	KEGG pathway
Q16539	Mitogen-activated protein kinase 14	MAPK14	KEGG pathway
P27361	Mitogen-activated protein kinase 3	MAPK3	KEGG pathway
P45983	Mitogen-activated protein kinase 8	MAPK8	KEGG pathway
P45984	Mitogen-activated protein kinase 9	MAPK9	KEGG pathway
Q99558	Mitogen-activated protein kinase kinase kinase 14	MAP3K14	KEGG pathway
Q99683	Mitogen-activated protein kinase kinase kinase 5	MAP3K5	KEGG pathway
O43318	Mitogen-activated protein kinase kinase kinase 7	MAP3K7	KEGG pathway
P41279	Mitogen-activated protein kinase kinase kinase 8	MAP3K8	KEGG pathway
Q8NB16	Mixed lineage kinase domain-like protein	MLKL	KEGG pathway
P26038	Moesin	MSN	Wagatsuma 1996
P08571	Monocyte differentiation antigen CD14	CD14	Liao 2004, Kang 2014, Park 2016
Q9H8L6	Multimerin-2	MMRN2	Serada 2010
P05164	Myeloperoxidase	MPO	Meng 2016, Baskol 2006
P02144	Myoglobin	MB	Sekigawa 2008
Q9UKX3	Myosin-13	MYH13	Noh 2014
O95167	NADH dehydrogenase (ubiquinone] 1 alpha subcomplex subunit 3	NDUFA3	Katano 2009
Q9Y6K9	NF-kappa-B essential modulator	IKBKG	KEGG pathway
P25963	NF-kappa-B inhibitor alpha	NFKBIA	KEGG pathway
O14745	Na(+)/H(+) exchange regulatory cofactor NHE-RF1	SLC9A3R1	Wang 2012, Schulz 2007
Q13491	Neuronal membrane glycoprotein MG-b	GPM6B	Katano 2009
P0C0P6	Neuropeptide S	NPS	Katano 2009
P59665	Neutrophil defensin 1	DEFA1	Mateos 2012
P80188	Neutrophil gelatinase-associated lipocalin	LCN2	Katano 2009, Mateos 2012
P43490	Nicotinamide phosphoribosyltransferase	NAMPT	Wang 2012
P19838	Nuclear factor NF-kappa-B p105 subunit	NFKB1	KEGG pathway
Q9HC29	Nucleotide-binding oligomerization domain-containing protein 2	NOD2	KEGG pathway
O00151	PDZ and LIMB domain protein 1	PDLIM1	Serada 2010
P04746	Pancreatic alpha-amylase	AMY2A	Sekigawa 2008
P12272	Parathyroid hormone-related protein	PTHLH	Okano 1996

UniProtKB ID	Protein name	Gene name	Reference
Q8TEW0	Partitioning defective 3 homologue	PARD3	Noh 2014
O75594	Peptidoglycan recognition protein 1	PGLYRP1	Sekigawa 2008
P23284	Peptidyl-prolyl cis-trans isomerase B	PPIB	Sekigawa 2008
Q00688	Peptidyl-prolyl cis-trans isomerase FKBP3	FKBP3	Wang 2012
Q06830	Peroxiredoxin-1	PRDX1	Serada 2010
Q13162	Peroxiredoxin-4	PRDX4	Chang 2009
P30044	Peroxiredoxin-5, mitochondrial	PRDX5	Giusti 2010
P30041	Peroxiredoxin-6	PRDX6	Wang 2012
P04180	Phosphatidylcholine-sterol acyltransferase	LCAT	Serada 2010
P27986	Phosphatidylinositol 3-kinase regulatory subunit alpha	PIK3R1	KEGG pathway
O00459	Phosphatidylinositol 3-kinase regulatory subunit beta	PIK3R2	KEGG pathway
Q92569	Phosphatidylinositol 3-kinase regulatory subunit gamma	PIK3R3	KEGG pathway
P42336	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform	PIK3CA	KEGG pathway
P42338	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit beta isoform	PIK3CB	KEGG pathway
O00329	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit delta isoform	PIK3CD	KEGG pathway
P05155	Plasma protease C1 inhibitor	SERPING1	Obry 2015, Serada 2010
P00747	Plasminogen	PLG	Cheng 2014
P13796	Plastin-2	LCP1	Mateos 2012
P02776	Platelet factor 4	PF4	Trocme 2009
Q13093	Platelet-activating factor acetylhydrolase	PLA2G7	Serada 2010
P20742	Pregnancy zone protein	PZP	Mateos 2012
Q9UHG3	Prenylcysteine oxidase 1	PCYOX1	Serada 2010
P01133	Pro-epidermal growth factor	EGF	Fabre 2008
Q14005	Pro-interleukin-16	IL16	Murota 2016
Q03837	Profilaggrin	FLG	Chandra 2011
P46013	Proliferation marker protein Ki-67	MKI67	Noh 2014
Q07954	Prolow-density lipoprotein receptor-related protein 1	LRP1	Wang 2012
P13674	Prolyl 4-hydroxylase subunit alpha-1	P4HA1	Yang 2015

UniProtKB ID	Protein name	Gene name	Reference
Q8NBP7	Proprotein convertase subtilisin/kexin 9	PCSK9	Serada 2010
P35354	Prostaglandin G/H synthase 2	PTGS2	KEGG pathway
P11171	Protein 4.1	EPB41	Noh 2014
Q5T0W9	Protein FAM83B	FAM83B	Noh 2014
P31949	Protein S100-A11	S100A11	Liao 2004
P80511	Protein S100-A12	S100A12	Baillet 2010
P05109	Protein S100-A8	S100A8	Katano 2009, Mateos 2012, Baillet 2010
P06702	Protein S100-A9	S100A9	Katano 2009, Baillet 2010
P78504	Protein jagged-1	JAG1	KEGG pathway
Q9Y2J8	Protein-arginine deiminase type-2	PADI2	De Rycke 2005
Q9UM07	Protein-arginine deiminase type-4	PADI4	Darrah 2017
P21980	Protein-glutamine gamma-glutamyltransferase 2	TGM2	Wang 2012
Q92954	Proteoglycan 4	PRG4	Mateos 2012
P01100	Proto-oncogene c-Fos	FOS	KEGG pathway
P31749	RAC-alpha serine/threonine-protein kinase	AKT1	KEGG pathway
P31751	RAC-beta serine/threonine-protein kinase	AKT2	KEGG pathway
Q9Y243	RAC-gamma serine/threonine-protein kinase	AKT3	KEGG pathway
P50395	Rab GDP dissociation inhibitor beta	GDI2	Noh 2014
P35241	Radixin	RDX	Wagatsuma 1996
P43487	Ran-specific GTPase-activating protein	RANBP1	Schulz 2007
P15153	Ras-related C3 botulinum toxin substrate 2	RAC2	Meng 2016
Q13546	Receptor-interacting serine/threonine-protein kinase 1	RIPK1	KEGG pathway
Q9Y572	Receptor-interacting serine/threonine-protein kinase 3	RIPK3	KEGG pathway
P15927	Replication protein A 32 kDa subunit	RPA2	Schulz 2007
Q9HD89	Resistin	RETN	Schaffler 2003
Q15293	Reticulocalbin-1	RCN1	Wang 2012
P51647	Retinal dehydrogenase 1	ALDH1A1	Chang 2009
	Rheumatoid factor		Chandra 2011
P42331	Rho GTPase-activating protein 25	ARHGAP25	Noh 2014
O75676	Ribosomal protein S6 kinase alpha-4	RPS6KA4	KEGG pathway

UniProtKB ID	Protein name	Gene name	Reference
O75582	Ribosomal protein S6 kinase alpha-5	RPS6KA5	KEGG pathway
P21817	Ryanodine receptor 1	RYR1	Noh 2014
A0A096LPE2	SAA2-SAA4 readthrough	SAA2-SAA4	Cheng 2014
Q8WVM8	Sec 1 family domain-containing protein 1	SCFD1	Chang 2009
P49908	Selenoprotein P	SELENOP	Serada 2010
O75326	Semaphorin-7A	SEMA7A	Kim 2006
Q92743	Serine protease HTRA1	HTRA1	Hueber 2009
Q96HS1	Serine/threonine-protein phosphatase PGAM5, mitochondrial	PGAM5	KEGG pathway
P02787	Serotransferrin	TF	Obry 2015
Q86U17	Serpin A11	SERPINA11	Grazio 2013
P02768	Serum albumin	ALB	Noh 2014
P0DJI8	Serum amyloid A-1 protein	SAA1	Serada 2010
P35542	Serum amyloid A-4 protein	SAA4	Seok 2017
P02743	Serum amyloid P-component	APCS	Cheng 2014
P27169	Serum paraoxonase/arylesterase 1	PON1	Yang 2018
Q15166	Serum paraoxonase/lactonase 3	PON3	Serada 2010
P04278	Sex hormone-binding globulin	SHBG	Yang 2018
P37108	Signal recognition particle 14kDa protein	SRP14	Wang 2012
O00193	Small acidic protein	SMAP	Wang 2012
P55854	Small ubiquitin-related modifier 3	SUMO3	Wang 2012
P05023	Sodium/potassium-transporting ATPase subunit alpha-1	ATP1A1	Wang 2012
Q9Y666	Solute carrier family 12 member 7	SLC12A7	Noh 2014
Q15465	Sonic hedgehog protein	SHH	Yang 2018
Q9Y5X1	Sorting nexin-9	SNX9	Wang 2012
Q13813	Spectrin alpha chain, brain	SPTAN1	Wang 2012
Q01130	Splicing factor, arginine/serine-rich 2	SFRS2	Wang 2012
Q16629	Splicing factor, arginine/serine-rich 7	SFRS7	Wang 2012
P08254	Stromelysin-1	MMP3	Mateos 2012
Q9Y6N5	Sulfide:quinone oxidoreductase, mitochondrial	SQOR	Wang 2012
P00441	Superoxide dismutase [Cu-Zn]	SOD1	Sekigawa 2008

UniProtKB ID	Protein name	Gene name	Reference
P04179	Superoxide dismutase [Mn], mitochondrial	SOD2	Wang 2012, Chang 2009
O14543	Suppressor of cytokine signaling 3	SOCS3	KEGG pathway
Q15750	TGF-beta-activated kinase 1 and MAP3K7-binding protein 1	TAB1	KEGG pathway
Q9NYJ8	TGF-beta-activated kinase 1 and MAP3K7-binding protein 2	TAB2	KEGG pathway
Q8N5C8	TGF-beta-activated kinase 1 and MAP3K7-binding protein 3	TAB3	KEGG pathway
Q12933	TNF receptor-associated factor 2	TRAF2	KEGG pathway
Q13114	TNF receptor-associated factor 3	TRAF3	KEGG pathway
O00463	TNF receptor-associated factor 5	TRAF5	KEGG pathway
Q9Y228	TRAF3-interacting JNK-activating modulator	TRAF3IP3	Noh 2014
Q8NBS9	Thioredoxin domain-containing protein 5	TXNDC5	Chang 2009
P30048	Thioredoxin-dependent peroxidase reductase, mitochondrial	PRDX3	Chang 2009
P07996	Thrombospondin-1	THBS1	Wang 2012
P04216	Thy-1 membrane glycoprotein	THY1	Wang 2012
P19971	Thymidine phosphorylase	TYMP	Wang 2012
Q15560	Transcription elongation factor A protein 2	TCEA2	Noh 2014
P05412	Transcription factor AP-1	JUN	KEGG pathway
P17275	Transcription factor jun-B	JUNB	KEGG pathway
Q04206	Transcription factor p65	RELA	Yamasaki 2001
Q15582	Transforming growth factor-beta-induced protein ig-h3	TGFBI	Mateos 2012
Q01995	Transgelin	TAGLN	Wang 2012
Q92973	Transportin-1	TNPO1	Noh 2014
P02766	Transthyretin	TTR	Yang 2018
P40939	Trifunctional enzyme subunit alpha, mitochondrial	HADHA	Wang 2012, Yang 2015
Q9NP99	Triggering receptor expressed on myeloid cells 1	TREM1	Sekigawa 2008
P60174	Triosephosphate isomerase	TPI1	Chang 2009
O14773	Tripeptidyl-peptidase 1	TPP1	Wang 2012
Q9NYL9	Tropomodulin-3	TMOD3	Sekigawa 2008
P09493	Tropomyosin alpha-1 chain	TPM1	Wang 2012
P67936	Tropomyosin alpha-4 chain	TPM4	Wang 2012
P07951	Tropomyosin beta chain	TPM2	Wang 2012

UniProtKB ID	Protein name	Gene name	Reference
P35030	Trypsin 3	PRSS3	Wang 2012
Q9BQE3	Tubulin alpha-1C chain	TUBA1C	Noh 2014
P07437	Tubulin beta chain	TUBB	Chang 2013
Q86V25	Tubulinyl-Tyr carboxypeptidase 2	VASH2	Noh 2014
P01375	Tumor necrosis factor	TNF	Olszewski 2001
P21580	Tumor necrosis factor alpha-induced protein 3	TNFAIP3	KEGG pathway
P19438	Tumor necrosis factor receptor superfamily member 1A	TNFRSF1A	KEGG pathway
P20333	Tumor necrosis factor receptor superfamily member 1B	TNFRSF1B	Cuppen 2017
P25445	Tumor necrosis factor receptor superfamily member 6	FAS	KEGG pathway
Q15628	Tumor necrosis factor receptor type 1-associated DEATH domain protein	TRADD	KEGG pathway
P54577	Tyrosine--tRNA ligase, cytoplasmic	YARS	Wang 2012
Q16851	UTP--glucose-1-phosphate uridylytransferase	UGP2	Wang 2012
P51965	Ubiquitin-conjugating enzyme E2 E1	UBE2E1	Katano 2009
P68036	Ubiquitin-conjugating enzyme E2 L3	UBE2L3	Wang 2012
P07911	Uromodulin	UMOD	Yang 2018
P26640	Valine--tRNA ligase	VARS	Wang 2012
P19320	Vascular cell adhesion protein 1	VCAM1	KEGG pathway
P49767	Vascular endothelial growth factor C	VEGFC	KEGG pathway
P50552	Vasodilator-stimulated phosphoprotein	VASP	Sekigawa 2008
P49748	Very long-chain specific acetyl-CoA dehydrogenase, mitochondrial	ACADVL	Wang 2012, Yang 2015
O75396	Vesicle-trafficking protein SEC22b	SEC22B	Wang 2012
P08670	Vimentin	VIM	Meng 2016, Noh 2014, Chang 2013, Tabushi 2008
P02774	Vitamin D-binding protein	GC	Yan 2012
P07225	Vitamin K-dependent protein S	PROS1	Obry 2015
P13010	X-ray repair cross-complementing protein 5	XRCC5	Schulz 2007
B2RXF5	Zinc finger and BTB domain-containing protein 42	ZBTB42	Noh 2014
Q14929	Zinc finger protein 169	ZNF169	Noh 2014
Q53GI3	Zinc finger protein 394	ZNF394	Noh 2014
Q8TA94	Zinc finger protein 563	ZNF563	Noh 2014

UniProtKB ID	Protein name	Gene name	Reference
Q5TYW1	Zinc finger protein 658	ZNF658	Noh 2014
Q499Z4	Zinc finger protein 672	ZNF672	Noh 2014
P25311	Zinc-alpha-2-glycoprotein	AZGP1	Obry 2015

Appendix Five References

- Swedlund *et al*. Ann Rheum Dis 1974;33:162-164[274].
- Degré *et al*. Ann Rheum Dis 1983;42:672-6[252].
- Biemond *et al*. Arthritis Rheum 1984[253].
- Gysen *et al*. Clin Rheumatol 1985;4(1):39-50[254].
- Firestein *et al*. J Immunol 1992;149(3):1054-62[266].
- Malyak *et al*. Arthritis Rheum 1993;36(6):781-9[255].
- Doran *et al*. Ann Rheum Dis 1995;54:466-70[273].
- Okano *et al*. Br J Rheumatol 1996;35:1056-62[256].
- Wagatsuma *et al*. Mol Immunol 1996;33(15):1171-6[283].
- Olszewski *et al*. Arthritis Rheum 2001;44(3):541-9[281].
- Yamasaki *et al*. Ann Rheum Dis 2001;60:678-84[267].
- Schäffler *et al*. JAMA 2003;290(13):1709-10[257].
- Forslind *et al*. Ann Rheum Dis 2004;63(9):1090-5[32].
- De Rycke *et al*. Arthritis Rheum 2005;52(8):2323-30[268].
- Hueber *et al*. Arthritis Rheum 2005;52(9):2645-55[75].
- Kim *et al*. J Korean Med Sci 2006;21:478-84[258].
- Baskol *et al*. Cell Biochem Funct 2006;24:307-11[275].
- Hueber *et al*. Ann Rheum Dis 2007;66(6):712-19[291].
- Schulz *et al*. J Proteome Res 2007;6(9):3752-9[278].
- Fabre *et al*. Clin Exp Immunol 2008;153(2):188-95[129].
- Sekigawa *et al*. Clin Exp Rheumatol 2008;26:261-7[126].
- Tabushi *et al*. Ann Clin Biochem 2008;45:413-17[259].
- Chang *et al*. J Rheumatol 2009;36(5):872-80[269].
- Fabre *et al*. Clin Exp Immunol 2009;155(3):395-402[132].
- Hueber *et al*. Arthritis Res Ther 2009;11(3):R76[130].
- Katano *et al*. Arthritis Res Ther 2009;11(1):R3[260].
- Trocme *et al*. Ann Rheum Dis 2009:68(8):1328-33[128].
- Baillet *et al*. Rheumatology 2010;49:671-82[261].
- Giusti *et al*. Proteomics Clin Appl 2010;4:315-24[288].
- Lu *et al*. Arthritis Rheum 2010;62(5):1213-23[279].
- Serada *et al*. Ann Rheum Dis 2010;69:770-4[127].
- Chandra *et al*. Arthritis Res Ther 2011;13:R102[284].
- Shi *et al*. Proc Natl Acad Sci USA 2011;108(42):17372-7[37].
- Mateos *et al*. J Proteomics 2012;75:2869-78[262].
- Wang *et al*. Arthritis Rheum 2012;64(4):993-1004[270].
- Yan *et al*. Clin Exp Rheumatol 2012;30(4):525-33[271].
- Chang *et al*. J Rheum 2013;40(3):219-27[272].
- Grazio *et al*. Clin Exp Rheumatol 2013;31(5):665-71[276].
- Cheng *et al*. Inflammation 2014;37(5):1459-67[292].
- Doroshevskaya *et al*. Bull Exp Biol Med 2014;156(3):377-80[282].
- Kang *et al*. J Proteome Res 2014;13(11):5206-17[293].
- Noh *et al*. J Microbiol Biotechnol 2014;24(1):119-26[263].
- Obry *et al*. Theranostics 2015;5(11):1214-24[131].
- Yang *et al*. Arthritis Res Ther 2015;17:140[290].
- Yang *et al*. Plos One. 2015;10(7):e0132695[264].
- Meng *et al*. Plos One 2016;11(10):e0165501[265].
- Murota *et al*. Cytokine 2016;78:87-93[134].
- Park *et al*. Exp Mol Med 2016;48:e211[294].
- Darrah *et al*. J Proteome Res 2017;16(1):355-65[280].

