Diagnosis of paediatric infectious diseases and the utility of host immune signatures

Peter John O'Reilly

A thesis presented for the degree of
Doctor of Philosophy

Department of Paediatrics
Wolfson College
University of Oxford
May 2024

Do mo mháthair...

Ná bris do loirgín ar stól nach bhfuil i do shlí.

Acknowledgements

The results in this thesis come from two cohort studies which took place at Patan Hospital, Nepal.

The first cohort was recruited as part of the Pneumonia Aetiology in Nepal: Diagnosis with Immunology and Transcriptomics (PANDIT) Study. Funding was provided by the Wellcome Trust and the European Union's Horizon 2020 Programme through the PErsonalised Risk assessment in Febrile illness to Optimise Real-life Management (PERFORM) consortium. Funding for the mass spectrometry proteomics work and the Meso Scale Discovery cytokine panel was provided through the European Union Innovative Medicines Initiative Respiratory Syncytial Virus Consortium in Europe (RESCEU).

Funding for the second cohort study was from the European Union's Horizon 2020 Research and Innovation Programme through a multi-site consortium, Diagnosis and Management of Febrile Illness using RNA Personalised Molecular Signature Diagnosis (DIAMONDS). Separate funding for the molecular testing carried out at Patan Hospital was provided by a European Society For Paediatric Infectious Diseases-bioMérieux Small Grant Award on Diagnostics. I was supported during this DPhil by a Medical Sciences Graduate School Studentship, and I am grateful to the donors and the University of Oxford for providing this.

I would like to express my sincere thanks to all those who helped me throughout this DPhil. Thank you to my family and friends who provided great support and reminded me that there was a life outside of a DPhil.

To all of my colleagues at the University of Oxford, the Patan Academy of Health Sciences and Imperial College London, I am hugely grateful for all of the advice and assistance over recent years.

Thank you to the Pediatric Research Team at Patan Hospital who coordinated both of the cohort studies presented in this thesis and who continue to produce high-quality clinical research. I am indebted to the laboratory teams at the Oxford Vaccine Group and Patan Academy of Health Sciences for teaching me

invaluable laboratory skills. I have learned a huge amount and am grateful to the bioinformatics teams at the Oxford Vaccine Group and the Department of Infectious Disease, Imperial College London. I have been a part of the Overseas Team at the Oxford Vaccine Group since 2019, it has been a privilege working with each member of this team.

I will be forever grateful to my team of supervisors, Dr Daniel O'Connor, Dr Michael Carter, Professor Shrijana Shrestha, Professor Aubrey Cunnington and Professor Sir Andrew Pollard. I was very lucky to have this group of knowledgeable people to guide me through this process, I appreciate all of the wise words and helpful advice.

Finally, I would like to acknowledge the children who participated in the clinical studies and their parents and guardians, without whom the work presented in this thesis would not have been possible.

Declaration

I, Peter John O'Reilly, declare that this thesis is a product of my own work, and, where relevant, I have done my best to acknowledge the work of others at the appropriate points.

My role in the clinical studies

The enrolment to the first cohort study described in Chapter 2 was complete when I started this DPhil. I became involved at the analysis stage. I took the lead in analysing the plasma samples from study participants using two platforms for protein analysis, mass spectrometry proteomics and a cytokine panel. I performed the laboratory work for the cytokine panels with the help of research assistant, Zara Valliji. I re-analysed the RNA sequencing (RNA-seq) data to align with my analyses using the protein platforms.

Regarding the second cohort study included in this thesis and described in Chapter 6, I have had a prominent role from the inception of the study. I have been involved in planning the study design, drafting study documents, training and advice to the study team, quality control of the data, the molecular laboratory work at Patan Hospital and analysing the data.

COVID-19 pandemic mitigation

I began my DPhil work during the COVID-19 pandemic. The main effect this had on my DPhil was that the start of the cohort study described in Chapter 6 was delayed as the research team had to find a way to deliver the study with the safety of staff and participants in mind. Extra protocols had to be put in place to deliver the study safely and get agreement from hospital authorities, all of which delayed the study start. Restrictions on the transport of SARS-CoV-2 positive samples have also restricted the laboratory work we could perform on these samples.

Abstract

Background It can be difficult to distinguish between bacterial and viral infections in children. Current diagnostics lack the sensitivity to confidently rule out serious bacterial infections. The aims of this thesis include discovery of novel host biomarkers to identify bacterial infections, and evaluation of the role of expanded molecular testing in detecting pathogens.

Methods The data in this thesis come from two prospective cohort studies recruiting children admitted to Patan Hospital, a tertiary-level hospital in the Kathmandu Valley, Nepal. The first cohort study recruited children with lower respiratory tract infections (LRTIs). Whole blood samples from these children were sent for RNA-sequencing and plasma samples were sent for mass spectrometry proteomics and cytokine analysis. The LRTI cases were grouped together based on aetiology; RNA and protein levels were compared between different LRTI groups. Data were divided into training and test datasets; using a partial least squares approach signatures were identified to differentiate between bacterial and viral LRTIs. The data in the RNA and protein platforms were integrated to identify a multi-platform signature.

The second cohort study recruited children with any signs of infection. Causes of infection were described using routine diagnostics; samples were tested using four additional research molecular panels.

Results In the LRTI study, samples from 258 LRTI cases were tested. The median age was 2 years and 38% were female. The most common pathogens detected were RSV, influenza, and *Streptococcus pneumoniae*. Using a training subset of cases classified as bacterial, n = 47, or viral, n = 53, a three-gene RNA signature was identified; this signature identified bacterial infections, in the test dataset, with a sensitivity of 89% and a specificity of 100%. Integrating data from the RNA and protein platforms identified a signature with 100% sensitivity and 94% specificity for identifying bacterial LRTIs.

In the second cohort study, 574 cases were enrolled, median age was 3 years and 35% were female, 498 cases had additional molecular testing. Using routine diagnostics only, 28.7% (143/498) of cases were classified as having a confirmed bacterial or viral aetiology. After additional molecular testing, 45.6% (227/498) of cases had a likely pathogen identified. Molecular testing for RSV infection, influenza virus infection, *Neisseria meningitidis*, enterovirus infection and dengue fever showed potential diagnostic utility.

Conclusions The signatures identified in this thesis have the potential to improve the identification of serious bacterial infections. A small number of molecular tests which had the potential to alter clinical management were identified.

Contents

List of Figures

List of Tables

List of Abbreviations

A1AT	alpha-1-antitrypsin
A2GL	leucine-rich alpha-2-glycoprotein
AACT	alpha-1-antichymotrypsin
ACP3	acid phosphatase 3
ALOX5	arachidonate 5-lipoxygenase
ALOX5AP	arachidonate 5-lipoxygenase activating protein
ALPL	alkaline phosphatase, biomineralisation associated
ARL4C	ADP ribosylation factor-like GTPase 4C
AUC	area under curve
B3GNT5	UDP-GlcNAc:betaGal beta-1, 3-N-acetylglucosaminyltransferase 5
BH	Benjamini-Hochberg
bp	base pair
BP	Biological Process
BTB	Broad-complex, Tramtrack, and Bric-à-brac
BTS	British Thoracic Society
CAP	community acquired pneumonia
CAPN13	calpain 13
CC	Creative Commons

CCVTM	Centre for Clinical Vaccinology and Tropical Medicine
cDNA	complementary DNA
CI	confidence interval
CLUS	clusterin
COVID-19	coronavirus disease 2019
CO9	complement component C9
CPM	counts per million
CRP	C-reactive protein
CSF	cerebrospinal fluid
CSFs	colony-stimulating factors
CV	coefficient of variation
CXCL8	C-X-C motif chemokine ligand 8
CXCL9	C-X-C motif chemokine ligand 9
CXCL10	C-X-C motif chemokine ligand 10
DDA	data-dependent acquisition
DE	differential expression
DEA	differential expression analysis
DEG	differentially expressed gene
DGE	differential gene expression
DIAMONDS	Diagnosis and Management of Febrile Illness using RNA Personalised Molecular Signature Diagnosis
DIA	data-independent acquisition
DIABLO	Data Integration Analysis for Biomarker discovery using Latent cOmponents
DIA-NN	data-independent acquisition by neural networks

DNA	deoxyribonucleic acid
EBV	Epstein–Barr virus
EDTA	ethylenediaminetetraacetic acid
ELISA	enzyme-linked immunosorbent assay
EU	European Union
FAM151B	family with sequence similarity 151 member B
FC	fold-change
FDR	false discovery rate
FSTL4	follistatin like 4
G-CSF	granulocyte colony-stimulating factor
GELS	gelsolin
GM-CSF	granulocyte/macrophage colony-stimulating factor
GO	Gene Ontology
GSEA	gene-set enrichment analysis
GTP	guanosine triphosphate
Hib	*Haemophilus influenzae* type b
HIV	human immunodeficiency virus
HV434	immunoglobulin heavy variable 4-34
ICH-GCP	International Conference on Harmonisation-Good Clinical Practice
IDH3A	isocitrate dehydrogenase [NAD] subunit alpha, mitochondrial
IL	interleukin
IL-6	interleukin 6
IgG	immunoglobulin G
IgM	immunoglobulin M

IC	internal control
ICU	intensive care unit
IFN	interferon
IP-10	interferon gamma-induced protein 10
IQR	inter-quartile range
IRC	Institutional Review Committee
IRR	incidence rate ratio
KVD12	immunoglobulin kappa variable 1D-12
LBP	lipopolysaccharide binding protein
LC	liquid chromatography
LFC	log-2-fold-change
LILR	leukocyte immunoglobulin-like receptor
LILRA2	leukocyte immunoglobulin-like receptor A2
LLOD	lower limit of detection
LMICs	low- and middle-income countries
log2CPM	log2-counts per million
LRTI	lower respiratory tract infection
M-CSF	macrophage colony-stimulating factor
MIS-C	multisystem inflammatory syndrome in children
MMR	measles, mumps and rubella
mRNA	messenger RNA
MS	mass spectrometry
MSD	Meso Scale Discovery
MSigDB	Molecular Signature Database

MTB	*Mycobacterium tuberculosis*
NCBI	National Center for Biotechnology Information
NES	normalised enrichment score
NF-kappaB	nuclear factor-kappa-B
NHRC	Nepal Health Research Council
NK	natural killer
NS1	non-structural protein one
ORA	over-representation analysis
OxTREC	Oxford Tropical Research Ethics Committee
OR	odds ratio
PaO2	partial pressure of oxygen
pAUC	partial area under the ROC curve
PBMC	peripheral blood mononuclear cell
PC	principal component
PCA	principal component analysis
PCCC	Paediatric Chronic Complex Condition
PCR	polymerase chain reaction
PCV	pneumococcal conjugate vaccine
PCV10	10-valent pneumococcal conjugate vaccine
PERCH	Pneumonia Etiology Research for Child Health
PGRP2	N-acetylmuramoyl-L-alanine amidase
PICU	paediatric intensive care unit
PLTP	phospholipid transfer protein
PLS	partial least squares

PLS-DA	partial least squares - discriminant analysis
QC	quality control
QSOX1	sulfhydryl oxidase 1
RCT	randomised controlled trial
ReSVinet	Respiratory Syncytial Virus Network
RMSE	root mean squared error
ROC	receiver operating characteristic
RIF	rifampicin
RMSE	root mean squared error
RNA	ribonucleic acid
RNase	ribonuclease
RNA-seq	RNA sequencing
RPP	Respiratory Pathogen Panel
RSV	respiratory syncytial virus
rRNA	ribosomal ribonucleic acid
RT-PCR	reverse transcription polymerase chain reaction
SAA1	serum amyloid A-1
SAA2	serum amyloid A-2
SAGE	Strategic Advisory Group of Experts on Immunization
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SOFA	Sequential (Sepsis-related) Organ Failure Assessment
STGG	skim milk, tryptone, glucose and glycerol
TCV	typhoid conjugate vaccine
TDI	Target Discovery Institute

TLR	toll-like receptor
TLR6	toll-like receptor 6
TMM	trimmed mean of M values
TNF	tumour necrosis factor
TRAF	TNF receptor-associated factor
TRAF6	TNF receptor-associated factor 6
TRAIL	TNF-related apoptosis-inducing ligand
UI	uncertainty interval
UK	United Kingdom
VSN	variance stabilising normalisation
VWF	von Willebrand factor
WASH	water, sanitation and hygiene
WCC	white cell count
WHO	World Health Organisation
WTCHG	Wellcome Trust Centre for Human Genetics
XDR	extensively drug resistant
ZBT12	zinc finger and BTB domain-containing protein 12

Chapter 1

Introduction

1.1 Infections in children

Children regularly present to medical professionals with non-specific signs, and parents are often told that their child has a viral infection, without confirmation of this diagnosis. Mostly, these children will get better with minimal support, but occasionally children will deteriorate because the child's immune system cannot deal with the viral infection, lack of physiological reserve, or a serious bacterial infection was missed.

A high proportion of paediatric infectious presentations are linked to a viral illness, however, the consequences of missing a serious bacterial infection have led to high levels of antibiotic use in children.[3] As antibiotic resistance rates rise globally, the widespread use of antibiotics in medical and non-medical fields means that antibiotics will become less and less effective at treating serious bacterial infections.[4, 5, 6] Antimicrobial resistance has been highlighted by the WHO as one of the top 10 healthcare threats; new diagnostic tests with improved sensitivity for identifying serious bacterial infections could help to reduce antibiotic use and avoid a world where common antibiotics are no longer effective.[7]

Millions of children continue to die every year, many from preventable infectious causes. Globally, combined infectious diseases remain a leading cause of under five years of age mortality, along with prematurity and birth asphyxia. [8, 9]

Of the 5 million children, under five years of age who died in 2021, infectious causes were prominent with LRTIs (14.2%), malaria (8.8%) and diarrhoeal diseases (8.7%) making large contributions. These deaths are not spread evenly

throughout the world with 80% of deaths occurring in Sub-Saharan Africa and South Asia.[10]

With improvements to living standards and healthcare, infections do not need to be the most common cause of mortality in children. In a study of American children, injuries were the most common causes of death, with infections much less prominent.[11]

1.2 Lower respiratory tract infections

Respiratory tract infections are both common and important, and are often divided into upper and lower respiratory tract infections.[10, 12] Upper respiratory tract infections tend to be self-limiting, do not extend beyond the large airways, and will not be discussed further.[13] Improved understanding and diagnosis of LRTIs is a major focus of this thesis and are discussed in the following sections.

The lower respiratory tract is made up of the trachea, bronchi, bronchioles and alveoli. LRTIs can be classified based on which anatomical area is predominantly involved. Tracheitis is inflammation of the trachea, bronchitis is inflammation of the bronchi, bronchiolitis is inflammation of the bronchioles, and inflammation of the alveoli and surrounding tissues is classified as pneumonia. However, these distinctions are not always clear as a pathogen can infect multiple components of the respiratory tract simultaneously.[14]

Bronchiolitis is common in infants and is associated with viral infections. However, antibiotics are often given in severe cases due to the inability to clinically discriminate between viral bronchiolitis and bacterial LRTI.[15, 16]

Pneumonia can be caused by bacteria, viruses or fungi. A large case-control study of pneumonia cases in children under five years of age, requiring hospital admission, estimated that viruses account for 61.4% of pneumonia cases. Bacterial causes were estimated to account for 27.3% of cases.[17]

As pneumonia is the most common type of LRTI, and a major cause of mortality, different definitions have been suggested to identify pneumonia cases.