- Seok *et al*. Molecules 2017;22:805[286].
- Siebert *et al*. Scientific Reports 2017;7:40473[289].
- Urbaniak *et al*. Anal Biochem 2017;525:29-37[285].
- Cuppen *et al*. Scand J Rheumatol 2018;47(1):12-21[137].
- Kim *et al*. Int J Biol Macromol 2018;109:704-10[287].
- Yang *et al*. Clin Rheumatol 2018;37:1773-82[277].
- KEGG pathway: TNF signalling pathway [accessed 17th June 2019][202].

APPENDIX SIX: SIGNIFICANT RESULTS FROM THE CASE-CONTROL PROTEOMICS ANALYSIS

Significantly differentially expressed proteins in patients with RA compared to HC.

UniProt ID	Mean expression RA	Mean expression HC	Difference in means (95% CI)	Standard error	p-value	% missing before imputation	Library
Q99497	13.50	7.06	6.44 (6.41-6.47)	1.40E-02	2.83E-111	72.05	RA
P06733	11.28	2.15	9.13 (9.08-9.19)	2.62E-02	9.28E-104	64.10	RA
Q9Y3I0	12.39	5.00	7.39 (7.35-7.43)	2.18E-02	4.93E-103	67.18	RA
P12956	12.44	4.90	7.54 (7.49-7.58)	2.30E-02	3.99E-102	60.26	RA
Q14012	13.99	6.23	7.76 (7.71-7.81)	2.41E-02	1.31E-101	62.82	RA
Q05682	11.96	4.39	7.57 (7.52-7.62)	2.37E-02	1.84E-101	76.67	RA
P62937	13.07	4.52	8.55 (8.49-8.60)	2.69E-02	2.56E-101	60.77	RA
Q9H4M9	12.83	5.15	7.68 (7.63-7.73)	2.45E-02	6.33E-101	72.82	RA
P05109	15.46	3.40	12.05 (11.95-12.16)	5.19E-02	6.90E-93	34.36	RA
Q15404	11.81	5.41	6.40 (6.34-6.47)	3.35E-02	1.37E-87	58.21	RA
P43490	12.70	3.91	8.80 (8.66-8.93)	6.80E-02	3.94E-77	53.08	RA
P02786	12.87	4.14	8.72 (8.58-8.87)	7.30E-02	5.39E-75	39.23	RA
P08246	11.17	4.82	6.34 (6.21-6.47)	6.52E-02	1.58E-69	62.31	RA
Q8NCW5	10.49	3.78	6.71 (6.50-6.91)	1.03E-01	7.90E-59	65.13	RA
P30043	10.56	4.77	5.77 (5.58-5.96)	9.45E-02	4.22E-57	77.95	RA
P23284	10.14	3.97	6.17 (5.89-6.46)	1.43E-01	7.22E-48	52.31	RA
P20774	8.75	9.72	-0.97 (-1.03-(-0.91))	2.92E-02	2.74E-41	77.69	RA
P02730	10.60	8.85	1.75 (1.60-1.89)	7.17E-02	1.94E-33	82.05	RA
P06703	10.21	8.94	1.26 (1.17-1.36)	4.87E-02	2.66E-29	68.72	RA
P78347	13.03	11.48	1.55 (1.39-1.71)	8.02E-02	3.92E-28	70.51	RA
P06702	11.83	9.45	2.39 (2.12-2.65)	1.32E-01	2.48E-26	43.08	RA
P07900	15.63	14.48	1.14 (1.00-1.29)	8.03E-02	3.23E-26	66.67	RA
Q32MZ4	14.81	13.88	0.92 (0.81-1.04)	5.99E-02	6.68E-25	72.56	RA
Q15084	13.04	11.51	1.52 (1.33-1.72)	9.74E-02	3.34E-23	58.46	RA
O43505	17.12	15.77	1.35 (1.16-1.53)	9.25E-02	1.09E-21	70.26	RA
P12110	13.36	14.72	-1.36 (-1.56-(-1.16))	1.00E-01	1.40E-21	60.77	RA
Q9Y6R7	12.41	11.20	1.21 (1.04-1.39)	8.70E-02	3.14E-21	77.69	RA
P67936	10.70	8.64	2.06 (1.76-2.36)	1.50E-01	8.54E-21	28.97	RA
P06753	15.15	14.10	1.05 (0.86-1.23)	9.13E-02	3.90E-17	71.54	RA
Q13490	8.93	7.72	1.20 (1.01-1.39)	9.38E-02	2.95E-16	73.59	Plasma
Q9P2E9	14.65	13.91	0.74 (0.60-0.87)	6.78E-02	4.09E-16	56.67	RA
Q13136	10.55	14.29	-3.74 (-4.47-(-3.00)	3.69E-01	1.82E-15	33.08	Plasma
P02788	10.92	9.14	1.78 (1.42-1.42)	1.82E-01	1.24E-13	51.03	RA
Q92952	10.31	9.29	1.02 (0.79-1.25)	1.15E-01	2.49E-13	43.08	Plasma
Q9H299	9.81	8.42	1.39 (1.14-1.64)	1.24E-01	7.30E-13	10.51	RA
P12109	12.87	11.46	1.41 (1.09-1.72)	1.59E-01	1.10E-12	58.46	RA
Q15185	14.17	15.72	-155 (-1.92-(-1.18))	1.85E-01	3.29E-12	28.97	RA
P37108	9.94	7.49	2.45 (2.03-2.86)	2.03E-01	3.97E-12	56.92	Plasma
Q14116	13.02	14.80	-1.78 (-2.16-(-1.39))	1.91E-01	4.27E-12	58.46	RA
O00151	12.93	11.18	1.75 (1.33-2.17)	2.09E-01	4.62E-12	40.51	RA
P63104	9.58	8.79	0.79 (0.60-0.99)	9.91E-02	3.05E-11	45.90	RA
Q92851	9.34	10.76	-1.43 (-1.79-(-1.06))	1.83E-01	3.30E-11	27.18	Plasma
Q13045	14.08	14.76	-0.68 (-0.85-(-0.51))	8.62E-02	5.80E-11	71.28	RA

UniProt ID	Mean expression RA	Mean expression HC	Difference in means (95% CI)	Standard error	p-value	% missing before imputation	Library
O15111	11.35	13.11	-1.75 (-2.19-(-1.30))	2.23E-01	6.70E-11	48.46	Plasma
P29144	13.64	14.50	-0.86 (-1.08-(-0.64))	1.11E-01	7.31E-11	38.72	RA
P81605	9.87	8.24	1.63 (1.20-2.05)	2.11E-01	7.43E-11	54.10	RA
P49247	11.02	9.78	1.24 (0.91-1.57)	1.64E-01	9.65E-11	33.08	RA
P62158	13.04	10.69	2.34 (1.72-2.96)	3.11E-01	1.33E-10	5.64	RA
O75347	12.27	13.82	-1.54 (-1.95-(-1.13)	2.05E-01	1.40E-10	35.90	RA
Q00610	14.27	13.30	0.97 (0.72-1.22)	1.26E-01	7.86E-10	57.44	RA
P13796	15.11	15.97	-0.86 (-1.10-(-0.62))	1.15E-01	3.45E-09	48.72	RA
P36222	13.69	15.46	-1.77 (-2.28-(-1.27))	2.51E-01	7.18E-09	24.62	RA
Q12906	10.46	11.10	-0.64 (-0.83-(-0.44)	9.63E-02	7.54E-09	63.08	RA
P07737	12.56	11.07	1.48 (1.04-1.93)	2.23E-01	2.35E-08	4.87	RA
O95168	12.42	9.56	2.86 (2.24-3.49)	2.97E-01	2.47E-08	43.85	Plasma
P04406	15.29	15.96	-0.67 (-0.89-(-0.46)	1.08E-01	3.64E-08	47.69	RA
P14174	10.46	9.54	0.92 (0.71-1.13)	1.00E-01	4.32E-08	71.54	RA
P17980	16.56	15.48	1.08 (0.78-1.38)	1.46E-01	4.63E-08	55.13	RA
O95445	16.30	17.02	-0.72 (-0.94-(-0.50))	1.07E-01	4.74E-08	12.31	RA
Q12797	11.53	13.27	-1.74 (-2.26-(-1.23)	2.53E-01	7.95E-08	41.54	Plasma
Q13740	15.06	15.82	-0.76 (-1.01-(-0.51))	1.24E-01	1.72E-07	50.77	RA
P13797	13.31	11.76	1.55 (1.03-2.07)	2.58E-01	1.90E-07	38.21	RA
P46939	14.50	15.41	-0.92 (-1.18-(-0.65))	1.29E-01	2.07E-07	19.49	RA
P0DJI9	12.50	10.26	2.25 (1.59-2.90)	3.17E-01	2.53E-07	65.90	RA
P46821	11.75	10.09	1.66 (1.06-2.26)	3.01E-01	5.08E-07	58.46	RA
P18428	13.49	12.28	1.21 (0.81-1.60)	1.94E-01	6.35E-07	0.51	RA
Q9NY33	12.38	13.53	-1.16 (-1.46-(-0.85))	1.43E-01	6.51E-07	51.28	RA
Q562R1	11.94	10.77	1.17 (0.82-1.53)	1.70E-01	7.46E-07	1.54	RA
P08603	17.63	17.05	0.59 (0.37-0.80)	1.06E-01	9.08E-07	0.00	RA
P98179	17.50	18.42	-0.92 (-1.24-(-0.60))	1.58E-01	1.15E-06	38.72	RA
Q14766	13.45	12.36	1.09 (0.79-1.40)	1.42E-01	1.39E-06	60.00	RA
P05023	12.78	13.78	-1.00 (-1.35-(-0.65))	1.74E-01	1.66E-06	7.18	Plasma
P04040	8.73	8.25	0.48 (0.29-0.66)	9.22E-02	2.01E-06	61.54	RA
Q86U17	8.04	8.60	-0.56 (-0.78-(-0.35))	1.01E-01	2.09E-06	68.46	Plasma
P21399	16.34	17.32	-0.98 (-1.36-(-0.60))	1.92E-01	2.48E-06	53.33	RA
P01610	10.17	10.55	-0.38 (-0.52-(-0.23))	7.32E-02	2.99E-06	73.85	RA
Q01518	10.14	9.03	1.11 (0.72-1.49)	1.87E-01	3.26E-06	18.46	RA
Q6UX71	9.04	8.78	0.27 (0.16-0.37)	5.30E-02	4.54E-06	68.21	RA
Q02985	9.48	8.52	0.96 (0.64-1.28)	1.52E-01	4.65E-06	66.67	RA
P02741	14.00	11.18	2.82 (1.79-3.84)	4.98E-01	5.95E-06	10.51	RA
P18510	10.21	9.17	1.04 (0.64-1.44)	1.96E-01	8.49E-06	15.90	Plasma
Q05639	15.09	13.85	1.24 (0.85-1.63)	1.84E-01	8.97E-06	73.85	RA
Q5JQC9	10.43	9.33	1.09 (0.65-1.54)	2.21E-01	1.17E-05	34.87	Plasma
P12955	14.20	13.51	0.70 (0.40-0.99)	1.49E-01	1.38E-05	56.67	RA
Q16539	11.61	8.86	2.76 (1.85-3.66)	4.21E-01	1.40E-05	53.59	Plasma
P04264	13.89	12.67	1.22 (0.79-1.64)	2.00E-01	1.48E-05	56.41	RA
Q96AE4	11.92	14.37	-2.45 (-3.37-(-1.53))	4.39E-01	1.90E-05	64.10	Plasma
P26640	13.15	12.51	0.64 (0.39-0.90)	1.25E-01	2.42E-05	65.38	RA
Q9Y283	13.18	12.18	1.00 (0.56-1.45)	2.23E-01	3.11E-05	54.87	Plasma
P02654	15.78	16.94	-1.16 (01.64-(-0.68))	2.35E-01	3.13E-05	0.51	RA
P04207	9.85	10.90	-1.05 (-1.46-(-0.65))	1.91E-01	3.26E-05	77.69	RA

UniProt ID	Mean expression RA	Mean expression HC	Difference in means (95% CI)	Standard error	p-value	% missing before imputation	Library
P05155	17.70	19.35	-1.64 (-2.32-(-0.97))	3.28E-01	4.27E-05	0.00	RA
P63313	19.18	18.55	0.63 (0.34-0.92)	1.44E-01	4.86E-05	75.13	RA
O75083	10.67	6.78	3.89 (2.35-5.43)	7.28E-01	5.51E-05	65.64	RA
P12111	15.76	14.93	0.84 (0.49-1.18)	1.66E-01	6.15E-05	42.05	RA
P05019	8.97	10.31	-1.34 (-1.88-(-0.80))	2.56E-01	7.23E-05	78.97	RA
Q8NBS9	16.10	15.35	0.75 (0.41-1.10)	1.71E-01	7.39E-05	38.21	RA
P13473	9.27	8.53	0.75 (0.41-1.08)	1.62E-01	8.08E-05	28.21	RA
P04114	18.20	18.92	-0.72 (-1.05-(-0.39))	1.61E-01	8.25E-05	0.00	RA
Q9NYU2	9.89	8.23	1.66 (0.97-2.35)	3.27E-01	9.34E-05	69.74	RA
P04075	9.88	11.69	-1.81 (-2.62-(-0.99))	3.94E-01	1.38E-04	84.62	RA
P02655	15.81	17.06	-1.25 (-1.83-(-0.67))	2.83E-01	1.47E-04	0.00	RA
Q01813	7.48	6.39	1.09 (0.56-1.63)	2.64E-01	1.66E-04	26.41	Plasma
Q13263	10.70	10.07	0.62 (0.31-0.93)	1.56E-01	1.91E-04	62.31	RA
Q12933	11.35	10.69	0.66 (0.33-0.99)	1.63E-01	2.21E-04	36.92	Plasma
P00736	15.00	14.47	0.53 (0.27-0.80)	1.30E-01	2.45E-04	0.00	RA
Q9H0W9	17.25	16.49	0.76 (0.37-1.15)	1.95E-01	2.46E-04	49.23	RA
Q9Y446	14.06	12.46	1.59 (0.86-2.33)	3.47E-01	2.70E-04	43.85	RA
Q14204	19.05	18.78	-0.72 (-1.10-(-0.35))	1.88E-01	2.73E-04	25.64	RA
P50454	20.06	19.73	0.33 (0.16-0.50)	8.69E-02	2.79E-04	32.31	RA
Q5VWQ8	7.32	10.93	-3.61 (-5.34-(-1.87))	8.26E-01	3.69E-04	33.33	Plasma
Q15560	10.49	10.71	-0.23 (-0.35-(-0.10))	6.07E-02	4.07E-04	47.18	Plasma
P05451	11.16	10.74	0.42 (0.20-0.63)	1.04E-01	4.26E-04	43.33	RA
Q9BXR6	9.80	9.13	0.68 (0.33-1.02)	1.69E-01	4.78E-04	41.03	RA
P08294	8.70	7.97	0.73 (0.33-1.13)	2.01E-01	5.15E-04	50.77	RA
Q12778	9.70	9.07	0.63 (0.29-0.96)	1.68E-01	5.48E-04	64.10	Plasma
P07384	11.17	11.66	-0.49 (-0.77-(-0.22))	1.37E-01	7.02E-04	54.87	RA
P06727	19.78	18.42	1.36 (0.62-2.10)	3.60E-01	8.00E-04	13.33	RA
Q9HDC9	11.26	11.92	-0.65 (-1.00-(-0.31))	1.68E-01	8.44E-04	57.95	RA
P17948	13.56	13.03	0.53 (0.24-0.83)	1.46E-01	8.80E-04	43.85	RA
P61626	12.48	11.05	1.42 (0.68-2.17)	3.52E-01	9.98E-04	4.10	RA
O00329	9.17	7.36	1.81 (0.80-2.82)	4.87E-01	1.13E-03	38.72	Plasma
Q99460	13.90	12.60	1.30 (0.55-2.04)	3.64E-01	1.28E-03	33.08	RA
P00558	12.05	10.87	1.18 (0.51-1.85)	3.27E-01	1.34E-03	38.21	RA
P09172	9.66	10.40	-0.74 (-1.17-(-0.32))	2.08E-01	1.38E-03	35.64	RA
P22352	13.10	13.68	-0.59 (-0.94-(-0.23))	1.76E-01	1.62E-03	0.00	RA
Q14019	11.20	10.51	0.70 (0.30-1.09)	1.90E-01	1.63E-03	48.46	RA
P20700	12.85	13.74	-0.90 (-1.40-(-0.39))	2.36E-01	1.70E-03	21.03	RA
P01375	11.00	9.46	1.54 (0.68-2.40)	4.00E-01	1.76E-03	40.77	Plasma
P40939	9.73	10.28	-0.55 (-0.88-(-0.21))	1.67E-01	1.75E-03	56.92	RA
O43852	11.38	10.64	0.74 (0.31-1.16)	2.01E-01	1.84E-03	47.69	RA
P27918	12.40	11.46	0394 (0.39-1.49)	2.66E-01	2.05E-03	0.00	RA
P43487	8.96	8.62	0.34 (0.13-0.54)	9.93E-02	2.06E-03	67.18	Plasma
Q9Y450	13.62	14.13	-0.51 (-0.82-(-0.20))	1.51E-01	2.12E-03	19.49	Plasma
P0C0L5	16.09	17.02	-0.93 (-1.50-(-0.36))	2.78E-01	2.18E-03	6.92	RA
P07339	19.85	16.67	3.18 (1.30-5.06)	8.95E-01	2.35E-03	54.62	RA
P45983	10.38	11.16	-0.78 (-1.27-(-0.29))	2.42E-01	2.48E-03	54.36	RA
Q99832	14.73	13.74	0.99 (0.41-1.56)	2.70E-01	2.55E-03	51.28	RA
Q09666	19.61	19.14	0.46 (0.17-0.75)	1.44E-01	2.91E-03	44.36	RA