The British Thoracic Society (BTS) divides community acquired pneumonia (CAP) from hospital-acquired pneumonia. The BTS states that CAP can be defined as signs and symptoms of pneumonia in a previously healthy child due to an infection acquired outside of the hospital. Signs and symptoms of pneumonia include fever, cough, tachypnoea, breathlessness or difficulty breathing, wheeze, or chest pain. Consolidation on chest X-ray can confirm these findings but the BTS recommends that chest X-rays should not be performed in children who do not require hospital admission. The BTS also states that an ideal definition would include the isolation of a pathogen, but there are difficulties in isolating a pathogen, as discussed below.[2]

The World Health Organisation (WHO) defines pneumonia based on a set of clinical signs. In children aged 2 months to 5 years of age, pneumonia is defined as a cough with fast breathing and/or chest indrawing. [18] This definition is useful as it does not rely on X-rays or other diagnostic tests, and so can be used outside of hospital and in low-resource settings. The WHO definition is likely to overestimate pneumonia cases, but this definition is designed so that possible pneumonia cases can be quickly identified, and treatment given.[19]

Rather than defining LRTIs based on their location or clinical signs, it is impor-

tant clinically to identify whether LRTIs are bacterial or viral in origin. The BTS recommends treating all cases of pneumonia with antibiotics as bacterial pneumonia cannot be ruled out using current diagnostic tests.[2] A diagnostic test that could confidently rule out bacterial infection could change this advice.

For the rest of this thesis, LRTIs are generally grouped together and classified based on whether they are bacterial or viral in origin. Where cited evidence uses different terms for LRTIs I use the terms used by the authors, with the caveat that often different definitions are used for LRTIs within different publications.[2]

1.2.1 Paediatric lower respiratory tract infection epidemiology

Respiratory infections are common in children, a prospective cohort study estimated that children have a mean of 22 respiratory infections from birth to 12 years of age, with most occurring in the first five years of life.[12] LRTIs are one of the main causes of morbidity and mortality in young children, especially in low- and middle-income countries (LMICs).[20] In 2021, LRTIs contributed to almost 800,000 deaths globally in children under five years of age.[10] Globally, in children under five years of age, there were 103.3 (95% uncertainty intervals (UIs) 92.8-114.0) deaths per 1,000 LRTI cases in 2016. Children under one year of age had the highest reported mortality from LRTIs.[21]

Pneumonia, bronchiolitis and croup make up the majority of acute respiratory cases.[22] Viruses are linked to a higher proportion of paediatric LRTIs, particularly in the first two years of life where viral bronchiolitis is common.[3, 17]

Bacterial LRTIs, while less common, are associated with worse outcomes.[20, 23] A US multi-centre study which recorded paediatric pneumonia data over a 12-year period reported that bacterial CAP was associated with longer hospital stays and a higher risk of intensive care admission.[23]. Due to the risks associated with missing a bacterial LRTI, many children are treated with antibiotics without strong evidence of a bacterial cause.[3]

Bacteria or viruses cause the vast majority of LRTIs in immunocompetent children. Diagnosing LRTI aetiology often relies on identifying the pathogen, which can be difficult. Currently, available infectious disease diagnostics lack the sensitivity to confidently rule out serious bacterial infections.[2] Traditional culture-based diagnostics do not provide answers in time to effect initial clinical management decisions.[24] These issues are discussed further in the Section 1.4. Where a pathogen is found, *Streptococcus pneumoniae* and *Haemophilus influenzae* type b (Hib) are the most commonly identified bacterial causes of LRTIs worldwide, followed by *Staphylococcus aureus* and *Klebsiella pneumoniae*.[25] However, this is an evolving picture as the vaccination status of a population alters the pathogen profile of pneumonia in childhood. In a large case-control study of children under five years of age with pneumonia, different bacteria were found on blood culture depending on the study site, see Figure 1.1.[17]

Vaccinations are available to prevent many major LRTI-associated pathogens including *S. pneumoniae*, Hib infections, pertussis, measles and influenza viruses.[26] pneumococcal conjugate vaccines (PCVs) have been shown to have an impact on pneumonia cases. A systematic review of interventions to address childhood pneumonia reported that across six randomised controlled trials (RCTs) PCV

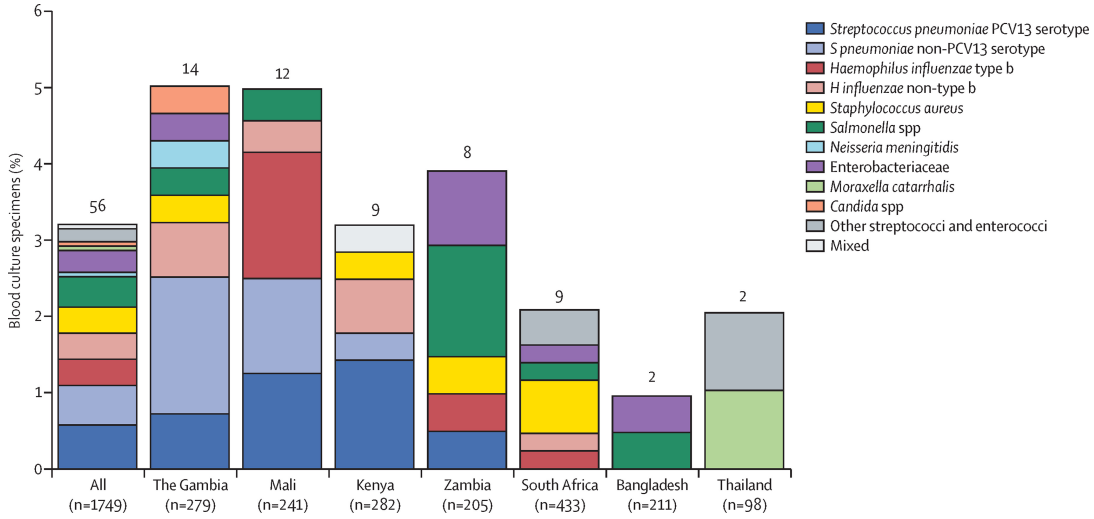


Figure 1.1: Results from a large case-control study looking at the causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia - PERCH Study Group. Blood culture results from children under five years of age with pneumonia. Results presented are from cases with a chest x-ray consistent with pneumonia and without HIV infection.[17] Figure reproduced under Creative Commons (CC) license, DOI:https://doi.org/10.1016/S0140-6736(19)30721-4.

introduction was associated with a 29% reduction in radiologically-confirmed pneumonia. Across four RCTs and two case-control studies, the introduction of Hib vaccinations resulted in a 6% reduction in severe pnuemonia and a 7% reduction in pneumonia mortality.[27]

Viral LRTIs are more common than bacterial with at least 26 viruses reported to be associated with CAP in adults or children.[28] Respiratory syncytial virus (RSV) is the pathogen most commonly linked to paediatric LRTI cases; in a multi-site case-control study of paediatric pneumonia 31% of all identified pathogens were RSV.[17] Another study in low-resource settings attributed RSV to 15-40% of admitted paediatric pneumonia or bronchiolitis cases.[25] Other viruses which have been reported to make a substantial contribution to viral LRTIs include influenza, parainfluenza, human metapneumovirus and rhinovirus.[17, 29] The odds ratio for different pathogens in a large case-control study of children with pneumonia requiring hospital admission are shown in Figure 1.2. RSV has the highest odds ratio (OR). Human metapneumovirus, certain influenza and parainfluenza viruses and rhinovirus also have significant odds ratios. However, the results would suggest less confidence in rhinovirus as a cause of LRTIs. While the OR reaches significance for rhinovirus, 1.32 (95% confidence interval (CI) 1.12, 1.54), the level of rhinovirus detection is only slightly higher in cases than controls, with approximately 20% detection in cases and controls.[17]

Vaccines are also available to prevent viral LRTIs. Influenza vaccines are commonly used to prevent infection and have been introduced to national immunisation programmes in many high-income countries.[26] Recent developments in

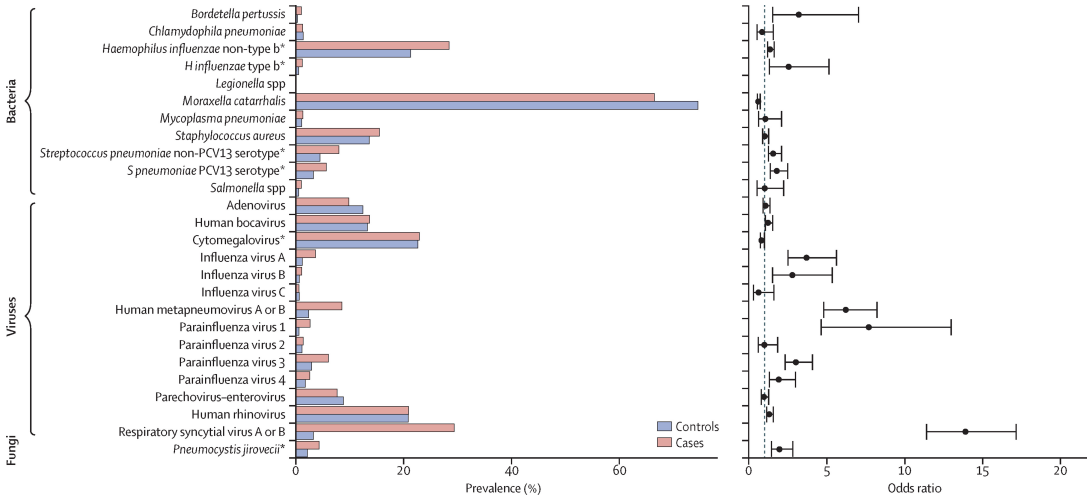


Figure 1.2: Nasopharyngeal-oropharyngeal (NP-OP) pathogen prevalence and adjusted ORs in children with severe pneumonia requiring hospital admission from a large case-control study examining the causes of pneumonia in children without HIV infection from Africa and Asia. Nasopharyngeal-oropharyngeal (NP-OP) pathogen prevalence and adjusted in children under five years of age admitted to hospital with pneumonia, with a chest X-ray consistent with pneumonia and no known human immunodeficiency virus (HIV). Results from a multi-centre case-control study reported by the PERCH Study Group.[17] Figure reproduced under CC license, DOI:https://doi.org/10.1016/S0140-6736(19)30721-4.

RSV vaccines mean that the burden of RSV may be significantly reduced in the coming years.[30, 31]

1.3 Common and important infections in Nepal

The data in this thesis come from studies in Nepal. The early thesis chapters focus on LRTIs, but Chapter 6 presents results from a cohort of children with a broader range of febrile illnesses.

Many of the causes of infection in Nepali children are caused by pathogens that are present in most global settings, including the common respiratory pathogens discussed in previous sections. LRTIs are an important causes of mortality in Nepali children. The mortality associated with LRTIs in Nepal is higher than the global average; there were 119 (95% UIs 81.7-169.6) deaths per 1,000 LRTI cases in children under five years of age in the same year. *Streptococcus pneumoniae* is the leading cause of LRTI mortality globally.[32] The 10-valent pneumococcal conjugate vaccine (PCV10) was introduced to Nepal's routine vaccination program during 2015. The LRTI cohort study described in Chapter 2 began recruitment prior to PCV10 introduction and continued during the initial years of PCV10 use. This means that the aetiology of LRTIs was likely changing as the cohort study was enrolling participants.

Data from adult studies, and studies focused on individual pathogens, highlight other important diseases in this Nepali context. Many pathogens found in Nepal are important in most settings globally, such as common respiratory viruses, e.g., RSV, influenza, and important bacteria, e.g. *S. pneumoniae* and

N. meningitidis.[17, 33] Other pathogens which are important in Nepal include dengue virus, *Salmonella* enterica serovar Typhi, *Salmonella* enterica serovar Paratyphi, *Rickettsia* and *Leptospira* species.[33, 34, 35, 36, 37] Data regarding the prevalence of these conditions in Nepali children is limited, and prevalence rates are inferred from adult studies.[38, 37] However, a systematic review of scrub typhus cases in Nepal included children. The seroprevalence of scrub typhus in Nepal was 19.31% (95% CI: 18.55–20.08%), and the highest prevalence was seen in children and young people under 20 years of age.[39]

Dengue fever is discussed here as an outbreak of dengue occurred during recruitment to the cohort study described in Chapter 6. Dengue virus is a flavivirus, primarily carried by the female *Aedes aegypti* mosquito.[40] In 2019, there were an estimated 56 million dengue cases globally, and up to 40,000 deaths.[41] Most dengue infections are asymptomatic, 70-80%. However, clinical presentations range from a mild febrile illness to dengue haemorrhagic fever and shock.[42, 43]

Dengue virus has four serotypes, 1-4, these serotypes are genetically similar but antigenically distinct.[44] Following the first dengue infection, a person is protected against the dengue serotype caused by that infection only. A second infection with a different dengue serotype carries with it a much higher risk of severe dengue infection.[45] Currently, there is no reliable biomarker to predict severe dengue disease.[46]

Recent progress in vaccination studies offers an avenue to reduce the burden of dengue disease. To date, positive results have been reported for three different tetravalent vaccines in phase III trials.[47, 48, 49] The WHO Strategic Advisory Group of Experts on Immunization (SAGE) has recommended that persons

over nine years of age who have evidence of previous dengue infection should receive three doses of one of these dengue vaccines, Dengvaxia. Evidence of previous infection is required as Dengvaxia was shown to increase the risk of severe dengue in individuals who experience their first dengue infection after vaccination. Dengvaxia was introduced to the routine immunisation programme in the Philippines in 2016 but was subsequently withdrawn when it was found that there was an increased risk of severe dengue infection following vaccination in children 2 to 16 years of age.[48] A second vaccine, TAK-003, has also been recently recommended for use in children aged 6 to 16 years of age in endemic regions with high transmission.[50] More work is needed before dengue vaccines can be offered to the youngest children.

As global temperatures increase, the vector for dengue fever is being found in new areas, and cases of dengue fever are spreading to new locations. The incidence of dengue is increasing in the South Asia region which includes Nepal.[41]

In Nepal, the first case of dengue was reported in 2004. Since 2010, dengue outbreaks have been reported yearly in the warmer low-land regions.[35] The geographical distribution of dengue outbreaks in Nepal is increasing. Dengue cases had previously been relatively rare in the Kathmandu Valley but at the end of 2022 there was a large number of cases reported in Kathmandu.[51] All four serotypes have previously been reported in Nepal.[52]

Should outbreaks of dengue occur regularly in Kathmandu we can expect an increase in the number of severe dengue cases presenting to hospitals. Cases of severe dengue are more likely following a second infection with a different dengue serotype.[53] Dengue is likely to become an even more important global

pathogen in the coming decades.

1.4 Diagnostics in different clinical presentations

Improving outcomes in children with infections requires improvements in diagnosis and management. Improved diagnostics would allow clinicians to more accurately assign the aetiology of disease, and offer more targeted treatment. Better diagnostics means finite healthcare resources can be targeted to the patients who need them most.

As the results presented in this thesis focus on LRTIs and infections relevant to a Nepali context, I will discuss paediatric infectious disease diagnostics with a focus on LRTIs, and later, other infections relevant to Nepal.

1.4.1 Lower respiratory tract infection diagnostics

The investigations performed in paediatric LRTIs depends on the setting and whether the child is unwell enough to be admitted to hospital. In paediatric LRTIs, samples are rarely obtained from the site of infection. Samples from other sites are used to infer the pathogen causing infection in the lower respiratory tract.

Depending on the setting, samples from the upper respiratory tract are often taken in LRTIs. Nasal, throat or nasopharyngeal samples have the advantage of being easily obtained. Organisms in the upper respiratory tract can represent the

cause of LRTIs.[17, 54] These samples are often tested using molecular diagnostics which are not routinely available at many centres.