UniProt ID	Mean expression RA	Mean expression HC	Difference in means (95% CI)	Standard error	p-value	% missing before imputation	Library
P41250	14.54	13.96	0.58 (0.22-0.94)	1.74E-01	3.03E-03	65.13	RA
P13010	9.06	9.92	-0.85 (-1.38-(-0.32))	2.56E-01	3.09E-03	15.13	Plasma
P40926	13.11	12.82	0.29 (0.10-0.49)	9.49E-02	3.09E-03	67.69	RA
P27361	8.88	9.53	-0.65 (-1.07-(-0.24))	2.02E-01	3.20E-03	65.38	Plasma
P06737	13.86	14.17	-0.31 (-0.50-(-0.11))	9.42E-02	3.42E-03	69.74	RA
Q9UKX3	8.99	9.67	-0.68 (-1.12-(-0.24))	2.15E-01	3.45E-03	64.10	Plasma
Q15746	13.05	12.45	0.59 (0.22-0.96)	1.77E-01	3.66E-03	54.36	RA
Q99784	10.30	9.98	0.32 (0.11-0.52)	9.90E-02	4.43E-03	51.28	RA
Q8TEW0	11.80	10.39	1.41 (0.49-2.33)	4.39E-01	4.62E-03	33.59	Plasma
Q04917	9.12	8.17	0.95 (0.33-1.57)	2.98E-01	4.89E-03	72.05	Plasma
Q9BPX6	9.76	10.60	-0.85 (-1.41-(-0.28))	2.73E-01	5.47E-03	42.82	Plasma
P10909	17.89	17.53	0.35 (0.11-0.60)	1.21E-01	5.66E-03	0.00	RA
Q92820	9.99	9.36	0.63 (0.20-1.06)	2.11E-01	5.72E-03	32.56	RA
P05546	18.27	17.85	0.42 (0.13-0.71)	1.44E-01	5.75E-03	0.00	RA
Q14CX7	12.61	11.15	1.46 (0.49-2.43)	4.55E-01	5.81E-03	48.97	Plasma
Q9UGM5	12.22	11.60	0.62 (0.19-1.06)	2.13E-01	6.61E-03	1.79	RA
P54578	14.18	14.54	-0.36 (-0.61-(-0.11))	1.23E-01	6.79E-03	57.69	RA
Q9HC38	13.54	14.11	-0.57 (-0.99-(-0.16))	2.06E-01	6.83E-03	41.28	RA
P23396	10.91	9.28	1.63 (0.52-2.75)	5.25E-01	6.96E-03	21.79	RA
P01776	18.57	17.08	1.50 (0.44-2.55)	5.06E-01	7.65E-03	57.44	RA
P05160	11.78	11.35	0.42 (0.1-0.72)	1.37E-01	7.69E-03	56.92	RA
P02656	16.44	17.38	-0.94 (-1.61-(-0.28))	3.22E-01	7.80E-03	0.00	RA
P14780	10.33	9.69	0.64 (0.18-1.09)	2.26E-01	7.96E-03	26.67	RA
P27169	16.15	16.73	-0.58 (-0.99-(-0.16))	1.97E-01	9.05E-03	0.00	RA
Q7Z3U7	18.80	18.00	0.80 (0.21-1.39)	2.87E-01	9.68E-03	8.21	RA
Q8WVM8	15.68	13.50	2.18 (0.61-3.75)	7.35E-01	9.90E-03	34.10	Plasma
P41222	10.00	10.44	-0.43 (-0.75-(-0.11))	1.53E-01	1.03E-02	72.31	RA
P31749	9.64	9.15	0.49 (0.12-0.85)	1.75E-01	1.08E-02	67.44	Plasma
P52272	16.46	16.24	0.21 (0.05-0.38)	8.19E-02	1.10E-02	68.46	RA
P01011	20.94	20.40	0.53 (0.13-0.93)	1.92E-01	1.11E-02	0.00	RA
Q7KZF4	16.88	15.53	1.34 (0.34-2.35)	4.76E-01	1.13E-02	24.36	RA
P19838	15.14	14.13	1.01 (0.25-1.78)	3.75E-01	1.15E-02	10.77	Plasma
P00918	8.68	8.18	0.50 (0.12-0.88)	1.86E-01	1.19E-02	19.49	RA
Q13201	12.78	12.25	0.53 (0.13-0.94)	1.95E-01	1.19E-02	63.85	RA
Q03591	12.05	10.81	1.24 (0.31-2.17)	4.38E-01	1.20E-02	11.54	RA
Q8NB16	11.47	11.15	0.32 (0.08-0.57)	1.17E-01	1.22E-02	26.67	Plasma
P02748	17.03	16.47	0.56 (0.13-0.99)	2.07E-01	1.23E-02	0.00	RA
P26599	11.69	12.16	-0.47 (-0.83-(-0.18))	1.82E-01	1.25E-02	58.21	RA
P11766	16.66	16.40	0.26 (0.06-0.46)	9.78E-02	1.37E-02	56.67	RA
P52566	14.53	13.67	0.86 (0.18-1.54)	3.31E-01	1.48E-02	52.56	RA
O43143	11.77	13.14	-1.37 (-2.45-(-0.30))	5.01E-01	1.57E-02	57.69	RA
P22897	12.35	12.94	-0.59 (-1.05-(-0.13))	2.17E-01	1.60E-02	43.33	RA
P80188	10.65	10.86	-0.21 (-0.38-(-0.04))	8.40E-02	1.62E-02	76.15	RA
P48740	11.02	11.60	-0.58 (-1.05-(-0.11))	2.29E-01	1.73E-02	6.15	RA
P03951	12.02	11.00	1.02 (0.20-1.84)	3.91E-01	1.75E-02	2.82	RA
P15144	12.58	11.79	0.80 (0.16-1.44)	3.00E-01	1.75E-02	41.28	RA
P42765	12.00	11.66	0.34 (0.06-0.63)	1.36E-01	2.13E-02	45.38	RA
P02776	12.66	12.15	0.51 (0.08-0.94)	2.10E-01	2.14E-02	8.46	RA

UniProt ID	Mean expression RA	Mean expression HC	Difference in means (95% CI)	Standard error	p-value	% missing before imputation	Library
P0DJI8	11.21	9.88	1.33 (0.21-2.46)	5.45E-01	2.25E-02	26.92	RA
O60716	11.38	10.27	1.10 (0.17-2.04)	4.47E-01	2.37E-02	37.69	RA
P02452	11.31	10.67	0.64 (0.10-1.18)	2.59E-01	2.39E-02	61.79	RA
Q86VB7	10.32	9.79	0.54 (0.07-1.00)	2.23E-01	2.46E-02	32.05	RA
Q01668	9.82	9.20	0.62 (0.09-1.16)	2.48E-01	2.51E-02	61.79	Plasma
P30048	15.17	12.28	2.89 (0.42-5.36)	1.15E-01	2.53E-02	13.59	RA
O75369	16.46	16.84	-0.38 (-0.71-(-0.05)	1.56E-01	2.57E-02	52.82	RA
P02743	16.78	16.26	0.52 (0.07-0.97)	2.17E-01	2.62E-02	0.26	RA
P46734	10.73	10.41	0.31 (0.04-0.59)	1.37E-01	2.63E-02	68.97	Plasma
Q86VP6	14.37	14.83	-0.46 (-0.87-(-0.01))	1.95E-01	2.94E-02	48.72	RA
P07359	13.26	12.66	0.60 (0.07-1.14)	2.59E-01	2.95E-02	8.21	RA
P05090	17.27	17.65	-0.38 (-0.72-(-0.04))	1.69E-02	2.97E-02	0.00	RA
P02763	19.57	18.90	0.68 (0.07-1.28)	2.97E-01	3.04E-02	0.00	RA
P62701	13.40	13.08	0.32 (0.03-0.61)	1.39E-01	3.30E-02	12.56	RA
P02753	17.33	17.70	-0.37 (-0.70-(-0.03))	1.65E-01	3.35E-02	0.00	RA
P04180	13.23	12.99	0.24 (0.02-0.47)	1.10 E-01	3.36E-02	0.00	RA
P04275	14.45	14.60	-0.15 (-0.28-(-0.01))	6.48E-02	3.38E-02	54.87	RA
P10809	15.50	16.02	-0.53 (-1.01-(-0.04))	2.33E-01	3.44E-02	48.21	RA
P49913	10.98	10.45	0.53 (0.04-1.02)	2.38E-01	3.64E-02	20.25	RA
P04004	18.83	18.45	0.37 (0.02-0.72)	1.68E-01	3.68E-02	0.00	RA
P49908	11.27	11.70	-0.42 (-0.82-(-0.03))	1.94E-01	3.69E-02	6.92	RA
P09871	15.62	15.29	0.32 (0.02-0.63)	1.51E-01	4.05E-02	0.00	RA
P68871	17.82	16.92	0.90 (0.04-1.75)	4.09E-01	4.09E-02	0.00	RA
P09104	16.42	16.57	-0.15 (-0.29-(-0.00))	7.20E-02	4.32E-02	55.90	RA
Q9UNW1	11.33	10.96	0.37 (0.01-0.73)	1.71E-01	4.39E-02	53.33	RA
Q7Z794	11.75	12.13	-0.38 (-0.74-(-0.01))	1.83E-01	4.43E-02	41.03	RA
P49327	17.85	17.39	0.45 (0.01-0.90)	2.13E-01	4.50E-02	18.97	RA
P08697	16.25	16.69	-0.44 (-0.87-(-0.01))	2.05E-01	4.68E-02	0.00	RA
P58107	15.13	15.40	-0.27 (-0.54-(-0.00))	1.32E-01	4.82E-02	64.36	RA

APPENDIX SEVEN: SIGNIFICANT PROTEINS ASSOCIATED WITH RA DISEASE OUTCOMES

Table 1. Proteins measured before treatment with etanercept significantly associated with DAS28 at baseline, univariate analysis.

Protein (UniProt ID)	β-coefficient (95% CI)	p-value	Adjusted p-value	% missing before imputation
P01011	0.25 (0.13 – 0.37)	0.0001	0.0206	0.00
P46734	0.21 (0.10 – 0.33)	0.0003	0.0206	68.97
P02741	0.74 (0.35 – 1.14)	0.0003	0.0206	10.51
P18428	0.28 (0.13 – 0.44)	0.0004	0.0206	0.51
P30048	0.56 (0.24 – 0.87)	0.0007	0.0283	13.59
P02763	0.37 (0.16 – 0.57)	0.0008	0.0290	0.00
P17948	0.21 (0.09 – 0.33)	0.0010	0.0294	43.85
P02748	0.22 (0.09 – 0.35)	0.0012	0.0305	0.00
Q32MZ4	-0.18 (-0.28 – (-0.07))	0.0013	0.0305	72.56
P54578	0.16 (0.06 – 0.26)	0.0014	0.0305	57.69
P00558	-0.39 (-0.63 – (-0.15))	0.0016	0.0305	38.21
P0DJI8	0.63 (0.24 – 1.02)	0.0017	0.0305	26.92
Q04917	-0.30 (-0.49 – (-0.12))	0.0018	0.0305	72.05
Q05682	-0.50 (-0.81 – (-0.18))	0.0022	0.0347	76.67
Q99460	-0.44 (-0.81 – (-0.18))	0.0030	0.0437	33.08

ABBREVIATIONS: Confidence interval (CI), Disease Activity Score of 28 Joints (DAS28), identifier (ID).

Table 2. Proteins measured before treatment with etanercept significantly associated with DAS28 at baseline, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	β-coefficient$_{adj}$ (95% CI)	p-value	Adjusted p-value	% missing before imputation
P01011	0.24 (0.12 – 0.37)	0.0002	0.0260	0.00
P46734	0.21 (0.10 – 0.33)	0.0003	0.0260	68.97
P02741	0.74 (0.34 – 1.13)	0.0004	0.0260	10.51
P18428	0.28 (0.12 – 0.43)	0.0006	0.0273	0.51
P30048	0.55 (0.24 – 0.87)	0.0007	0.0273	13.59
P02763	0.36 (0.15 – 0.58)	0.0009	0.0273	0.00
Q32MZ4	-0.18 (-0.29 – (-0.08))	0.0010	0.0273	72.56
P0DJI8	0.65 (0.26 – 1.03)	0.0013	0.0273	26.92
P17948	0.20 (0.08 – 0.33)	0.0013	0.0273	43.85
Q04917	-0.31 (-0.50 – (-0.12))	0.0014	0.0273	72.05
P54578	0.16 (0.06 – 0.26)	0.0015	0.0273	57.69
P00558	-0.40 (-0.64 – (-0.16))	0.0016	0.0273	38.21
P02748	0.21 (0.08 – 0.34)	0.0017	0.0273	0.00
Q05682	-0.50 (-0.81 – (-0.19))	0.0018	0.0273	76.67
Q99460	-0.44 (-0.73 – (-0.15))	0.0031	0.0443	33.08
P02786	-0.36 (-0.60 – (-0.12))	0.0036	0.0489	39.23

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), Disease Activity Score of 28 Joints (DAS28), identifier (ID), rheumatoid arthritis (RA).

Table 3. Proteins measured before treatment with etanercept significantly associated with DAS28 at baseline, adjusted for age, multivariable model.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**% missing before imputation**
P01011	0.05 (-0.32 – 0.42)	0.7961	0.00
P46734	0.36 (0.09 – 0.64)	0.0110	68.97
P02741	0.03 (-0.06 – 0.12)	0.5016	10.51
P18428	0.05 (-0.18 – 0.29)	0.6588	0.51
P30048	0.01 (-0.09 – 0.12)	0.8252	13.59
P02763	0.07 (-0.10 – 0.25)	0.4044	0.00
Q32MZ4	-0.16 (-0.49 – 0.17)	0.3389	72.56
P0DJI8	0.09 (0.00 – 0.17)	0.0496	26.92
P17948	0.14 (-0.15 – 0.43)	0.3462	43.85
Q04917	-0.18 (-0.32 – (-0.04))	0.0123	72.05
P54578	0.02 (-0.31 – 0.35)	0.8922	57.69
P00558	-0.04 (-0.17 – 0.09)	0.5602	38.21
P02748	-0.41 (-0.82 – 0.01)	0.0579	0.00
Q05682	0.03 (-0.09 – 0.14)	0.6765	76.67
Q99460	0.00 (-0.10 – 0.11)	0.9487	33.08
P02786	-0.15 (-0.26 – (-0.04))	0.0080	39.23
Age at baseline	-0.01 (-0.02 – 0.01)	0.3087	N/A
Male sex	-0.01 (-0.33 – 0.31)	0.9632	N/A
Disease duration	0.01 (-0.00 – 0.03)	0.0568	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), Disease Activity Score of 28 Joints (DAS28), identifier (ID).

Table 4. Proteins measured before treatment with etanercept significantly associated with DAS28 after three months of treatment, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
Q9H4M9	0.32 (0.17 – 0.47)	6.53E-05	0.0117	72.82
Q99832	0.11 (0.05 – 0.16)	0.0001	0.0017	51.28

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), Disease Activity Score of 28 Joints (DAS28), identifier (ID).

Table 5. Proteins measured before treatment with etanercept significantly associated with DAS28 at three months, multivariable model.

Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**% missing before imputation**
Q9H4M9	0.25 (0.09 – 0.41)	0.0030	72.82
Q99832	0.67 (0.21 – 1.13)	0.0050	51.28
Age at baseline	0.02 (0.00 – 0.04)	0.0158	N/A
Male sex	-0.30 (-0.81 – 0.20)	0.2373	N/A
Disease duration	-0.01 (-0.04 – 0.01)	0.1787	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), Disease Activity Score of 28 Joints (DAS28).

Table 6. Proteins measured after three months of treatment with etanercept associated with DAS28 at three months, univariate model.