Commercial panels are available to test a wide range of potential pathogens in upper respiratory tract samples.[55] Care must be taken in interpreting positive results when many molecular targets are tested simultaneously. Large case-control studies found that detection of RSV, influenza, parainfluenza viruses and human metapneumovirus in the nasopharynx was associated with community acquired pneumonia. However, children often carry potentially pathogenic organisms in their upper respiratory tract, and these may not be the cause of their LRTI. In the case-control studies, several viruses, including adenovirus, bocavirus and non-SARS coronaviruses were were detected at similar levels in cases and controls.[17, 54]

Healthy children commonly carry pneumococci in their upper airway, and detection of pneumococci in upper airway samples is not used as a diagnostic marker. However, the detection of certain pneumococcal serotypes is rare in healthy children, and more common in children with invasive pneumococcal disease. Serotypes 1 and 5 are common serotypes found in children with invasive pneumococcal disease, however, these are rarely found in healthy children. [56, 57, 58]

A clinical practice guideline from two infectious disease societies in America recommends only taking blood cultures when a child with pneumonia has failed initial antibiotic treatment or is unwell enough to require hospital admission.[59] This is common practice in many settings, with blood cultures only taken if initial treatment fails, and if the facilities to take blood cultures are available.

While a positive blood culture can offer a confident diagnosis, blood cultures have a low sensitivity in LRTIs, and it can be difficult to obtain an adequate volume of blood in young children. Blood culture positivity in children hospitalised with LRTIs ranges from 2% to 5%.[17, 60, 61]

During pneumonia we expect infectious pathogens to be present at the alveolar level. However, it is difficult to obtain a sample directly from the lower respiratory tract. Bronchoalveolar lavage is performed during bronchoscopy or while patients are intubated. While useful to add diagnostic information in severely unwell cases, this is not a practical procedure in most LRTIs.[62]

Pleural fluid sampling is rarely performed in LRTIs due to concerns around acceptability of the procedure and perceived safety concerns. In clinical practice, pleural fluid is not sampled unless a substantial pleural effusion or empyema is present. A study in The Gambia enrolled a cohort of children with lobar pneumonia or empyema; the authors reported that blood culture alone yielded a potential pathogen in 18% of cases, whereas pathogen detection increased to 52% when percutaneous lung or pleural fluid aspiration was added to routine clinical investigations. [63]

Chest X-ray is used to diagnose pneumonia radiologically, but chest X-rays are not recommended for patients with LRTIs that are well enough to be managed in the community.[2, 59] Even when infants with bronchiolitis require admission, chest X-rays are not routinely recommended.[64]

Toikka *et al.* reported that in paediatric pneumonia patients with marked X-ray changes, a bacterial pathogen was detected in 69% of cases, and a viral pathogen in 17% of cases.[65]. Similarly, Virkki *et al.* reported that alveolar infiltrates

were seen in 71% of paediatric pneumonia cases where a bacterial pathogen was detected and in 49% of cases where a viral pathogen was detected.[66] These studies show that while a higher proportion of bacterial cases have significant changes on chest X-ray, a substantial minority of viral pneumonia cases also have similar X-ray changes.

Co-infections

Even when a likely pathogen is found, co-infections further complicate the diagnostic picture. The presence of a pathogenic virus does not rule out a bacterial co-infection. In an American surveillance study of 6,769 children hospitalised with laboratory-confirmed influenza, 2% also had a positive blood culture for a bacterial pathogen.[67] And as discussed earlier in this section, the low sensitivity of blood cultures means that this is likely an underestimate of the actual incidence of co-infections.[17, 60, 61]

Children are also at higher risk for a bacterial infection following a viral illness. Following a review of pneumococcal cases across six years at an American hospital, Ampofo *et al.* reported a positive correlation between invasive pneumococcal disease and RSV infections for up to four weeks after peak incidence of RSV cases.[68] In a study of hospital records in the USA across 36 states, Weinberger *et al.* found a temporal association between peaks in hospital admissions of viral cases (RSV and influenza) and pneumococcal pneumonia in children under 2 years of age. The authors report that 20% of the hospitalised pneumococcal cases were attributable to increased RSV activity.[69] These studies again point to the need for improved diagnostics to identify co-infections that can complicate

clinical management.

1.5 Molecular diagnostics

Molecular testing has been used in high-resource settings for several years to augment traditional culture-based diagnostic tests. Molecular tests, in general, have increased sensitivity compared with culture-based diagnostics and provide faster results.[70, 71]

During the coronavirus disease 2019 (COVID-19) pandemic, molecular testing for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) became an important test in many settings. To allow for this, the facilities to perform polymerase chain reaction (PCR) testing became available in many new settings. Staff were trained in nucleic acid extraction and use of PCR machines.[72, 73] In the wake of the COVID-19 pandemic the ability to perform molecular testing in these centres can be used to identify different pathogen targets. In Chapter 6, I describe the use of additional molecular testing panels in a Nepali setting where access to PCR diagnostics became available during the COVID-19 pandemic.

Molecular panels allow for the detection of multiple pathogen targets within one test. Syndromic panels test molecular targets likely to be associated with particular clinical presentations. Panels for bacterial sepsis, meningitis and LRTIs are used in hospitals in high-resource settings.[74, 75]

The benefits to using molecular panels include smaller sample volumes required to test multiple targets, reduced laboratory time and a potential financial benefit compared with testing individual targets.

The disadvantages include that the set of pathogen targets on molecular panels may not be suitable for the population being tested. It may be difficult to interpret molecular panel results where several different pathogens are positive in the same individual. The utility of molecular panels is discussed in more detail in Chapter 6, where molecular panels are used to augment routine investigations in a cohort study. The value of the molecular targets tested depends on which pathogens are relevant to the population, the clinical picture and if the appropriate sample type was tested. These issues are discussed in the following sections.

Molecular testing of respiratory samples

As discussed in Section 1.4, the detection of certain respiratory viruses has been associated with LRTIs. In large case-control trials of paediatric LRTIs, higher detection rates were found in cases for RSV, influenza, parainfluenza viruses and human metapneumovirus compared with healthy controls.[17, 54]

The Shah *et al.* study of molecular results in a cohort of European children with febrile illnesses compared children with bacterial and viral infections. I have used a similar classification system in Section 2.2.3. Children were classified into bacterial or viral groups based on routine investigations and then molecular testing was carried out on respiratory and blood samples. They used the Luminex NxTAG Respiratory Pathogen Panel to look for several molecular targets, including RSV, influenza, parainfluenza, human metapneumovirus, non-SARS-CoV coronaviruses and rhinovirus, including RSV, influenza, parainfluenza, human metapneumovirus, non-SARS-CoV coronaviruses and rhinovirus, see Figure 2.2. Influenza A, B and RSV were the only viral targets which had a

significantly higher odds ratio of being detected in the viral group compared with the bacterial group.[76] The odds of *Mycoplasma pneumoniae* detection in respiratory samples were higher in the bacterial compared with the viral group. However, *M. pneumoniae* detection in the upper airway was not associated with paediatric pneumonia in the large case-control studies discussed previously.[17, 54]

Molecular testing in blood for bacterial infections

Molecular testing can provide a faster diagnosis, with increased specificity when compared with culture-based or serological testing. However, detection of nucleic acid in the blood does not always mean that the active pathogen has been identified. Shah *et al.* reported results from molecular testing in a large cohort of European children admitted to hospital with febrile illnesses. Blood samples from febrile children were tested for a panel of common bacterial PCR targets. Of the healthy control cases, 8% (79/996) had a positive result for at least one bacterial target.[76] In the cohort study described in Chapter 6 samples are tested for a similar panel of bacterial PCR targets, and molecular panel testing is discussed further in Section 6.1.3.

1.6 Host biomarkers

Considering the above difficulties with identifying a pathogen in paediatric infections, it is reasonable to examine different approaches. One such approach is to look for changes in the host immune response, which consistently occur during infection. C-reactive protein (CRP) and, more recently, procalcitonin are

routinely used to aid diagnosis and classification of infections, depending on the setting. CRP and procalcitonin can help to differentiate between bacterial and viral LRTIs, but these tests have their limitations.

CRP is an acute phase protein produced in the liver.[77] In an Australian study of hospitalised children, CRP, $\geq$ 72 mg/L had a sensitivity of 75% and a specificity of 84% for differentiating between bacterial and viral pneumonia, taking into account the difficulties with confidently assigning pneumonia aetiology.[78] Virkki *et al.* reported that in children with pneumonias, CRP levels were significantly different between bacterial and viral groups at cut-offs of >40 mg/L, >80 mg/L, and >120 mg/L, however the sensitivity for detecting bacterial pneumonia was too low for use in clinical practice.[66] Largman-Chalamish *et al.* have suggested that combing the duration of symptoms with CRP level improves the differentiation between bacterial and viral infections.[77]

Müller *et al.* report that procalcitonin outperforms CRP in identifying patients with sepsis, sensitivity of 89% and specificity of 94% using a procalcitonin cut-off of 1 ng/ml.[79]

A meta-analysis of studies comparing procalcitonin and CRP to identify bacterial infections in patients with fever of unknown origin reported a pooled sensitivity for procalcitonin of 0.76 (95% CI, 0.72 to 0.80) and pooled specificity of 0.69 (95% CI, 0.64 to 0.72) to identify bacteraemia. For CRP, the pooled sensitivity was 0.85 (95% CI, 0.78 to 0.91) and pooled specificity of 0.80 (95% CI, 0.65 to 0.90).[80]

Toikka *et al.* report that while CRP and procalcitonin were raised in bacterial LRTIs, the values overlapped significantly and could not be used to reliably

differentiate between bacterial and viral LRTIs.[65] There have been studies of many other potential biomarkers including TNF-related apoptosis-inducing ligand (TRAIL), interleukin 6 (IL-6) and interferon gamma-induced protein 10 (IP-10).[81, 82] A combination of biomarkers may offer higher diagnostic accuracy. The combination of TRAIL, IP-10 and CRP improved accuracy when differentiating bacterial and viral infections, compared with these biomarkers individually.[82, 83, 84]

While some of these biomarker studies have shown promising results, many individual biomarkers lack the sensitivity or specificity to alter decisions in clinical cases. An expert consensus publication outlined the target product profile for a diagnostic test to differentiate between bacterial and non-bacterial infections in low-resource settings, with the aim of reducing antibiotic overuse. The recommended key performance requirements of any new test were a sensitivity >90% and a specificity >80%. This consensus paper also outlined other ideal characteristics of a new test including a test which can be used in community healthcare centres with less than two days of training required and time-to-result of less than 10 minutes.[85]

Different biomarkers at various thresholds have been used to improve classification of infection. As host biomarkers are a key subject for this thesis, I discuss novel host biomarkers in Section 1.7.

1.6.1 Infectious disease diagnostics in a Nepali context

The site of the studies presented in this thesis, Patan Hospital, relies on the culture-based diagnostics discussed in Section 1.4 for a pathogen diagnosis. During the COVID-19 pandemic Patan Hospital acquired molecular testing capabilities but this is currently only used for SARS-CoV-2 testing in infectious disease diagnostics.

The limitations of culture-based diagnostics combined with Nepali hospitals not having access to routine molecular testing mean that there is a paucity of data regarding the causes of infection in Nepal.[38]

Nepal's geography is diverse with low-lying areas where malaria and dengue are common, and regions at higher altitude where vector-borne diseases are non-existent. Historically, the Kathmandu valley, where the cohorts in this thesis were recruited, at an altitude of 1,324 metres was spared dengue virus outbreaks, but this picture has changed in recent years.[86]

A large outbreak of dengue occurred in the Kathmandu Valley in the second half of 2022.[51] This outbreak occurred during the recruitment period for the cohort study presented in Chapter 6

Dengue virus infection can be asymptomatic, present with flu-like non-specific symptoms or as severe dengue disease. Dengue fever can present with similar signs to other common infections which makes diagnosis difficult. Mild symptoms can progress quickly to severe dengue shock or haemorrhagic fever so early diagnosis is important.[87] There are three important types of dengue fever diagnostics, molecular, antibody and antigen, which are discussed in detail in

Section 6.1.4. Routine testing for dengue fever at Patan Hospital involves NS1 antigen, immunoglobulin M (IgM) and immunoglobulin G (IgG) dengue antibody testing. However, molecular testing is considered the most sensitive and specific test available for dengue fever.[1]

Enteric fever is an important infectious syndrome in South Asia but is difficult to diagnose. Enteric fever is caused by ingesting food or water contaminated with *Salmonella enterica* Typhi (*S.* Typhi), typhoid fever, or *Salmonella enterica* Paratyphi (*S.* Paratyphi) A, B, and C, paratyphoid fever. Despite global improvements in water, sanitation and hygiene (WASH) and new vaccinations, enteric fever remains an important disease in many settings.[88] The incidence of typhoid fever in children aged 5 to 15 years at a Kathmandu site was estimated to be 570 (95% CI, 310–1124) per 100,000 person-years.[89] Enteric fever is likely an important cause of fever in the cohorts presented in this thesis.

Blood culture in enteric fever has low sensitivity, estimated to be 50% but is likely lower in children, and does not provide a result in time to decide on initiation of treatment. The number of viable bacteria in the blood is low in typhoidal salmonella infections, $\leq$1 colony forming unit per millilitre. Bone marrow cultures offer the highest sensitivity, over 80%, but are rarely performed.[90] Nucleic acid testing for *Salmonella*-specific DNA sequences provides quicker results but with sensitivity levels ranging from 40-100%. Often several millilitres of blood are required for enteric fever diagnosis which can be problematic in paediatric patients.[91]

Other potentially important causes of infection, such as leptospirosis, scrub typhus and rickettsial disease are less commonly diagnosed, with limitations of available

diagnostics likely a contributing factor; diagnosis of these conditions relies on a combination of culture, serological and nucleic acid testing if available.[92, 93, 94]

Antibody testing is useful in certain situations. Specific antibodies rise after exposure to a given antigen. For example, IgM antibodies are useful in diagnosing hepatitis A, leptospirosis and dengue.[92, 93, 94]

Antigen testing can be useful, where levels of certain antigen rise quickly during acute infection. SARS-CoV-2 antigen testing was extensively used during the COVID-19 pandemic.[95, 96] non-structural protein one (NS1) antigen testing for dengue fever diagnosis is discussed in Section 6.1.4.

Molecular testing capabilities have become more common in Nepali hospitals in the wake of the COVID-19 pandemic.[73] This should allow for a better understanding of infectious disease causes in Nepal.

PCR is considered the gold standard for dengue fever testing and is discussed in Section 6.1.4. Nucleic acid testing for *Salmonella*-specific DNA sequences has shown variable sensitivity levels, likely due to the low volume of *Salmonella* organisms in blood and is often not available in low resource settings.[91]

Rickettsial DNA is difficult to detect in blood. This is because rickettsial species commonly infect endothelial cells and not circulating blood cells; this leads to low levels of circulating *Rickettsial* species.[97] Sensitivity of *Rickettsia* PCR testing of whole blood ranged from 6-69% when compared with serological diagnosis.[98]

PCR for leptospirosis was reported to be less sensitive than serological testing, 79% compared with 62%, over the course of illness. However, PCR was more sensitive early in disease.[99] These tests are not routinely used in most Nepali

healthcare settings.

1.7 Novel host biomarkers

Different research approaches have attempted to use the host immune response to assign aetiology of disease.[100, 101, 84] The principle being that different disease processes will trigger specific, consistent immune responses. These responses can then be used to identify the cause of the clinical presentation.

1.7.1 Transcriptomics

The transcriptome is the complete set of all RNA transcripts produced in a tissue or organism under specific physiological conditions.[102] Different gene expression profiling techniques allow us to measure the ribonucleic acid (RNA) transcripts produced. RNA-seq is a commonly used technique for biomarker discovery. The essential stages in RNA-seq are preparation of the sequencing library, sequencing of library results and analysis. These stages are discussed in more detail in Section 2.1.2.

RNA-seq allows us to measure the quantity of different RNA transcripts produced. This allows for DGE analysis. In DGE analysis we can compare the transcripts produced in different disease states, such as bacterial and viral infections.