Protein (UniProt ID)	β-coefficient (95% CI)	p-value	Adjusted p-value	% missing before imputation
Q9H4M9	0.41 (0.27 – 0.56)	1.47E-07	3.20E-05	72.82
P02741	0.54 (0.34 – 0.74)	3.49E-07	3.80E-05	10.51
Q12797	-0.26 (-0.36 – (-0.15))	2.32E-06	0.0002	41.54
P18428	0.19 (0.11 – 0.28)	7.85E-06	0.0004	0.51
Q02985	0.10 (0.05 – 0.14)	5.10E-05	0.0017	66.67
P02763	0.26 (0.13 – 0.38)	5.57E-05	0.0017	0.00
P05019	-0.17 (-0.25 – (-0.09))	5.67E-05	0.0017	78.97
P01375	-0.18 (-0.26 – (-0.09))	0.0001	0.0038	40.77
P0DJI9	0.14 (0.07 – 0.22)	0.0003	0.0058	65.90
P78347	-0.14 (-0.22 – (-0.07))	0.0003	0.0058	70.51
Q12906	-0.09 (-0.14 – (-0.04))	0.0003	0.0058	63.08
P43490	-0.39 (-0.59 – (-0.18))	0.0003	0.0059	53.08
P06737	0.09 (0.04 – 0.14)	0.0005	0.0080	69.74
P08603	0.12 (0.05 – 0.18)	0.0006	0.0094	0.00
P0DJI8	0.28 (0.12 – 0.44)	0.0007	0.0094	26.92
P30043	-0.17 (-0.27 – (-0.07))	0.0007	0.0094	77.95
P36222	-0.24 (-0.38 – (-0.10))	0.0014	0.0178	24.62
Q14CX7	-0.26 (-0.42 – (-0.10))	0.0018	0.0221	48.97
P02748	0.13 (0.05 – 0.21)	0.0020	0.0226	0.00
Q9Y446	-0.20 (-0.33 – (-0.07))	0.0021	0.0230	43.85
P31749	-0.13 (-0.21 – (-0.05))	0.0023	0.0232	67.44
P12956	-0.23 (-0.37 – (-0.08))	0.0025	0.0238	60.26
P02743	0.14 (0.05 – 0.23)	0.0025	0.0238	0.26
Q00610	0.10 (0.03 – 0.16)	0.0029	0.0256	57.44
P10809	-0.11 (-0.18 – (-0.04))	0.0030	0.0256	48.21
P04075	-0.18 (-0.31 – (-0.06))	0.0031	0.0259	84.62
P01776	-0.32 (-0.52 – (-0.11))	0.0033	0.0264	57.44
Q8NB16	0.06 (0.02 – 0.10)	0.0035	0.0267	26.67
Q15185	-0.20 (-0.33 – (-0.07))	0.0041	0.0308	28.97
P29144	-0.21 (-0.35 – (-0.07))	0.0046	0.0329	38.72
Q9HDC9	0.10 (0.03 – 0.17)	0.0048	0.0336	57.95
Q99832	0.07 (0.02 – 0.13)	0.0053	0.0359	51.28
P62937	-0.24 (-0.42 – (-0.07))	0.0058	0.0378	60.77
P06753	0.10 (0.03 – 0.17)	0.0067	0.0424	71.54
Q6UX71	0.05 (0.01 – 0.08)	0.0074	0.0455	68.21
O95445	-0.10 (-0.17 – (-0.03))	0.0079	0.0476	12.31

ABBREVIATIONS: Confidence interval (CI), Disease Activity Score of 28 Joints (DAS28), identifier (ID).

Table 7. Proteins measured after three months of treatment with etanercept associated with DAS28 at three months, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P02741	0.55 (0.35 – 0.76)	3.51E-07	5.50E-05	10.51
Q9H4M9	0.41 (0.25 – 0.56)	5.09E-07	5.50E-05	72.82
P18428	0.20 (0.11 – 0.28)	1.01E-05	0.0007	0.51
Q12797	-0.24 (-0.34 – (-0.13))	1.64E-05	0.0009	41.54
P05019	-0.17 (-0.26 – (-0.09))	7.48E-05	0.0031	78.97
Q02985	0.10 (0.05 – 0.14)	8.72E-05	0.0031	66.67
P01375	-0.17 (-0.27 – (-0.08))	0.0003	0.0077	40.77
P02763	0.23 (0.11 – 0.36)	0.0003	0.0077	0.00
P43490	-0.38 (-0.59 – (_0.17))	0.0006	0.0122	53.08
P08603	0.12 (0.05 – 0.19)	0.0006	0.0122	0.00
Q12906	-0.09 (-0.13 – (-0.04))	0.0006	0.0122	63.08
P30043	-0.18 (-0.28 – (-0.08))	0.0007	0.0125	77.95
P06737	0.09 (0.04 – 0.14)	0.0009	0.0150	69.74
P0DJI8	0.27 (0.11 – 0.43)	0.0011	0.0164	26.92
Q14CX7	-0.27 (-0.44 – (-0.11))	0.0013	0.0181	48.97
P02743	0.15 (0.06 – 0.24)	0.0014	0.0189	0.26
P36222	-0.24 (-0.39 – (-0.10))	0.0015	0.0192	24.62
P0DJI9	0.12 (0.05 – 0.20)	0.0017	0.0200	65.90
P78347	-0.12 (-0.20 – (-0.05))	0.0020	0.0229	70.51
Q8NB16	0.07 (0.02 – 0.11)	0.0024	0.0250	26.67
P31749	-0.13 (-0.22 – (-0.05))	0.0024	0.0250	67.44
Q9Y446	-0.19 (-0.32 – (-0.06))	0.0039	0.0367	43.85
P01776	-0.31 (-0.53 – (-0.10))	0.0039	0.0367	57.44
P12956	-0.21 (-0.36 – (-0.07))	0.0049	0.0438	60.26
P02748	0.12 (0.04 – 0.20)	0.0053	0.0449	0
Q92820	0.15 (0.05 – 0.25)	0.0054	0.0449	32.56
Q00610	0.09 (0.03 – 0.16)	0.0056	0.0449	57.44
P0C0L5	0.17 (0.05 – 0.29)	0.0058	0.0449	6.92

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), Disease Activity Score of 28 Joint (DAS28), identifier (ID), rheumatoid arthritis (RA).

Table 8. Proteins measured after three months of treatment with etanercept significantly associated with DAS28 at three months, multivariable model.

Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**% missing before imputation**
P12956	0.05 (-0.16 – 0.26)	0.6174	60.26
Q9H4M9	0.17 (0.02 – 0.32)	0.0239	72.82
P43490	0.04 (-0.11 – 0.19)	0.4923	53.08
P30043	-0.22 (-0.47 – 0.03)	0.0851	77.95
P78347	-0.10 (-0.50 – 0.31)	0.6364	70.51
Q00610	0.08 (-0.27 – 0.43)	0.6456	57.44
P36222	-0.08 (-0.26 – 0.10)	0.3849	24.62
Q12906	-0.41(-0.90 – 0.09)	0.1097	63.08
Q12797	-0.28 (-0.53 – (-0.04))	0.0249	41.54
P0DJI9	0.18 (-0.15 – 0.50)	0.2895	65.90
P18428	0.06 (-0.27 – 0.40)	0.7128	0.51
P08603	0.14 (-0.33 – 0.61)	0.5605	0.00
Q02985	0.64 (0.10 – 1.17)	0.0209	66.67
P02741	0.12 (-0.02 – 0.26)	0.1055	10.51
P05019	-0.28 (-0.54 – (-0.01))	0.0423	78.97
Q9Y446	-0.15 (-0.33 – 0.04)	0.1182	43.85
P01375	-0.23 (-0.47 – (-0.00))	0.0507	40.77
P0C0L5	-0.05 (-0.24 – 0.15)	0.6424	6.92
P06737	-0.13 (-0.61 – 0.35)	0.6002	69.74
Q92820	0.06 (-0.18 – 0.30)	0.6327	32.56
Q14CX7	-0.11 (-0.29 – 0.07)	0.2342	48.97
P01776	0.13 (-0.06 – 0.31)	0.1805	57.44
P31749	-0.14 (-0.41 – 0.13)	0.3166	67.44
Q8NB16	0.06 (-0.48 – 0.60)	0.8259	26.67
P02748	-0.37 (-0.74 – 0.01)	0.0587	0.00
P0DJI8	0.04 (-0.13 – 0.21)	0.6599	26.92
P02743	-0.14 (-0.47 – 0.19)	0.4162	0.26
P02763	0.19 (-0.07 – 0.45)	0.1558	0.00
Age at baseline	0.01 (-0.01 – 0.02)	0.5393	N/A
Male sex	-0.16 (-0.62 – 0.30)	0.5005	N/A
Disease duration	0.00 (-0.02 – 0.02)	0.7589	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), Disease Activity Score of 28 Joints (DAS28).

Table 9. Proteins measured after three months of treatment with etanercept associated with DAS28 at six months, univariate analysis.

Protein (UniProt ID)	**β-coefficient (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P02741	0.43 (0.25 – 0.61)	7.32E-06	0.0011	10.51
Q12797	-0.21 (-0.30 – (-0.12))	9.81E-06	0.0011	41.54
P0DJI8	0.28 (0.14 – 0.43)	0.0001	0.0100	26.92
P02748	0.14 (0.07 – 0.21)	0.0002	0.0131	0.00
P18428	0.14 (0.07 – 0.22)	0.0003	0.0151	0.51
P17980	0.11 (0.05 – 0.17)	0.0004	0.0153	55.13
Q02985	0.08 (0.03 – 0.12)	0.0006	0.0190	66.67
P62158	-0.43 (-0.68 – (-0.18))	0.0008	0.0216	5.64
P18510	0.16 (0.06 – 0.26)	0.015	0.0366	15.90
P01375	-0.13 (-0.21 – (-0.05))	0.0018	0.0386	40.77

ABBREVIATIONS: Confidence interval (CI), Disease Activity Score of 28 Joints (DAS28), identifier (ID).

Table 10. Proteins measured after three months of treatment with etanercept associated with DAS28 at six months, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P02741	0.41 (0.23 – 0.60)	2.27E-05	0.0049	10.51
Q12797	-0.20 (-0.29 – (-0.10))	6.67E-05	0.0072	41.54
P18428	0.14 (0.07 – 0.22)	0.0004	0.0294	0.51
P02748	0.13 (0.06 – 0.20)	0.0007	0.0294	0.00
P17980	0.11 (0.05 – 0.17)	0.0007	0.0294	55.13
P0DJI8	0.25 (0.11 – 0.40)	0.0008	0.0294	26.92
Q02985	0.07 (0.03 – 0.12)	0.0012	0.0355	66.67

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), Disease Activity Score of 28 Joints (DAS28), identifier (ID), rheumatoid arthritis (RA).

Table 11. Proteins measured after three months of treatment with etanercept significantly associated with DAS28 at six months, multivariable model.

Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**% missing before imputation**
P17980	0.35 (-0.05 – 0.75)	0.0868	55.13
Q12797	-0.34 (-0.59 – (-0.10))	0.0076	41.54
P18428	0.18 (-0.20 – 0.56)	0.3492	0.51
Q02985	0.72 (0.15 – 1.28)	0.0145	66.67
P02741	0.05 (-0.12 – 0.21)	0.5921	10.51
P02748	-0.05 (-0.45 – 0.36)	0.8235	0.00
P0DJI8	0.16 (-0.01 – 0.33)	0.0684	26.92
Age at baseline	0.01 (-0.01 – 0.03)	0.2920	N/A
Male sex	-0.22 (-0.76 – 0.33)	0.4344	N/A
Disease duration	-0.01 (-0.03 – 0.01)	0.3240	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), Disease Activity Score of 28 Joints (DAS28).

Table 12. Proteins measured after three months of treatment with etanercept associated with ΔDAS28 at three months, univariate analysis.

Protein (UniProt ID)	**β-coefficient (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
Q9H4M9	-0.37 (-0.50 – (-0.24))	7.15E-08	1.50E-05	72.82
P02741	-0.36 (-0.54 – (-0.17))	0.0002	0.0201	10.51
P06737	-0.08 (-0.12 – (-0.04))	0.0004	0.0276	69.74
P0DJI8	-0.24 (-0.38- (-0.10))	0.0010	0.0356	26.92
Q12906	0.07 (0.03 – 0.11)	0.0010	0.0356	63.08
Q02985	-0.07 (-0.11 – (-0.03))	0.0011	0.0356	66.67
Q12797	0.16 (0.06 – 0.25)	0.0012	0.0356	41.54
P78347	0.11 (0.04 – 0.18)	0.0013	0.0356	70.51

ABBREVIATIONS: Change in Disease Activity Score of 28 Joints after treatment (ΔDAS28), confidence interval (CI), identifier (ID).

Table 13. Proteins measured after three months of treatment with etanercept associated with ΔDAS28 at three months, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
Q9H4M9	-0.37 (-0.50 – (-0.23))	3.05E-07	6.60E-05	72.82
P02741	-0.35 (-0.54 – (-0.17))	0.0003	0.0308	10.51

ABBREVIATIONS: Change in Disease Activity Score of 28 Joints after treatment (ΔDAS28), confidence interval (CI), identifier (ID).

Table 14. Proteins measured after three months of treatment with etanercept significantly associated with ΔDAS28 at three months, multivariable model.

Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**% missing before imputation**
Q9H4M9	-0.38 (-0.54 – (-0.22))	4.66E-06	72.82
P02741	-0.17 (-0.29 – (-0.05))	0.0047	10.51
Age at baseline	-0.02 (-0.05 – (-0.00))	0.0308	N/A
Male sex	0.20 (-0.34 – 0.74)	0.4635	N/A
Disease duration	0.01 (-0.01 – 0.03)	0.4756	N/A

ABBREVIATIONS: Adjusted (adj), change in Disease Activity Score of 28 Joints after treatment (ΔDAS28), confidence interval (CI).

Table 15. Proteins measured after three months of treatment with etanercept significantly associated with TJC at three months, univariate analysis.

Protein (UniProt ID)	**β-coefficient (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P43490	-0.09 (-0.14 – (-0.05))	3.96E-05	0.0086	53.08
Q12797	-0.05 (-0.07 – (-0.02)	0.0001	0.0130	41.54
P01375	-0.04 (-0.06 – (-0.02))	0.0002	0.0137	40.77
P05019	-0.04 (-0.06 – (-0.02))	0.0003	0.0166	78.97
P14174	-0.02 (-0.03 – (-0.01))	0.0004	0.0166	71.54
Q9Y446	-0.05 (-0.07 – (-0.02))	0.0006	0.0220	43.85
P30043	-0.04 (-0.06 – (-0.02))	0.0007	0.0220	77.95
P12956	-0.05 (-0.08 – (-0.02))	0.0010	0.0282	60.26
Q9H4M9	0.06 (0.02 – 0.09)	0.0014	0.0347	72.82
P0DJI9	0.03 (0.01 – 0.04)	0.0023	0.0463	65.90
Q12906	-0.02 (-0.03 – (-0.01))	0.0024	0.0463	63.08

ABBREVIATIONS: Confidence interval (CI), identifier (ID), tender joint count (TJC).

Table 16. Proteins measured after three months of treatment with etanercept significantly associated with TJC at three months, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P43490	-0.09 (-0.14 – (-0.05))	7.36E-05	0.0159	53.08
P01375	-0.04 (-0.06 – (-0.02))	0.0003	0.0215	40.77
Q12797	-0.04 (-0.06 – (-0.02))	0.0004	0.0215	41.54
P05019	-0.03 (-0.05 – (-0.01))	0.0006	0.0215	78.97
Q9Y446	-0.05 (-0.08 – (-0.02))	0.0006	0.0215	43.85
P14174	-0.02 (-0.03 – (-0.01))	0.0006	0.0215	71.54
P30043	-0.04 (-0.06 – (-0.02))	0.0010	0.0311	77.95
Q9H4M9	0.06 (0.02 – 0.09)	0.0013	0.0359	72.82

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), identifier (ID), rheumatoid arthritis (RA), tender joint count (TJC).

Table 17. Proteins measured after three months of treatment with etanercept significantly associated with TJC at three months, multivariable model.

Variable	β-coefficient$_{adj}$ (95% CI)	p-value	% missing before imputation
P43490	-0.12 (-0.72 - .47)	0.6910	53.08
P01375	-1.49 (-2.58 – (-0.41))	0.0079	40.77
Q12797	-0.99 (-2.04 – 0.07)	0.0700	41.54
P05019	-0.83 (-2.09 – 0.43)	0.1976	78.97
Q9Y446	-0.70 (-1.54 – 0.14)	0.1056	43.85
P14174	-4.21 (-6.45 – (-1.98))	0.0003	71.54
P30043	-0.81 (-1.88 – 0.25)	0.1361	77.95
Q9H4M9	0.82 (0.16 – 1.48)	0.0158	72.82
Age at baseline	-0.01 (-0.10 – 0.07)	0.7631	N/A
Male sex	-1.82 (-3.96 – 0.32)	0.0979	N/A
Disease duration	-0.02 (-0.11 – 0.06)	0.5939	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), tender joint count (TJC).

Table 18. Proteins measured after three months of treatment with etanercept significantly associated with SJC at three months, univariate analysis.

Protein (UniProt ID)	β-coefficient (95% CI)	p-value	Adjusted p-value	% missing before imputation
Q12906	-0.04 (-0.06 – (-0.02))	3.88E-05	0.0078	63.08
P05019	-0.07 (-0.11 – (-0.04))	7.25E-05	0.0078	78.97
Q02985	0.04 (0.02 – 0.06)	0.0001	0.0080	66.67
P02741	0.17 (0.09 – 0.26)	0.0001	0.0080	10.51
P01375	-0.07 (-0.11 – (-0.03))	0.0003	0.0110	40.77
P43490	-0.15 (-0.24 – (-0.07))	0.0006	0.0220	53.08
Q12797	-0.08 (-0.12 – (-0.03))	0.0012	0.0357	41.54
P18428	0.06 (0.02 – 0.09)	0.0018	0.0482	0.51
O95445	-0.05 (-0.08 – (-0.02))	0.0023	0.0482	12.31
P78347	-0.05 (-0.08 – (-0.02))	0.0023	0.0482	70.51
P02748	0.05 (0.02 – 0.09)	0.0025	0.0482	0.00

ABBREVIATIONS: Confidence interval (CI), identifier (ID), swollen joint count (SJC).

Table 19. Proteins measured after three months of treatment with etanercept significantly associated with SJC at three months, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	β-coefficient$_{adj}$ (95% CI)	p-value	Adjusted p-value	% missing before imputation
Q12906	-0.04 (-0.06 – (-0.02))	8.10E-05	0.0100	63.08
P05019	-0.07 (-0.11 – (-0.04))	0.0001	0.0100	78.97
Q02985	0.04 (0.02 – 0.06)	0.0002	0.0100	66.67
P02741	0.17 (0.08 – 0.26)	0.0002	0.0100	10.51
P01375	-0.07 (-0.11 – (-0.03)	0.0004	0.0189	40.77
P43490	-0.15 (-0.24 – (-0.06))	0.0011	0.0405	53.08

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), identifier (ID), rheumatoid arthritis (RA), swollen joint count (SJC).

Table 20. Proteins measured after three months of treatment with etanercept significantly associated with SJC at three months, multivariable model.

Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**% missing before imputation**
Q12906	-1.75 (-2.82 – (-0.69))	0.0016	63.08
P05019	-0.81 (-1.41 – (-0.21))	0.0091	78.97
Q02985	1.39 (0.29 – 2.49)	0.0143	66.67
P02741	0.34 (0.10 – 0.59)	0.0075	10.51
P01375	-0.63 (-1.19 – (-0.07))	0.0290	40.77
P43490	-0.10 (-0.35 – 0.15)	0.4260	53.08
Age at baseline	0.03 (-0.01 – 0.07)	0.1944	N/A
Male sex	-0.95 (-2.04 – 0.14)	0.0909	N/A
Disease duration	0.00 (-0.04 – 0.05)	0.8350	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), swollen joint count (SJC).

Table 21. Proteins measured before treatment with etanercept significantly associated with CRP at baseline, univariate analysis.

Protein (UniProt ID)	β-coefficient (95% CI)	p-value	Adjusted p-value	% missing before imputation
P0DJI8	0.06 (0.05 – 0.07)	1.19E-22	<1E-06	26.92
P17980	0.02 (0.01 – 0.02)	3.52E-16	<1E-06	55.13
P02763	0.03 (0.02 – 0.03)	1.51E-13	<1E-06	0.00
P0DJI9	0.03 (0.02 – 0.04)	1.55E-13	<1E-06	65.90
Q00610	0.02 (0.01 – 0.02)	2,99E-11	<1E-06	57.44
Q92952	0.02 (0.01 – 0.02)	4.20E-11	<1E-06	43.08
P06727	0.03 (0.02 – 0.04)	8.03E-10	<1E-06	13.33
P18428	0.02 (0.01 – 0.02)	1.61E-08	<1E-06	0.51
P02748	0.01 (0.01 – 0.02)	2.68E-08	1.00E-06	0.00
P01011	0.01 (0.01 – 0.02)	2.68E-08	3.00E-06	0.00
P19838	0.02 (0.01 – 0.03)	2.43E-06	4.40E-05	10.77
P49908	-0.01 (-0.02 – (-0.01))	7.03E-06	0.0001	6.92
O95445	-0.01 (-0.01 – (-0.00))	2.73E-05	0.0004	12.31
Q9BXR6	0.01 (0.01 – 0.02)	3.96E-05	0.0006	41.03
P06753	0.01 (0.00 – 0.01)	0.0001	0.0019	71.54
P46939	-0.01 (-0.01 – (-0.00))	0.0002	0.0020	19.49
P04114	-0.01 (-0.01 – (-0.00))	0.0003	0.0040	0.00
P02753	-0.01 (-0.01 – (-0.00))	0.0004	0.0040	0.00
Q9Y6R7	-0.01 (-0.01 – (-0.00))	0.0004	0.0041	77.69
Q92851	-0.02 (-0.03 – (-0.01))	0.0004	0.0041	27.18
P37108	0.01 (0.00 – 0.01)	0.0006	0.0052	56.92
P54578	0.01 (0.00 – 0.01)	0.0006	0.0052	57.69
P06737	-0.01 (-0.01 – (-0.00)	0.0007	0.0067	69.74
P46734	0.01 (0.00- 0.01)	0.0008	0.0069	68.97
P02654	-0.01 (-0.02 – (-0.00))	0.0011	0.0086	0.51
P01776	-0.02 (-0.03 – (-0.01))	0.0011	0.0086	57.44
P98179	-0.01 (-0.02 – (-0.00))	0.0011	0.0086	38.72
Q13740	-0.01 (-0.01 – (-0.00))	0.0012	0.0086	50.77
P20700	-0.01 (-0.01 – (-0.00))	0.0015	0.0107	21.03
P02743	0.01 (0.00 – 0.01)	0.0015	0.0107	0.26
P02452	0.01 (0.00 – 0.01)	0.0018	0.0123	61.79
Q99497	0.02 (0.01 – 0.02)	0.0019	0.0127	72.05
P08603	0.01 (0.00 – 0.01)	0.0025	0.0155	0.00
P27169	-0.01 (-0.01 – (-0.00))	0.0025	0.0155	0.00
P05109	0.02 (0.01 – 0.04)	0.0035	0.0208	34.36
Q9HC38	-0.01 (-0.02 – (-0.00))	0.0036	0.0208	41.28
Q14116	-0.01 (-0.02 – (-0.00))	0.0037	0.0208	58.46
O00329	0.02 (0.01 – 0.03)	0.0046	0.0253	38.72
P17948	0.01 (0.00 – 0.01)	0.0053	0.0284	43.85
Q05682	-0.02 (-0.03 – (-0.00))	0.0058	0.0305	76.67
Q99832	0.00 (0.00 – 0.01)	0.0070	0.0358	51.28
P40939	0.02 (0.00 – 0.03)	0.0085	0.0427	56.92
Q15185	-0.01 (-0.02 – (-0.00))	0.0087	0.0429	28.97
Q15746	0.00 (0.00 – 0.01)	0.0094	0.0451	54.36
Q9Y283	0.02 (0.00 – 0.03)	0.0102	0.0478	54.87
Q9H0W9	-0.01 (-0.02 – (-0.00))	0.0108	0.0490	49.23
P02656	-0.01 (-0.02 – (-0.00))	0.0109	0.0490	0.00
Q15084	0.01 (0.00 – 0.01)	0.0112	0.0490	58.46
P29144	-0.01 (-0.02 – (-0.00))	0.0113	0.0490	38.72

ABBREVIATIONS: Confidence interval (CI), C-reactive protein (CRP), identifier (ID).

Table 22. Proteins measured before treatment with etanercept significantly associated with CRP at baseline, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P0DJI8	0.06 (0.05 – 0.07)	7.37E-22	<1E-06	26.92
P17980	0.02 (0.01 – 0.02)	1.03E-15	<1E-06	55.13
P02763	0.03 (0.02 – 0.03)	3.34E-13	<1E-06	0.00
P0DJI9	0.03 (0.02 – 0.04)	4.41E-13	<1E-06	65.90
Q00610	0.02 (0.01 – 0.02)	5.41E-11	<1E-06	57.44
Q92952	0.02 (0.01 – 0.02)	6.62E-11	<1E-06	43.08
P06727	0.03 (0.02 – 0.04)	9.84E-10	<1E-06	13.33
P02748	0.01 (0.01 – 0.02)	2.04E-08	<1E-06	0.00
P18428	0.02 (0.01 – 0.02)	2.10E-08	<1E-06	0.51
P01011	0.01 (0.01 – 0.02)	1.46E-07	3.00E-06	0.00
P19838	0.03 (0.02 – 0.04)	2.62E-06	4.70E-05	10.77
P49908	-0.01 (-0.02 – (-0.01))	7.37E-06	0.0001	6.92
O95445	-0.01 (-0.01 – (-0.00))	3.80E-05	0.0006	12.31
Q9BXR6	0.01 (0.01 – 0.02)	4.99E-05	0.0007	41.03
P46939	-0.01 (-0.01 – (-0.00))	0.0001	0.0015	19.49
P06753	0.01 (0.00 – 0.01)	0.0002	0.0026	71.54
Q9Y6R7	-0.01 (-0.01 – (-0.00))	0.0002	0.0029	77.69
P04114	-0.01 (-0.01 – (-0.00)	0.0003	0.0038	0.00
P02753	-0.01 (-0.01 – (-0.00))	0.0004	0.0040	0.00
Q92851	-0.02 (-0.03 – (-0.01))	0.0005	0.0049	27.18
P37108	0.01 (0.01 – 0.02)	0.0006	0.0059	56.92
P54578	0.01 (0.00 – 0.01)	0.0008	0.0074	57.69
P06737	-0.01 (-0.01 – (-0.00))	0.0008	0.0074	69.74
P01776	-0.02 (-0.03 – (-0.01))	0.0009	0.0082	57.44
P20700	-0.01 (-0.01 – (-0.00))	0.0010	0.0082	21.03
Q13740	-0.01 (-0.01 – (-0.00))	0.0011	0.0082	50.77
P46734	0.01 (0.00 – 0.01)	0.0011	0.0082	48.72
Q99497	0.02 (0.01 – 0.03)	0.0033	0.0082	72.05
P98179	-0.01 (-0.02 – (-0.00))	0.0016	0.0083	38.72
P02654	-0.01 (-0.02 – (-0.00))	0.0013	0.0093	0.51
P02743	0.01 (0.00 – 0.01)	0.0014	0.0093	0.26
P02452	0.01 (0.00 – 0.01)	0.0021	0.0135	61.79
P08603	0.01 (0.00 – 0.01)	0.0023	0.0151	0.00
O00329	0.02 (0.01 – 0.03)	0.0028	0.0170	38.72
P05109	0.02 (0.01 – 0.04)	0.0032	0.0192	34.36
Q05682	-0.02 (-0.03 – (-0.01))	0.0034	0.0198	76.67
P17948	0.01 (0.00 – 0.01)	0.0037	0.0210	43.85
P27169	-0.01 (-0.01 – (-0.00))	0.0040	0.0221	0.00
Q14116	-0.01 (-0.02 – (-0.00))	0.0041	0.0021	58.46
Q9HC38	-0.01 (-0.02 – (-0.00))	0.0044	0.0230	41.28
Q15746	0.00 (0.00 – 0.01)	0.0045	0.0230	54.36
P02656	-0.01 (-0.02 – (-0.00))	0.0062	0.0305	0.00
P29144	-0.01 (-0.02 – (-0.00))	0.0062	0.0305	38.72
Q99832	0.00 (0.00 – 0.01)	0.0080	0.0387	51.28
Q02985	0.00 (0.00 – 0.01)	0.0083	0.0388	66.67
Q15185	-0.01 (-0.02 – (-0.00))	0.0087	0.0402	28.97
P08246	0.01 (0.00 – 0.02)	0.0105	0.0464	62.31
P42765	0.00 (0.00 – 0.01)	0.0106	0.0464	45.38
Q9H0W9	-0.01 (-0.02 – (-0.00))	0.0107	0.0464	49.23
P02655	-0.01 (-0.02 – (-0.00))	0.0112	0.0474	0.00

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), C-reactive protein (CRP), identifier (ID), rheumatoid arthritis (RA).

Table 23. Proteins measured before treatment with etanercept significantly associated with CRP at baseline, multivariable model.

Variable	β-coefficient$_{adj}$ (95% CI)	p-value	% missing before imputation
P0DJI8	2.11 (0.40 – 3.82)	0.0172	26.92
P17980	1.17 (5.82 – 17.66)	0.0002	55.13
P02763	3.30 (-0.63 – 7.22)	0.1027	0.00
P0DJI9	3.48 (0.99 – 5.98)	0.0073	65.90
Q00610	-0.02 (-7.48 – 7.44)	0.9958	57.44
Q92952	-1.65 (-6.59 – 3.29)	0.5148	43.08
P06727	2.87 (0.42 – 5.32)	0.0241	13.33
P02748	2.88 (-5.72 – 11.49)	0.5130	0.00
P18428	0.29 (-4.01 – 4.60)	0.8943	0.51
P01011	-4.39 (-11.83 – 3.05)	0.2502	0.00
P19838	1.44 (-0.34 – 3.21)	0.1160	10.77
P49908	-2.64 (-6.04 – 0.76)	0.1314	6.92
O95445	-1.05 (-7.02 – 4.92)	0.7312	12.31
Q9BXR6	2.13 (-1.81 – 6.06)	0.2919	41.03
P46939	-5.52 (-11.74 – 0.71)	0.0855	19.49
P06753	-5.40 (-11.03 – 0.23)	0.0631	71.54
Q9Y6R7	-2.27 (-6.94 – 2.40)	0.3436	77.69
P04114	0.77 (-4.35 – 5.89)	0.7692	0.00
P02753	-9.47 (-14.33 – (-4.61))	0.0002	0.00
Q92851	-1.99 (-3.83 – (-0.15))	0.0365	27.18
P37108	2.17 (-0.43 – 4.77)	0.1049	56.92
P54578	4.95 (-2.68 – 12.58)	0.2068	57.69
P06737	2.91 (-3.75 – 9.57)	0.3931	69.74
P01776	1.65 (-0.86 – 4.17)	0.2012	57.44
P20700	-4.00 (-9.41 – 1.42)	0.1511	21.03
Q13740	1.85 (-3.18 – 6.87)	0.4733	50.77
P46734	1.58 (-4.04 – 7.19)	0.5836	68.97
Q99497	1.09 (-1.11 – 3.28)	0.3358	72.05
P98179	2.32 (-3.22 – 7.85)	0.4145	38.72
P02654	1.37 (-2.75 – 5.48)	0.5174	0.51
P02743	-2.06 (-9.42 – 5.29)	0.5837	0.26
P02452	3.15 (-1.01 – 7.30)	0.1410	61.79
P08603	-6.30 (-13.10 – 0.50)	0.0725	0.00
O00329	0.30 (-1.16 – 1.76)	0.6871	38.72
P05109	-0.76 (-2.04 – 0.52)	0.2498	34.36
Q05682	0.24 (-1.94 – 2.42)	0.8275	76.67
P17948	-1.13 (-6.65 – 4.39)	0.6888	43.85
P27169	3.85 (-2.27 – 9.98)	0.2207	0.00
Q14116	-0.21 (-4.00 – 3.58)	0.9143	58.46
Q9HC38	2.98 (-0.58 – 6.54)	0.1046	41.28
Q15746	-4.12 (-14.35 – 6.11)	0.4319	54.36
P02656	0.06 (-4.02 – 4.14)	0.9762	0.00
P29144	-6.30 (-10.63 – (-1.98))	0.0052	38.72
Q99832	3.41 (-3.09 – 9.91)	0.3066	51.28
Q02985	-7.04 (-12.96 – (-1.12))	0.0217	66.67
Q15185	-2.76 (-6.63 – 1.11)	0.1661	28.97
P08246	2.70 (0.33 – 5.08)	0.0279	62.31
P42765	2.13 (-6.27 – 10.53)	0.6202	45.38
Q9H0W9	-1.85 (-4.93 – 1.23)	0.2424	49.23
P02655	1.45 (-2.20 – 5.11)	0.4380	0.00
Age at baseline	0.03 (-0.23 – 0.28)	0.8370	N/A
Male sex	4.55 (-2.04 – 11.14)	0.1793	N/A
Disease duration	0.00 (-0.24 – 0.24)	0.9943	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), C-reactive protein (CRP).

Table 24. Proteins measured before treatment with etanercept significantly associated with CRP at baseline, univariate analysis.

Protein (UniProt ID)	**β-coefficient (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P49908	-0.01 (-0.02 – (-0.01))	6.41E-05	0.0139	6.92
Q00610	0.01 (0.01 – 0.02)	0.0004	0.0139	57.44
P46734	0.01 (0.00 – 0.02)	0.0005	0.0295	68.97

ABBREVIATIONS: Confidence interval (CI), C-reactive protein (CRP), identifier (ID).

Table 25. Proteins measured before treatment with etanercept significantly associated with CRP at baseline, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P49908	-0.01 (-0.03 – (-0.01))	0.0001	0.0226	6.92
Q00610	0.01 (0.01 – 0.02)	0.0005	0.0263	57.44
P46734	0.01 (0.00 – 0.02)	0.0004	0.0263	68.97

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), C-reactive protein (CRP), identifier (ID).

Table 26. Proteins measured before treatment with etanercept significantly associated with CRP at six months, multivariable model.

Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**% missing before imputation**
Q00610	4.55 (0.63 – 8.47)	0.0244	57.44
P46734	7.26 (3.09 – 11.43)	0.0008	68.97
P49908	-5.50 (-8.95 – (-2.05))	0.0022	6.92
Age at baseline	0.21 (-0.04 – 0.46)	0.0989	N/A
Male sex	-0.74 (-7.11 – 5.64)	0.8216	N/A
Disease duration	-0.29 (-0.56 – (-0.02))	0.0391	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), C-reactive protein (CRP).

Table 27. Proteins measured after three months of treatment with etanercept significantly associated with CRP at three months, univariate analysis.

Protein (UniProt ID)	β-coefficient (95% CI)	p-value	Adjusted p-value	% missing before imputation
P18428	0.03 (0.02 – 0.05)	1.73E-09	<1E-06	0.51
P0DJI8	0.06 (0.03 – 0.08)	6.21E-07	4.50E-05	26.92
P02763	0.04 (0.02 – 0.05)	1.93E-05	0.0010	0.00
P02743	0.03 (0.01 – 0.04)	2.32E-05	0.0010	0.26
P02748	0.02 (0.01 – 0.03)	3.19E-05	0.0010	0.00
Q00610	0.02 (0.01 – 0.03)	3.30E-05	0.0010	57.44
P06727	0.04 (0.02 – 0.06)	9.84E-05	0.0027	13.33
Q14CX7	-0.04 (-0.06 – (-0.02))	0.0001	0.0028	48.97
Q92952	0.02 (0.01 – 0.03)	0.0003	0.0062	43.08
P01011	0.02 (0.01 – 0.03)	0.0005	0.0093	0.00
P05109	0.06 (0.03 – 0.10)	0.0006	0.0105	34.36
P06753	0.02 (0.01 – 0.03)	0.0006	0.0108	71.54
P17980	0.02 (0.01 – 0.03)	0.0007	0.0112	55.13
P08603	0.01 (0.01 – 0.02)	0.0013	0.0179	0.00
Q02985	0.01 (0.00 – 0.02)	0.0013	0.0179	66.67
P0DJI9	0.02 (0.01 – 0.03)	0.0021	0.0265	65.90
P08294	0.03 (0.01 – 0.06)	0.0024	0.0284	50.77
Q9H4M9	0.03 (0.01 – 0.06)	0.0024	0.0303	72.82
P62701	0.01 (0.00 – 0.02)	0.0033	0.0358	12.56
Q9HC38	-0.03 (-0.05 – (-0.01))	0.0047	0.0479	41.28

ABBREVIATIONS: Confidence interval (CI), identifier (ID).