Transcriptomic biomarker studies

Gene expression profiling can be used to identify transcriptomic signatures which differentiate between different infections; this approach relies on different diseases producing different levels of RNA transcripts. The transcript patterns seen in different diseases are called RNA signatures. Biomarker discovery using RNA platforms has shown promising results. DGE analysis using RNA-seq has identified RNA signatures for several different infectious processes. Herberg *et al.* demonstrated that a two transcript RNA signature can differentiate accurately between bacterial and viral infection in febrile children, sensitivity of 100%, and specificity of 96.4%.[103]

Missing bacterial sepsis in neonates is a major concern for paediatricians. Mahajan *et al.* reported a ten-transcript signature with 94% sensitivity and 95% specificity for identifying bacterial infections in infants less than 60 days of age.[100]

Tuberculosis in children is particularly difficult to diagnose, Anderson *et al.* identified a 51-transcript RNA signature which had a sensitivity of 82.9% and a specificity of 83.6% for diagnosing culture-positive tuberculosis.[104] Kawasaki disease is another condition which often is associated with diagnostic uncertainty. Wright *et al.* used a 13-transcript RNA signature to help distinguish Kawasaki Disease from other febrile illnesses with a sensitivity of 85.9% and specificity of 89.1%.[105]

Further work is required to reproduce these results in different populations and to evaluate the use of these RNA signatures in clinical practice.[106] Promising

biomarker discovery studies can also be performed using protein-based methods, as discussed in the following sections.

1.7.2 Protein platforms

Different platforms can be used to examine protein profiles in different biological states. Depending on the research question it may be useful to examine the entire protein profile or look at specific proteins. Both approaches are used in this thesis. In Chapter 3, mass spectrometry proteomics is used to look at several hundred proteins. In Chapter 4, a cytokine panel is used to analyse the levels of 10 selected proteins. In the following sections I discuss these two different approaches.

1.7.3 Mass spectrometry proteomics

The ability to measure many proteins in a sample can enhance understanding of the overall disease process and offers greater opportunities for biomarker discovery. The proteome can be defined as the set of proteins expressed in a cell, tissue or organism at a specific time and in a specific physiological state.[107]

MS proteomics is a commonly used research method to measure and identify hundreds of proteins in a given sample. Mass spectrometers measure mass-to-charge values and signal intensities to estimate which proteins are present in a sample and in what quantity.[107, 108]

MS proteomics has been used to measure protein abundance in different biological states, protein-protein interactions, protein function and biomarker discovery.[107,

109]

MS can be used with other techniques to improve performance. Liquid chromatography (LC) is combined with MS to increase the precision of protein measurements. LC helps to separate the different proteins as molecules of the same compound will tend to group together as they pass through the LC column. The MS proteomics workflow is discussed in more detail in Chapter 3.

1.7.4 Proteomic studies in infectious diseases

MS proteomics has been used to classify and aid understanding of disease with some relevant studies are highlighted here. In a large cohort of adult sepsis patients which tested over 2,000 samples, Mi *et al.* reported proteins associated with different sepsis phenotypes and severity of disease.[110] Stukalov *et al.* analysed the proteome of cells infected with SARS-CoV-2. Their results showed pathways involved in SARS-CoV-2 infection, and potential treatment targets. [111]

Using MS proteomics, Yin *et al.* identified proteins which were upregulated in a group of 40 children with RSV infection compared with healthy controls and suggested a link between glycolysis and RSV disease. [112] Palma Medina *et al.* used targeted proteomics to identify proteins which were differentially abundant between COVID-19 and pneumonia-associated sepsis.[113]

Jackson *et al.* reported a six-protein signature which could reliably differentiate between bacterial and viral infections in children. This was a multi-cohort study which included participants across 16 European sites. This study combined results

from different protein platforms and reported an area under the receiver operating characteristic (ROC) curve of between 89·4% and 93·6% when distinguishing bacterial and viral infections in children.[101]

1.7.5 Cytokines

While MS proteomic methods can detect a large number of proteins, these methods fail to detect smaller proteins, such as cytokines. Cytokines due to their low molecular weight have lower relative abundances and are often not detected using traditional MS techniques.[114] Other methods are required to detect these proteins. In the following sections, I discuss the importance of cytokines in the immune response, and common methods used to detect them. A panel of cytokines are measured in Chapter 4.

Cytokines are a heterogeneous group of small but important soluble proteins. Cytokines stimulate different responses within the body including inflammatory responses to pathogens, cell-to-cell communication, and control of cell replication. While having a key role in the immune response to infection, cytokines also have important roles in other areas including cancer control.[115]

Cytokines allow for a coordinated immune response to infection. Various cytokines allow cells in both the innate and adaptive immune responses to communicate. A variety of immune and non-immune cells produce cytokines; a complex network of interacting cells, cytokines and receptors is required to produce a coordinated immune response.[115, 116] Due to the ability of cytokines to activate a cascade of reactions, a small quantity of cytokine can make significant

biological differences.[117]

1.7.6 Role of different types of cytokines

Cytokines are divided into different groups depending on their physical properties, cells from which they are secreted or their primary functions. Cytokines can be divided based on whether they have predominantly pro- or anti-inflammatory properties, as discussed below.[116] Some of the major cytokine groups are described in the following paragraphs.

Interleukins (ILs) were initially grouped together as cytokines which are expressed by leukocytes, but they are also produced by many other cell types. Interleukins can influence reactions in many different cells and tissues by binding with high affinity cell surface receptors.[117]

There are 17 known families of interleukins, each with different functions. Some interleukins have predominantly pro-inflammatory properties including IL-1-beta, IL-6 and IL-8; whereas IL-4, IL-10, IL-13 and IL-1 receptor antagonist are predominantly anti-inflammatory.[118] Cytokines can have both pro- and anti-inflammatory effects. For example, IL-6 is raised in many inflammatory conditions, but is also involved in regulatory and anti-inflammatory processes.[119]

Colony-stimulating factors (CSFs) are cytokines which promote haemopoiesis. Granulocyte/macrophage colony-stimulating factor (GM-CSF), macrophage colony-stimulating factor (M-CSF) and granulocyte colony-stimulating factor (G-CSF) were initially grouped together for their role in generating mature myeloid cells from haematopoietic stem cells. Other roles for these cytokines

have been discovered including a role in activation of mature myeloid cells.[120] Cerebrospinal fluids (CSFs) can stimulate macrophages to produce various cytokines or enhance phagocytosis in neutrophils.[121]

Cytokines in the tumour necrosis factor (TNF) family have important roles in many inflammatory and autoimmune conditions. TNFs can aid in stimulating lymphocyte activation. Different TNFs can either increase lymphocyte survival or induce cell death. Outside of the immune system TNFs drive inflammation in a range of different tissues. Multiple TNF-blocking agents are available to treat autoinflammatory conditions including juvenile idiopathic arthritis and inflammatory bowel disease.[122]

Interferons (IFNs) are named for their ability to interfere with viral replication. Interferons are divided into three groups. Type I interferons, such as IFN-alpha and IFN-beta, are produced by innate immune cells in response to viral infections.

IFN-gamma is the only type II interferon and is primarily secreted by T cells and natural killer cells. IFN-gamma has a more modest anti-viral activity but has a prominent role in cell-mediated immunity.[123]

Type III interferons have prominent antiviral activity but are restricted by their distribution and they function mainly at endothelial and epithelial surfaces.[124]

Chemokines are small proteins even compared with other cytokines (8-12 kd). Chemokines provide the chemical stimulus to cause movement of cells, defined as chemotaxis. Chemokines can cause cell migration of several different types of cells including neutrophils, eosinophils, monocytes, and lymphocytes.[123]

Interactions between cytokines

It is difficult to interpret the action of any one cytokine in isolation. Cytokines are involved in a variety of functions within the immune system, and production of one cytokine can often alter the release of other cytokines. With these complex interactions in mind, it is useful to look at different cytokines in parallel.

In clinical practice, no one test should be interpreted in isolation. Medical practitioners use a variety of observations in forming their diagnosis. Cytokine interpretation should be thought of in the same way. A combination of cytokine results, or cytokine results combined with other investigations may provide more accurate diagnostic tests.