Table 28. Proteins measured after three months of treatment with etanercept significantly associated with CRP at three months, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P18428	0.04 (0.02 – 0.05)	1.53E-09	<1E-06	0.51
P0DJI8	0.06 (0.04 – 0.08)	3.70E-07	2.70E-05	26.92
P02763	0.04 (0.02 – 0.06)	7.59E-06	0.0004	0.00
Q00610	0.02 (0.01 – 0.03)	5.54E-05	0.0018	57.44
P02748	0.02 (0.01 – 0.03)	5.66E-05	0.0018	0.00
P06727	0.04 (0.02 – 0.06)	6.05E-05	0.0018	54.87
P02743	0.03 (0.01 – 0.04)	6.74E-05	0.0018	0.26
Q14CX7	-0.04 (-0.07 – (-0.02)	0.0001	0.0027	48.97
P17980	0.02 (0.01 – 0.03)	0.0002	0.0045	55.13
Q92952	0.02 (0.01 – 0.03)	0.0005	0.0106	43.08
P01011	0.02 (0.01 – 0.03)	0.0006	0.0116	0.00
P05109	0.06 (0.03 – 0.10)	0.0009	0.0148	34.36
P08603	0.02 (0.01 – 0.02)	0.0010	0.0148	0.00
P06753	0.02 (0.1 – 0.03)	0.0010	0.0148	71.54
P0DJI9	0.02 (0.01 – 0.03)	0.0016	0.0220	65.90
P0C0L5	0.03 (0.01 – 0.04)	0.0021	0.0248	6.92
P08294	0.04 (0.01 – 0.06)	0.0024	0.0290	50.77
Q02985	0.01 (0.00 – 0.02)	0.0029	0.0328	66.67
P62701	0.01 (0.00 – 0.02)	0.0031	0.0328	12.56
Q9H4M9	0.03 (0.01 – 0.05)	0.0044	0.0456	72.82
P05546	0.02 (0.00 – 0.03)	0.0048	0.0475	0.00

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), identifier (ID), rheumatoid arthritis (RA).

Table 29. Proteins measured after three months of treatment with etanercept significantly associated with CRP at three months, multivariable model.

Variable	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**% missing before imputation**
P18428	4.12 (1.60 – 6.64)	0.0017	0.51
P0DJI8	1.10 (-0.12 – 2.32)	0.0783	26.92
P02763	1.16 (-0.68 – 2.99)	0.2185	0.00
Q00610	1.66 (-1.95 – 5.27)	0.3692	35.9
P02748	-1.21 (-4.84 – 2.41)	0.5132	0.00
P06727	0.88 (-0.42 – 2.17)	0.1882	13.33
P02743	0.61 (-1.96 – 3.18)	0.6413	0.26
Q14CX7	-1.39 (-2.63 – (-0.15))	0.0303	48.97
P17980	0.32 (-2.92 – 3.56)	0.8458	55.13
Q92952	0.06 (-2.48 – 2.59)	0.9655	43.08
P01011	-0.61 (-3.87 – 2.66)	0.7170	0.00
P05109	-0.07 (-0.77 – 0.63)	0.8433	34.36
P08603	-1.27 (-5.15 – 2.52)	0.5239	0.00
P06753	-0.38 (-3.39 – 2.64)	0.8077	71.54
P0DJI9	0.86 (-1.58 – 3.30)	0.4923	65.90
P0C0L5	0.38 (-1.16 – 1.93)	0.6300	6.92
P08294	0.69 (-0.41 – 1.78)	0.2226	50.77
Q02985	3.58 (-0.46 – 7.62)	0.0848	66.67
P62701	-0.62 (-5.48 – 4.24)	0.8026	12.56
Q9H4M9	0.25 (-0.80 – 1.31)	0.6359	72.82
P05546	-1.36 (-4.78 – 2.06)	0.4370	0.00
Age at baseline	-0.03 (-0.17 – 0.11)	0.6739	N/A
Male sex	3.85 (0.23 – 7.47)	0.0388	N/A
Disease duration	0.09 (-0.06 – 0.23)	0.2326	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI).

Table 30. Proteins measured after three months of treatment with etanercept significantly associated with CRP at six months, univariate analysis.

Protein (UniProt ID)	**β-coefficient (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P0DJI8	0.03 (0.02 – 0.04)	1.21E-05	0.0013	26.92
Q12797	-0.02 (-0.03 – (-0.01))	3.41E-05	0.0019	41.54
P02763	0.02 (0.01 – 0.03)	3.43E-05	0.0019	0.00
P01011	0.01 (0.01 – 0.02)	5.38E-05	0.0023	0.00
P46734	0.01 (0.01 – 0.02)	0.0001	0.0039	68.97
Q9BXR6	0.01 (0.01 – 0.02)	0.0002	0.0068	41.03
Q15746	0.01 (0.00 – 0.01)	0.0003	0.0077	54.36
P02748	0.01 (0.01 – 0.02)	0.0003	0.0082	0
Q7KZF4	-0.02 (-0.03 – (0.01))	0.0023	0.0487	24.36
Q9HDC9	0.01 (0.00 – 0.01)	0.0029	0.0487	57.95
Q00610	0.01 (0.00 – 0.01)	0.0030	0.0487	57.44
P17980	0.01 (0.00 – 0.01)	0.0031	0.0487	55.13
P19838	0.02 (0.01 – 0.03)	0.0034	0.0487	10.77
Q92952	0.01 (0.00 – 0.02)	0.0036	0.0487	43.08
Q6UX71	0.00 (0.00 – 0.01)	0.0036	0.0487	68.21

ABBREVIATIONS: Confidence interval (CI), identifier (ID).

Table 31. Proteins measured after three months of treatment with etanercept significantly associated with CRP at six months, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P0DJI8	0.03 (0.01 – 0.04)	5.25E-05	0.0056	26.92
P02763	0.02 (0.01 – 0.03)	9.60E-05	0.0056	0.00
P01011	0.01 (0.01 – 0.02)	0.0001	0.0056	0.00
P46734	0.01 (0.01 – 0.02)	0.0001	0.0056	68.97
Q12797	-0.02 (-0.02 – (-0.01))	0.0002	0.0056	41.54
Q15746	0.01 (0.00 – 0.01)	0.0002	0.0074	54.36
Q9BXR6	0.01 (0.01 – 0.02)	0.0006	0.0156	43.33
P02748	0.01 (0.00 – 0.02)	0.0009	0.0221	0.00
Q9HDC9	0.01 (0.00 – 0.01)	0.0022	0.0474	57.95

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), identifier (ID), rheumatoid arthritis (RA).

Table 32. Proteins measured after three months of treatment with etanercept significantly associated with CRP at six months, multivariable model.

Variable	β-coefficient$_{adj}$ (95% CI)	p-value	% missing before imputation
P0DJI8	2.90 (0.94 – 4.87)	0.0044	26.92
P02763	2.40 (-0.45 – 5.24)	0.1006	0.00
P01011	0.02 (-5.80 – 5.84)	0.9941	0.00
P46734	6.29 (1.88 – 10.70)	0.0060	68.97
Q12797	-2.43 (-5.41 – 0.54)	0.1106	41.54
Q15746	7.43 (-0.18 – 15.04)	0.0578	54.36
Q9BXR6	2.25 (-1.45 – 5.96)	0.2349	43.33
P02748	-3.28 (-9.04 – 2.48)	0.2660	0.00
Q9HDC9	3.13 (-1.14 – 7.39)	0.1527	57.95
Age at baseline	0.12 (-0.12 – 0.36)	0.3378	N/A
Male sex	-1.25 (-7.36 – 4.85)	0.6881	N/A
Disease duration	-0.05 (-0.31 – 0.20)	0.6826	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI).

Table 33. Proteins measured after six months of treatment with etanercept significantly associated with CRP at six months, univariate analysis.

Protein (UniProt ID)	β-coefficient (95% CI)	p-value	Adjusted p-value	% missing before imputation
P0DJI8	0.07 (0.05 – 0.10)	1.38E-06	0.0001	26.92
P02748	0.03 (0.02 – 0.04)	6.23E-06	0.0004	0.00
P0DJI9	0.04 (0.02 – 0.05)	1.43E-05	0.0008	65.90
P02763	0.04 (0.02 – 0.06)	2.47E-05	0.0010	0.00
Q14012	-0.08 (-0.12 – (-0.05))	2.87E-05	0.0010	62.82
P01011	0.02 (0.01 – 0.04)	4.94E-05	0.0014	0.00
P18428	0.03 (0.02 – 0.04)	5.10E-05	0.0014	0.51
P17980	0.02 (0.01 – 0.03)	5.70E-05	0.0014	55.13
P06727	0.04 (0.02 – 0.06)	0.0004	0.0028	13.33
P05109	0.08 (0.04 – 0.13)	0.0004	0.0082	34.36
Q9BXR6	0.02 (0.01 – 0.03)	0.0006	0.0109	41.03
P41250	-0.01 (-0.02 – (-0.00))	0.0011	0.0185	65.13
P42765	0.01 (0.00 – 0.02)	0.0014	0.0219	45.38
P31749	0.01 (0.00 – 0.02)	0.0016	0.0234	67.44
P62701	0.01 (0.00 – 0.02)	0.0023	0.0300	12.56
P78347	-0.02 (-0.03 – (-0.01))	0.0024	0.0300	70.51
Q04917	-0.03 (-0.04 – (-0.01))	0.0027	0.0328	72.05
P62937	-0.06 (-0.09 – (-0.02))	0.0032	0.0368	60.77
P14174	0.01 (0.00 – 0.02)	0.0039	0.0417	71.54
Q13263	-0.02 (-0.04 – (-0.01))	0.0044	0.0452	62.31
P11766	0.01 (0.00 – 0.02)	0.0049	0.0458	56.67
P30043	-0.04 (-0.06 – (-0.01))	0.0049	0.0458	77.95
Q05639	-0.02 (-0.03 – (-0.01))	0.0053	0.0458	73.85
P13797	-0.04 (-0.07 – (-0.01))	0.0053	0.0458	38.21

ABBREVIATIONS: Confidence interval (CI), identifier (ID).

Table 34. Proteins measured after six months of treatment with etanercept significantly associated with CRP at six months, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**β-coefficient$_{adj}$ (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P0DJI8	0.08 (0.05 – 0.10)	3.08E-06	0.0003	26.92
P02748	0.03 (0.02 – 0.04)	1.69E-05	0.0012	0.00
P0DJI9	0.04 (0.02 – 0.05)	3.86E-05	0.0019	65.90
P02763	0.04 (0.02 – 0.06)	4.41E-05	0.0019	0.00
Q14012	-0.08 (-0.12 – (-0.04))	8.05E-05	0.0026	62.82
P18428	0.03 (0.01 – 0.04)	8.57E-05	0.0026	0.51
P01011	0.02 (0.01 – 0.04)	0.0001	0.0039	0.00
P17980	0.02 (0.01 – 0.03)	0.0002	0.0047	55.13
P06727	0.03 (0.01 – 0.05)	0.0009	0.0180	13.33
P05109	0.08 (0.04 – 0.13)	0.0009	0.0180	34.36
Q9BXR6	0.01 (0.01 – 0.02)	0.0014	0.0258	41.03
P42765	0.01 (0.00 – 0.02)	0.0018	0.0297	45.38
P62937	-0.06 (-0.l0 – (-0.02))	0.0022	0.0342	60.77
P02743	0.02 (0.01 – 0.04)	0.0026	0.0370	0.26
P78347	-0.02 (-0.03 – (-0.01))	0.0027	0.0370	70.51
P41250	-0.01 (-0.02 – (-0.00))	0.0032	0.0394	65.13
P62701	0.01 (0.00 – 0.02)	0.0033	0.0394	12.56
Q04917	-0.03 (-0.04 – (-0.01))	0.0039	0.0422	72.05
P31749	0.01 (0.00 – 0.02)	0.0039	0.0422	67.44
P14174	0.01 (0.00 – 0.02)	0.0043	0.0436	71.54
P30043	-0.04 (-0.06 – (-0.01))	0.0044	0.0436	77.95
P11766	0.01 (0.00 – 0.02)	0.0052	0.0484	56.67

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), identifier (ID), rheumatoid arthritis (RA).

Table 35. Proteins measured after six months of treatment with etanercept significantly associated with CRP at six months, multivariable model.

Variable	β-coefficient$_{adj}$ (95% CI)	p-value	% missing before imputation
P0DJI8	-0.22 (-2.28 – 1.83)	0.8329	26.92
P02748	0.74 (-5.51 – 7.00)	0.8171	0.00
P0DJI9	2.41 (-1.20 – 6.02)	0.1986	65.90
P02763	3.81 (0.90 – 6.71)	0.0143	0.00
Q14012	-0.63 (-2.09 – 0.83)	0.4052	62.82
P18428	1.86 (-2.17 – 5.89)	0.3718	0.51
P01011	0.06 (-7.90 – 8.01)	0.9891	0.00
P17980	1.61 (-4.21 – 7.43)	0.5904	55.13
P06727	3.37 (0.23 – 6.51)	0.0422	13.33
P05109	-0.05 (-1.13 – 1.03)	0.9313	34.36
Q9BXR6	6.28 (1.15 – 11.42)	0.0215	41.03
P42765	0.50 (-8.83 – 9.84)	0.9164	45.38
P62937	-1.03 (-3.00 – 0.94)	0.3124	60.77
P02743	-2.13 (-7.81 – 3.55)	0.4675	0.26
P78347	-0.90 (-7.77 – 5.97)	0.7977	70.51
P41250	-18.02 (-26.21 – (-9.83))	0.0001	65.13
P62701	2.67 (-11.16 – 16.51)	0.7070	12.56
Q04917	2.32 (-2.85 – 7.49)	0.3850	72.05
P31749	3.62 (-4.72 – 11.95)	0.4006	67.44
P14174	-2.79 (-10.46 – 4.88)	0.4804	71.54
P30043	2.41 (-1.08 – 5.90)	0.1832	77.95
P11766	8.13 (0.14 – 16.11)	0.0534	56.67
Age at baseline	-0.11 (-0.36 – 0.13)	0.3733	N/A
Male sex	4.52 (-2.64 – 11.68)	0.2240	N/A
Disease duration	-0.00 (-0.26 – 0.26)	0.9886	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI).

Table 36. Proteins measured after three months of treatment with etanercept associated with poor EULAR response at three months, univariate analysis.

Protein (UniProt ID)	OR (95% CI)	p-value	Adjusted p-value	% missing before imputation
P43490	0.57 (0.45 – 0.72)	3.33E-06	0.0007	53.08
Q12797	0.42 (0.27 – 0.67)	0.0002	0.0246	41.54
Q12906	0.12 (0.04 – 0.40)	0.0004	0.0253	63.08
P05019	0.36 (0.20 – 0.64)	0.0006	0.0253	78.97
Q9Y446	0.54 (-0.38 – 0.77)	0.0006	0.0253	43.85
P62158	0.76 (0.64 – 0.90)	0.0011	0.0403	5.64
P12110	0.44 (0.27 – 0.73)	0.0014	0.0403	60.77
P02741	1.38 (1.13 – 1.68)	0.0015	0.0403	10.51
P0DJI8	1.51 (1.17 – 1.94)	0.0017	0.0408	26.92

ABBREVIATIONS: Confidence interval (CI), European League Against Rheumatism (EULAR), identifier (ID), odds ratio (OR).

Table 37. Proteins measured after three months of treatment with etanercept associated with poor EULAR response at three months, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**OR_{adj} (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P43490	0.58 (0.46 – 0.74)	7.59E-06	0.0016	53.08
Q12797	0.43 (0.27 – 0.69)	0.0004	0.0282	41.54
Q9Y446	0.52 (0.36 – 0.75)	0.0005	0.0282	43.85
Q12906	0.13 (0.04 – 0.41)	0.0005	0.0282	63.08
P05019	0.36 (0.20 – 0.65)	0.0007	0.0296	78.97

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), European League Against Rheumatism (EULAR), identifier (ID), rheumatoid arthritis (RA).

Table 38. Proteins measured after three months of treatment with etanercept associated with poor EULAR response at three months, multivariable model.

Variable	**OR_{adj} (95% CI)**	**p-value**	**% missing before imputation**
P43490	0.81 (0.59 – 1.13)	0.2201	53.08
Q12906	0.21 (0.06 – 0.64)	0.0080	63.08
Q12797	0.69 (0.37 – 1.19)	0.2020	41.54
P05019	0.44 (0.21 – 0.82)	0.0166	78.97
Q9Y446	0.57 (0.36 – 0.87)	0.0127	43.85
Age at baseline	1.00 (0.96 – 1.05)	0.8742	N/A
Male sex	0.45 (0.11 – 1.51)	0.2214	N/A
Disease duration	0.98 (0.93 – 1.03)	0.4241	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), European League Against Rheumatism (EULAR), odds ratio (OR).

Table 39. Proteins measured after three months of treatment with etanercept associated with failure to achieve a MCID in DAS28 at three months, univariate analysis.

Protein (UniProt ID)	**OR (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
Q12797	0.44 (0.28 – 0.67)	0.0001	0.0209	41.54
P43490	0.67 (0.25 – 0.83)	0.0002	0.0209	53.08

ABBREVIATIONS: Confidence interval (CI), Disease Activity Score of 28 Joints (DAS28), identifier (ID), minimum clinically important difference (MCID).

Table 40. Proteins measured after three months of treatment with etanercept associated with failure to achieve a MCID in DAS28 at three months, adjusted for age, biological sex and RA disease duration.

Protein (UniProt ID)	**OR_{adj} (95% CI)**	**p-value**	**Adjusted p-value**	**% missing before imputation**
Q12797	0.45 (0.29 – 0.70)	0.0003	0.0410	41.54
P43490	0.68 (0.55 – 0.84)	0.0004	0.0410	53.08

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), Disease Activity Score of 28 Joints (DAS28), identifier (ID), minimum clinically important difference (MCID).

Table 41. Proteins measured after three months of treatment with etanercept associated with failure to achieve MCID in DAS28 at three months, multivariable model.

Variable	**OR_{adj} (95% CI)**	**p-value**	**% missing before imputation**
P43490	0.79 (0.61 – 1.01)	0.0644	53.08
Q12797	0.61 (0.36 – 0.96)	0.0451	41.54
Age at baseline	1.02 (0.98 – 1.05)	0.3475	N/A
Male sex	0.98 (0.37 – 2.44)	0.9179	N/A
Disease duration	0.95 (0.94 – 1.02)	0.3765	N/A

ABBREVIATIONS: Adjusted (adj), confidence interval (CI), Disease Activity Score of 28 Joints (DAS28), minimally clinically important difference (MCID).

APPENDIX EIGHT: STATISTICALLY SIGNIFICANT DIFFERENTIALLY EXPRESSED PROTEINS IN THE BRAGGSS ETANERCEPT SUB-COHORT

Table 1. Differentially expressed proteins between baseline and 3 months of treatment with etanercept in EULAR good/moderate responders.

Protein	Log-fold change	Average expression	p-value	Adjusted p-value	% missing before imputation
P0DJI9	-1.36	11.97	3.74E-17	8.08E-15	65.90
P20774	-0.45	8.57	9.12E-17	9.85E-15	77.69
P12110	0.89	14.23	2.90E-13	2.09E-11	60.77
Q02985	-0.45	9.31	1.78E-10	8.89E-09	66.67
P06753	-0.67	14.72	2.06E-10	8.89E-09	71.54
Q92952	-0.69	9.93	1.67E-09	6.00E-08	43.08
P0DJI8	-1.63	10.63	2.60E-09	8.08E-08	26.92
P17980	-0.60	16.29	3.02E-09	8.15E-08	55.13
P02741	-1.70	13.27	2.16E-08	5.19E-07	10.51
P18428	-0.57	13.34	7.04E-07	1.52E-05	0.51
P02763	-0.80	19.38	1.12E-06	2.14E-05	0.00
P13796	-0.68	14.54	1.19E-06	2.14E-05	48.72
P05019	0.55	9.36	1.71E-06	2.84E-05	78.97
Q86U17	0.40	8.37	2.30E-06	3.55E-05	68.46
Q96AE4	0.90	12.90	9.08E-06	0.0001	64.10
Q9Y6R7	-0.38	12.17	3.35E-05	0.0005	77.69
Q9P2E9	-0.42	14.09	3.54E-05	0.0005	56.67
Q00610	-0.37	14.06	9.26E-05	0.0011	57.44
P06727	-0.81	19.48	9.43E-05	0.0011	13.33
P02748	-0.40	16.94	0.0001	0.0015	0.00
O95168	-0.61	11.70	0.0001	0.0015	43.85
P43487	-0.51	8.55	0.0002	0.0015	67.18
Q05682	0.79	11.21	0.0002	0.0021	76.67
Q14766	-0.29	12.98	0.0004	0.0035	60
P01011	-0.36	20.83	0.0004	0.0035	0.00
P42765	-0.27	11.84	0.0006	0.0047	45.38
P02753	0.34	17.48	0.0013	0.0104	0.00
P04264	-0.31	13.41	0.0018	0.0138	56.41
Q05639	0.20	15.13	0.0019	0.0138	73.85
P27169	0.25	16.30	0.0021	0.0150	0.00
P22897	0.23	12.42	0.0023	0.0158	43.33
P05451	-0.23	11.09	0.0023	0.0158	43.33
Q12797	0.48	12.04	0.0031	0.0196	41.54
P18510	-0.45	9.93	0.0032	0.0196	15.90
P02743	-0.33	16.65	0.0032	0.0196	0.26
P04114	0.32	18.41	0.0045	0.0264	0.00
Q99497	-0.53	13.41	0.0045	0.0264	72.05
P12109	-0.43	12.43	0.0058	0.0330	58.46
P45983	0.42	10.71	0.0060	0.0330	54.36
Q5VWQ8	0.86	7.70	0.0062	0.0336	33.33

Protein	Log-fold change	Average expression	p-value	Adjusted p-value	% missing before imputation
P46939	0.20	14.59	0.0069	0.0365	19.49
P30048	-0.72	14.96	0.0083	0.0420	13.59
Q6UX71	-0.12	9.07	0.0086	0.0420	68.21
P04207	0.28	10.26	0.0087	0.0420	77.69
P09172	0.34	9.95	0.0089	0.0420	35.64
P12955	-0.39	14.13	0.0091	0.0420	56.67
Q15084	-0.30	12.57	0.0092	0.0420	58.46
P41222	0.21	10.18	0.0093	0.0420	72.31
P06702	-0.44	11.56	0.0109	0.0480	43.08

ABBREVIATIONS: European League Against Rheumatism (EULAR).

Table 2. Differentially expressed proteins between baseline and three months of treatment with etanercept, all patients.

Protein	Log-fold change	Average expression	p-value	Adjusted p-value	% missing before imputation
P05019	1.03	9.36	1.63E-05	0.0035	78.97
O95168	-1.16	11.70	0.0005	0.0400	43.86
P0DJI9	-1.07	11.97	0.0007	0.0400	65.90
Q86U17	0.58	8.37	0.0007	0.0400	68.46

Table 3. Differentially expressed proteins between baseline and six months of treatment with etanercept in EULAR poor responders.

Protein	Log-fold change	Average expression	p-value	Adjusted p-value	% missing before imputation
P02730	1.46	10.85	9.76E-07	0.0002	82.05
Q99497	-2.93	13.19	1.58E-06	0.0002	72.05
Q32MZ4	0.83	14.46	1.75E-05	0.0013	72.56
P46734	0.86	10.83	0.0001	0.0061	68.97
Q02985	-0.74	9.36	0.0003	0.0124	66.67
Q05682	2.27	10.96	0.0006	0.0208	76.67
P04264	0.91	13.39	0.0008	0.0234	56.41
P30043	1.89	9.23	0.0009	0.0235	77.95

ABBREVIATIONS: European League Against Rheumatism (EULAR).

Table 4. Differentially expressed proteins between baseline and six months of treatment with etanercept in EULAR good/moderate responders.

Protein	Log-fold change	Average expression	p-value	Adjusted p-value	% missing before imputation
P46734	1.13	10.83	4.05E-16	8.75E-14	68.97
Q04917	1.68	8.50	7.89E-15	8.52E-13	72.05
P31749	-0.90	9.45	6.30E-12	4.53E-10	67.44
P14174	-0.59	10.33	6.80E-10	3.67E-08	71.54
P04075	-1.18	9.50	3.74E-09	1.62E-07	84.62
P17980	-0.81	16.37	2.12E-08	7.65E-07	55.13
Q15746	-0.77	12.89	4.33E-08	1.33E-06	54.36
Q32MZ4	0.60	14.46	5.88E-08	1.59E-06	72.56
Q9H4M9	1.70	12.30	3.00E-07	7.20E-06	72.82
P06727	-1.31	19.69	4.88E-07	1.05E-05	13.33
Q05682	1.89	10.96	5.85E-07	1.07E-05	76.67
P0DJI9	-1.38	12.20	5.95E-07	1.07E-05	65.90
P06753	-0.72	14.86	2.65E-06	4.41E-05	71.54
P80188	0.61	10.56	4.66E-06	7.19E-05	76.15
P02452	-0.73	11.02	5.27E-06	7.26E-05	61.79
P02763	-1.12	19.48	5.37E-06	7.26E-05	0.00
P18428	-0.76	13.41	8.27E-06	0.0001	0.51
Q99497	-1.46	13.19	1.34E-05	0.0002	72.05
Q6UX71	-0.30	9.08	1.35E-05	0.0002	68.21
P02748	-0.66	17.02	1.65E-05	0.0002	0.00
P02741	-1.87	13.45	2.83E-05	0.0003	10.51
P02730	0.68	10.85	2.93E-05	0.0003	82.05
P40926	0.44	13.08	6.58E-05	0.0006	67.69
Q02985	-0.39	9.36	0.0005	0.0049	66.67
Q13201	0.60	12.78	0.0006	0.0050	63.85
P01011	-0.53	20.88	0.0006	0.0050	0.00
Q9HDC9	0.41	11.56	0.0007	0.0054	57.95
P30043	1.06	9.23	0.0008	0.0061	77.95
Q9BPX6	-0.65	9.79	0.0014	0.0104	42.82
P08246	-0.97	10.48	0.0015	0.0106	62.31
P40939	1.24	9.25	0.0016	0.0115	56.92
O00329	-1.44	9.33	0.0019	0.0130	38.72
Q92952	-0.50	10.05	0.0020	0.0133	43.08
P0DJI8	-1.41	10.84	0.0029	0.0187	26.92
P04264	0.44	13.39	0.0031	0.0192	56.41

ABBREVIATIONS: European League Against Rheumatism (EULAR).

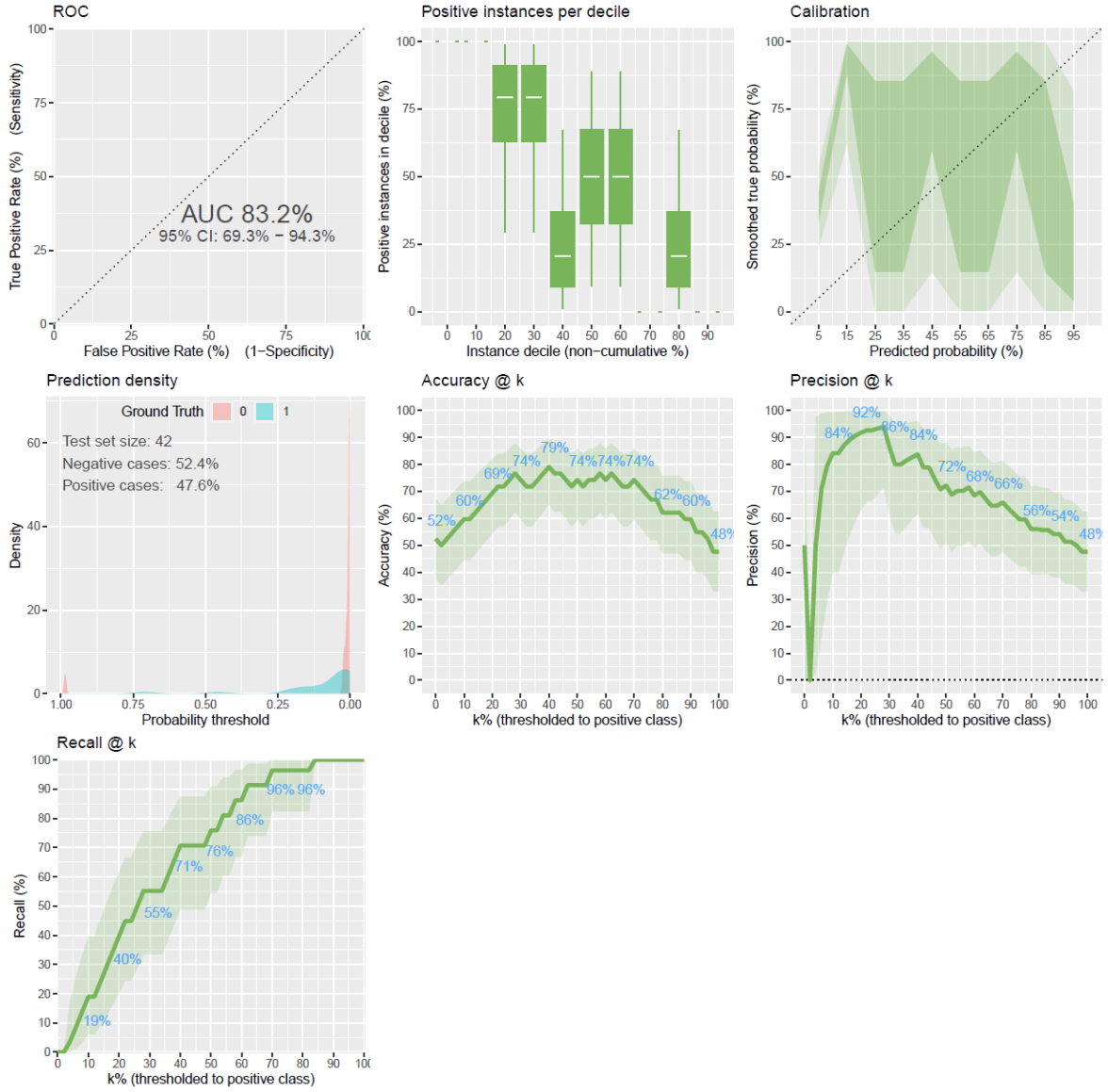
Table 5. Differentially expressed proteins between three and six months of treatment with etanercept in EULAR good/moderate responders.

Protein	**Log-fold change**	**Average expression**	**p-value**	**Adjusted p-value**	**% missing before imputation**
P46734	1.27	10.80	2.25E-18	4.87E-16	68.97
Q9HDC9	0.70	11.49	3.01E-14	3.25E-12	57.95
P04075	-1.77	9.76	4.69E-12	3.37E-10	84.62
Q9H4M9	1.78	12.20	3.36E-09	1.82E-07	72.82
P31749	-0.86	9.35	6.58E-09	2.84E-07	67.44
P06737	0.50	14.09	3.22E-08	1.16E-06	69.74
P02730	1.04	10.68	9.34E-08	2.88E-06	82.05
P14174	-0.52	10.30	3.09E-07	8.35E-06	71.54
Q15746	-0.67	12.81	3.18E-06	7.64E-05	54.36
Q9BPX6	-0.80	9.90	6.19E-05	0.0013	42.82
P43487	0.43	8.42	0.0002	0.0038	67.18
P05019	-0.53	9.48	0.0004	0.0078	78.97
P07384	-0.73	11.06	0.0008	0.0133	54.87
P20774	0.33	8.56	0.0022	0.0334	77.69
Q9UNW1	0.46	11.22	0.0029	0.0423	53.33
P08246	-0.85	10.45	0.0032	0.0438	62.31

ABBREVIATIONS: European League Against Rheumatism (EULAR).

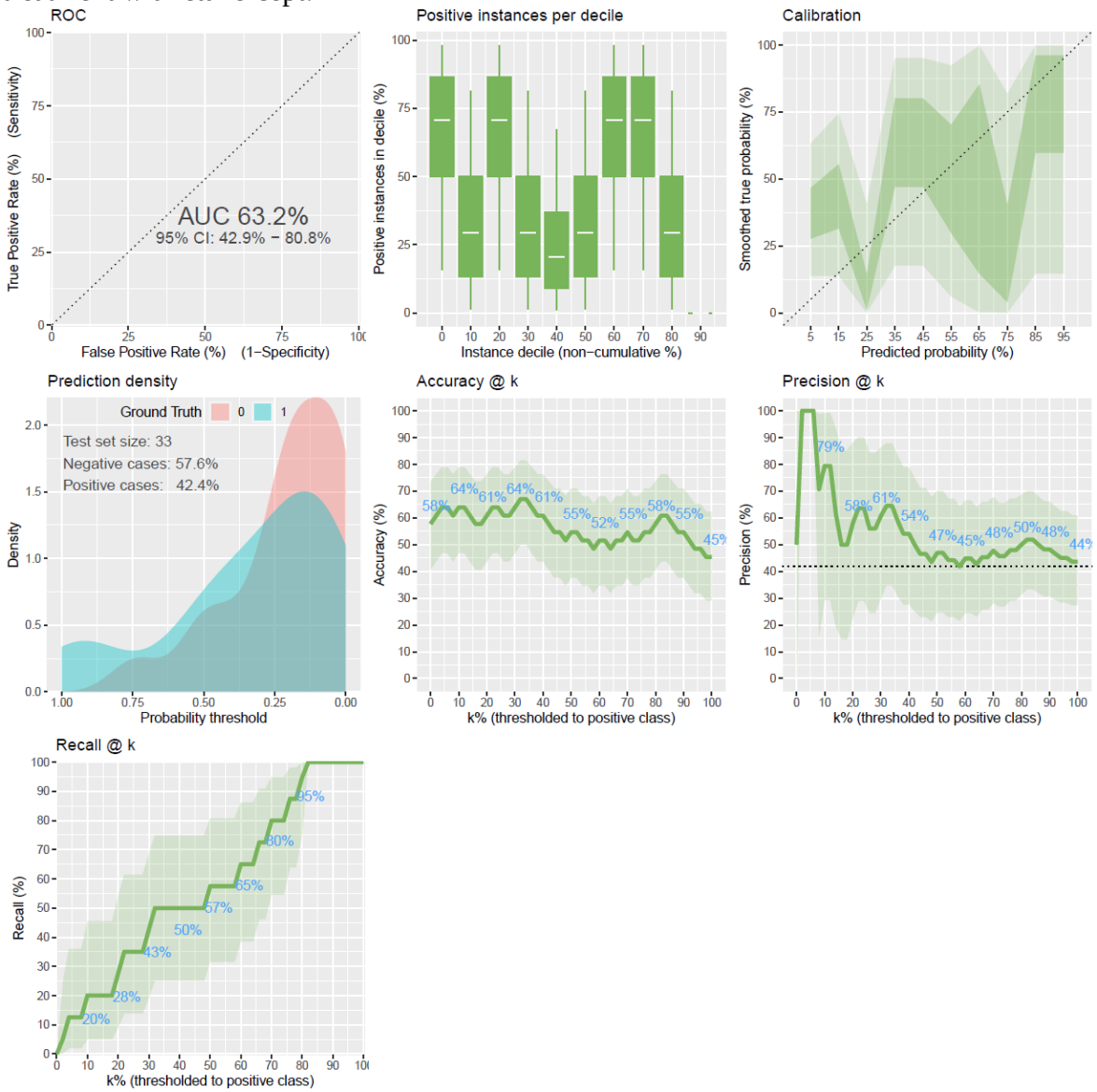
APPENDIX NINE: CLASSIFIER PLOTS FOR MODELS DEVELOPED DURING MACHINE LEARNING ANALYSIS

Figure 1. Baseline variables and prediction of poor EULAR response after three months of treatment with etanercept.



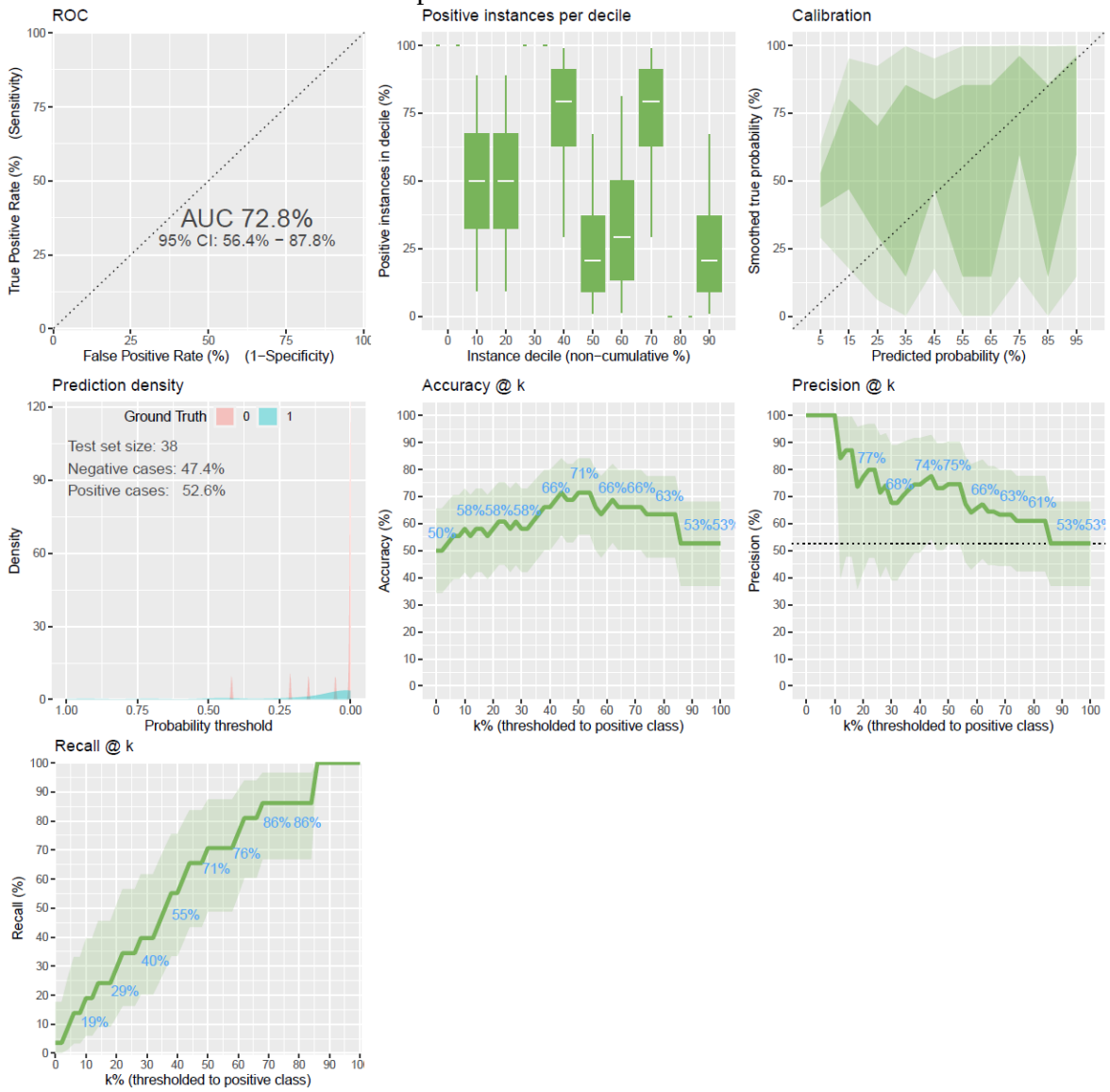
ABBREVIATIONS: Area under the receiver operating characteristic curve (AUC), European League Against Rheumatism (EULAR).

Figure 2. Baseline variables and prediction of poor EULAR response after six months of treatment with etanercept.



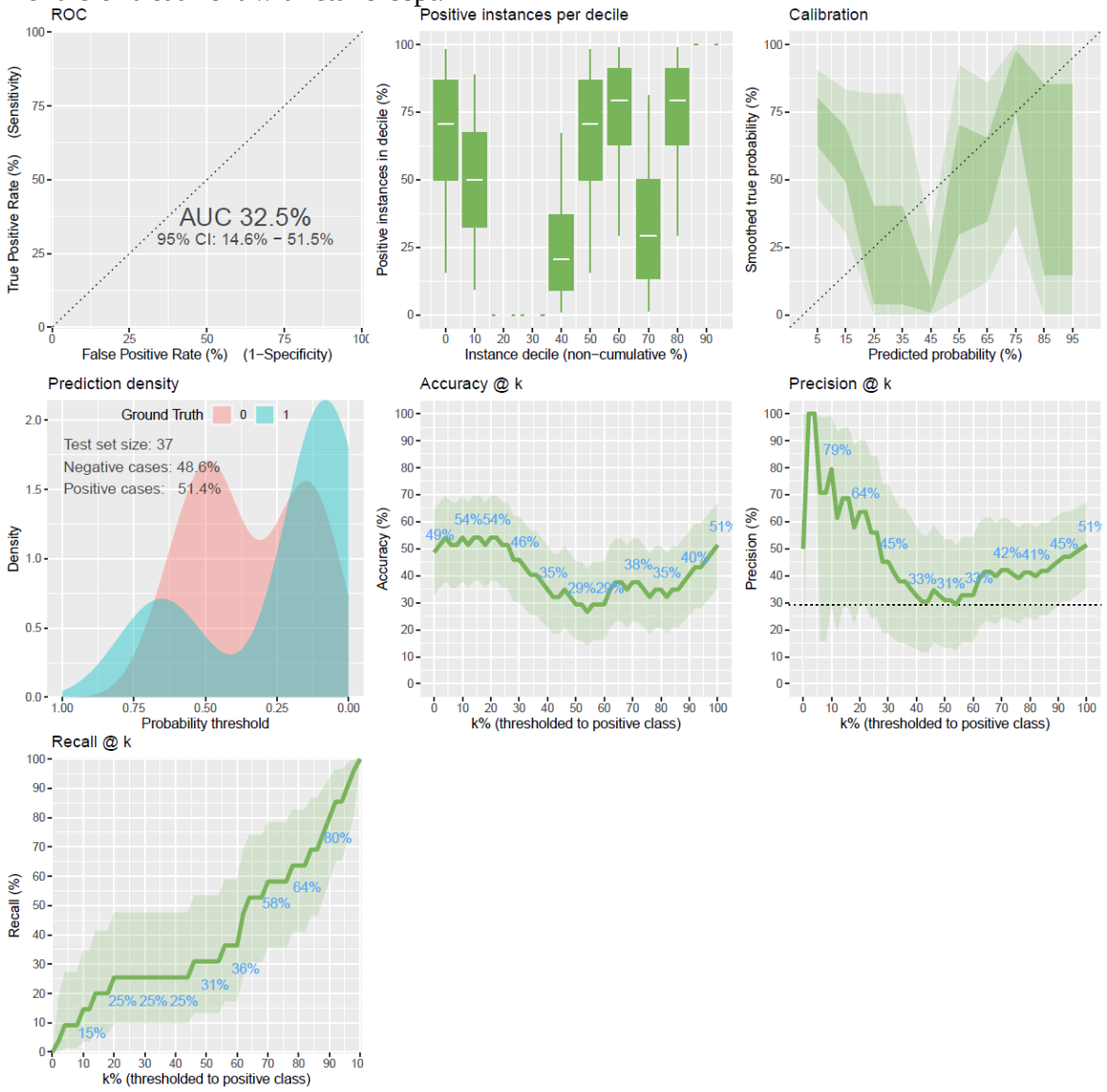
ABBREVIATIONS: Area under the receiver operating characteristic curve (AUC), European League Against Rheumatism (EULAR).

Figure 3. Baseline variables and prediction of failure to achieve MCID in DAS28 after three months of treatment with etanercept.



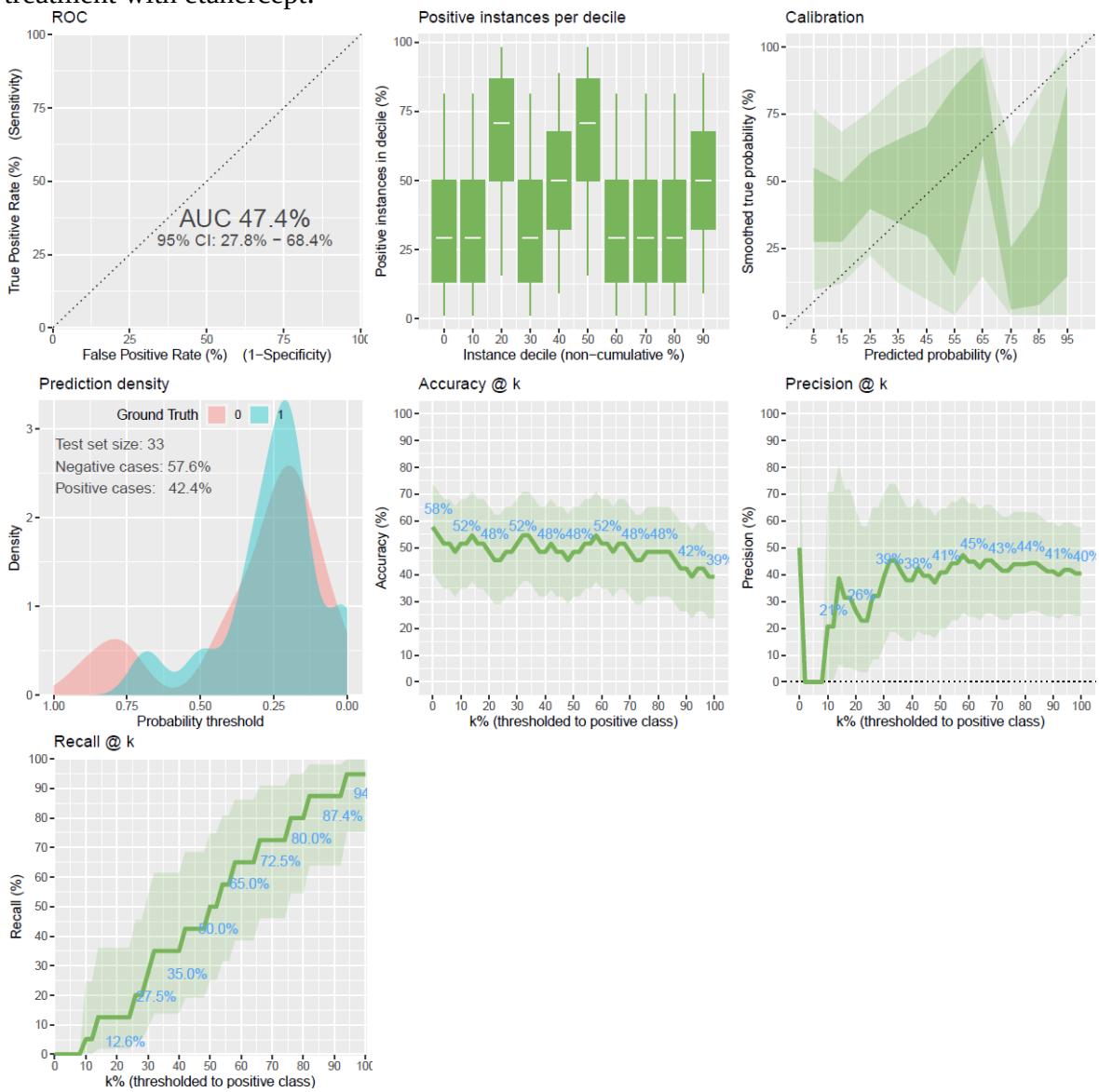
ABBREVIATIONS: Area under the receiver operating characteristic curve (AUC), Disease Activity Score of 28 Joints (DAS28), minimally clinically important difference (MCID).

Figure 4. Baseline variables and prediction of failure to achieve MCID in DAS28 after six months of treatment with etanercept.



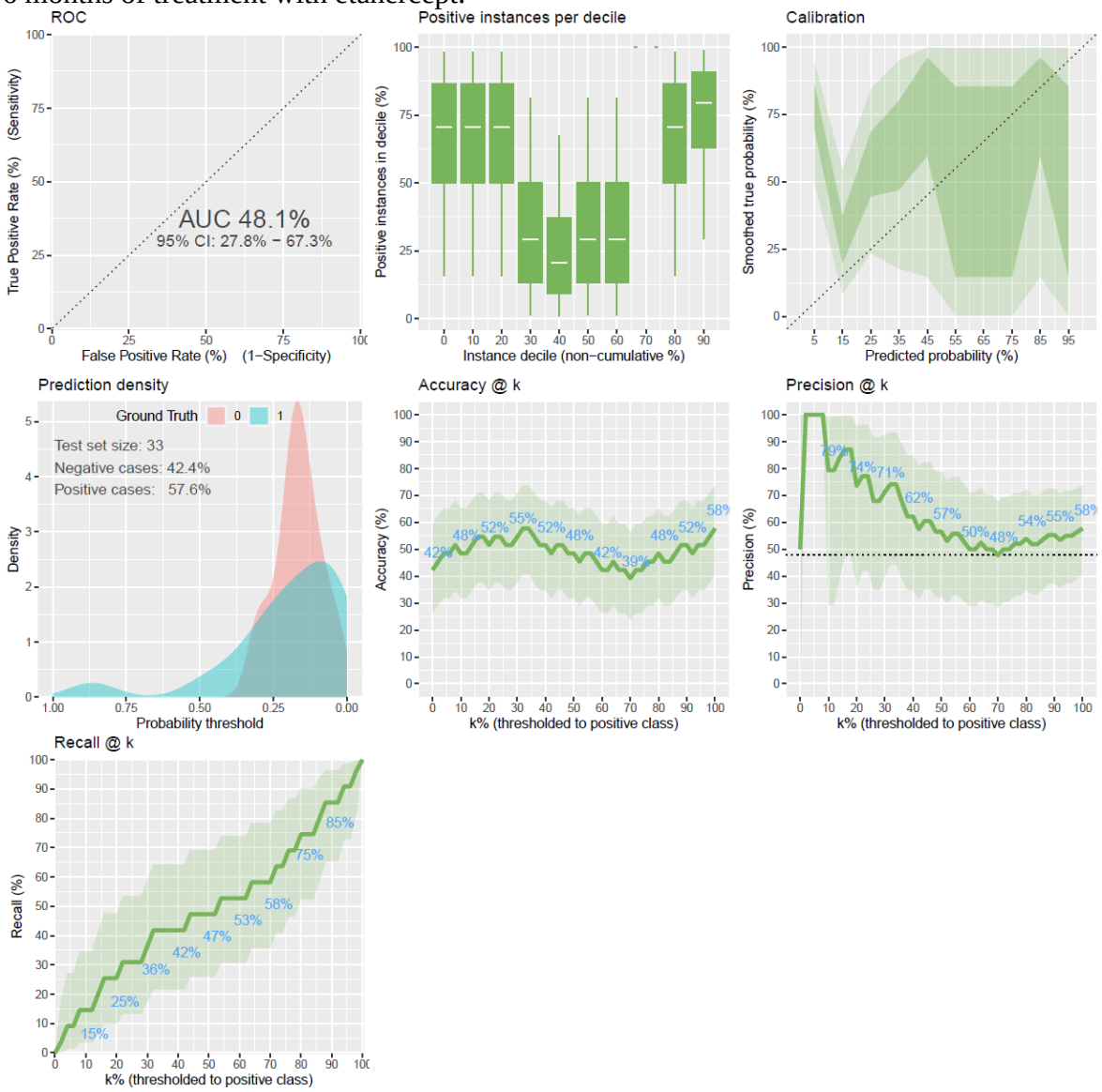
ABBREVIATIONS: Area under the receiver operating characteristic curve (AUC), Disease Activity Score of 28 Joints (DAS28), minimally clinically important difference (MCID).

Figure 5. Three-month variables and prediction of poor EULAR response after 6 months of treatment with etanercept.



ABBREVIATIONS: Area under the receiver operating characteristic curve (AUC), European League Against Rheumatism (EULAR).

Figure 6. Three-month variables and prediction of failure to achieve MCID in DAS28 after 6 months of treatment with etanercept.



ABBREVIATIONS: Area under the receiver operating characteristic curve (AUC), Disease Activity Score of 28 Joints (DAS28), minimally clinically important difference (MCID).

APPENDIX TEN: SIGNIFICANT *TRANS* PQTLS IDENTIFIED DURING PQTL ANALYSIS

Table 1. *trans* pQTLS before treatment with etanercept.

Top SNP	Protein	Total SNPs for protein	p-value	Adjusted p-value	Tissue of expression of corresponding eQTL
rs111538442	P08294	15	1.07E-24	6.28E-17	None
rs75213181	Q8NCW5	374	6.99E-24	6.28E-17	None

ABBREVIATIONS: Expression quantitative trait locus/loci (eQTL), protein quantitative trait locus/loci (pQTL), single nucleotide polymorphism (SNP).

Table 2. *trans* pQTLs after three months of treatment with etanercept.

Top SNP	Protein	Total SNPs for protein	p-value	Adjusted p-value	Tissue of expression of corresponding eQTL
rs9658041	O75369	358	2.53E-25	1.28E-18	None
rs77036310	P05186	13	1.74E-24	7.53E-18	None
rs79466292	P07900	1	1.28E-22	5.51E-16	None
rs143257148	P11766	440	2.91E-25	1.28E-18	None
rs1538304	P15144	529	1.75E-20	3.90E-14	None
rs6432156	P45983	8	2.21E-20	4.91E-14	None
rs527240128	Q14847	175	1.39E-20	3.90E-14	None
rs2069940	Q9UNN8	46	9.28E-22	3.94E-15	Cultured fibroblasts, skin, skeletal muscle, Epstein Barr virus-transformed lymphocytes, adipose tissue

ABBREVIATIONS: Expression quantitative trait locus/loci (eQTL), protein quantitative trait locus/loci (pQTL), single nucleotide polymorphism (SNP).