Ng *et al.* report that IL-6, TNF-alpha, and IL-1-beta were elevated in neonatal sepsis; combining these three cytokine results as one diagnostic test increased the accuracy of diagnosis (sensitivity, 95%; specificity, 84%) compared with using the individual cytokines.[125] Bergantini *et al.* report that a combination of IL-32, IL-6, IFN-gamma, and CRP had greater accuracy in differentiating severe COVID-19 disease compared with the individual biomarkers.[126]

In paediatric LRTIs, Zhu *et al.* found that combining CRP results with two other proteins (CD64 and CD35) increased the potential to differentiate between bacterial and viral LRTIs. However, adding IL-6 or procalcitonin did not improve the diagnostic accuracy.[127] ten Oever *et al.* reported results from another cohort of children with LRTIs; they report that combining CRP with proteins including IL-6 and IL-18 did not improve the ability to differentiate between bacterial and viral infections.[128] These results suggest that more work is needed to

understand the utility of cytokines in paediatric LRTIs.

Platforms for detecting cytokines

The reliable measurement of cytokines can be used to aid diagnosis and understanding of disease processes.

The variety of proteins detected using MS proteomic techniques is increasing, however, limitations still exist in detecting smaller proteins using this technology. Cytokines usually have a small molecular size and so contribute fewer peptides for MS detection.

The sensitivity of MS proteomics is usually insufficient to confidently detect the range of cytokines likely present. [114]

There are different methods available for detecting cytokines, as outlined in Section 4.1.3. Enzyme-linked immunosorbent assay (ELISA) is one of the most commonly used approaches for detecting analytes in biological samples. ELISA uses specific detection antibodies to measure analytes of interest.[129] ELISA methods measure light absorption to quantify analyte levels. In Chapter 4, I present results from a cytokine panel using the MSD platform. MSD uses similar methods to ELISA, except electrochemiluminescence is measured instead of light absorption.[130]

1.8 Multiomics - combining results from different platforms

Up to now we have discussed how individual RNA and protein platforms can be used to improve our understanding of disease processes. As messenger RNA is translated into proteins, the ability to analyse RNA and protein changes simultaneously should provide more information about the biological picture, compared with looking at RNA or proteins separately.

Integrating different -omic platforms is described as multiomics or integrative omics. Multiomics can involve the integration of results from many different platforms including genomics, epigenomics, transcriptomics, proteomics, and metabolomics.[131]

1.8.1 Advantages of multiomic approaches

The potential advantages of taking a multiomic approach include identifying important biological pathways, biomarkers or therapeutic targets which may be missed when analysing a single platform. Biological insights in one platform can corroborate findings in another. For example, epigenetic changes can be shown to influence the RNA transcripts produced, which in turn alters the proteins transcribed.[131, 132]

Multiomic studies can collect a large amount of data related to a particular biological process. This can be especially useful when trying to understand new diseases. During the COVID-19 pandemic, a United Kingdom (UK)-based

consortium developed a blood atlas of COVID-19, combining several different platforms. The authors were able to identify markers of disease severity across several platforms, as well as provide comparisons between COVID-19 infection and other diseases.[133] Stephenson *et al.* combined single-cell transcriptomics and proteomics with T and B cell antigen receptor analyses to define the immune response in early COVID-19 disease. These studies provide templates for the multiomic study of new diseases.[134]

Apart from COVID-19, Wozniak *et al.* combined proteomics and metabolomics to identify prognostic markers for *S. aureus* bacteraemia.[135] Multiomic approaches have been used to identify biomarkers for various non-infectious diseases including ovarian carcinoma, stroke, and gastrointestinal
cancers.[136, 137, 138]

1.8.2 Difficulties with multiomic approaches

Incorporating large datasets poses technical and statistical problems. Biases can be introduced when combining datasets with different numbers of features and different sensitivities. For example, proteomic datasets will consist of hundreds of proteins, with more abundant proteins more likely to be detected. This is compared with transcriptomic datasets consisting of thousands of transcripts, with detection of transcripts even at low abundance.[139] Integrating different omic platforms can increase the power to detect biomarkers of interest, however, this also increases the risk of false positive, spurious results. Missing data can cause significant issues when interpreting multiomic data. Many multiomic approaches require a complete dataset to accurately integrate different platforms.

Data can be missing for many reasons, and is more common in some platforms, e.g. proteomics, than others.[139, 140]

Fortunately, there are statistical methods which can be used to address these difficulties, but care must be taken in the analysis and interpretation of multiomic data. Methods to integrate different omics datasets are discussed in Section 5.1.2.

1.9 Biomarker studies and disease pathophysiology

While the primary aim of the biomarker studies described in previous sections was to discover new biomarkers, the results of these studies also improve our understanding of disease processes.

By using pathway analysis tools to analyse omic data a broader understanding of disease processes can be achieved. For example, the study of a multi-pathogen RNA signature from Habgood-Coote *et al.* used enrichment analyses to report the different biological pathways involved in bacterial infections, viral infections, malaria, and different inflammatory conditions.[141] Similarly, when presenting proteomic biomarkers which differentiate between COVID-19 disease and bacterial sepsis, Palma Medina *et al.* reported the differences found in the coagulation cascade in these conditions.[113]

In Chapters 2, 3 and 4, RNA and protein results in children with LRTIs are analysed. These results provide information on the immune response to LRTIs in

children. Improved understanding of disease processes can lead to discoveries in areas including vaccine development and novel therapeutics, as well as biomarker discovery.

1.10 Major events in Nepal during enrolment to the cohort studies

During the enrolment periods for both cohort studies presented in this thesis, some major events occurred in Nepal which have affected the results obtained. See Figure 1.3 for timeline of events relative to enrolment periods for the studies.

The first cohort study enrolled children with signs of pneumonia, and is referred to as the "LRTI cohort" for this thesis. Enrolment to the LRTI cohort began in March 2015 and ended in December 2017.

On 25th April 2015 and 12th May 2015 there were two major earthquakes in Nepal, followed by hundreds of aftershocks which caused devastation and destruction across Nepal.[142] The air quality in the Kathmandu Valley deteriorated due to dust from damaged buildings, and many people were forced to move.

During the LRTI cohort study, the PCV10 was introduced to the Nepali routine immunisation programme in a phased approach from January 2015. This vaccination programme began vaccinations in the Kathmandu Valley, during the second half of 2015.[143]

Between October 2015 and March 2016 there was a fuel shortage due to a blockade on the India-Nepal border.[144] This resulted in higher rates of solid

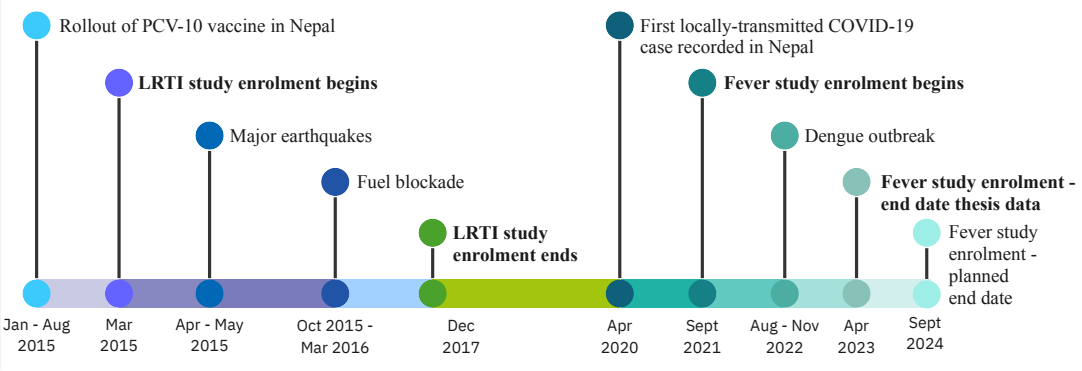


Figure 1.3: Timeline of major events which occurred in Nepal during the enrolment periods of both cohort studies included in this thesis. The first cohort study enrolled children with signs of pneumonia and is referred to as the LRTI cohort. The second cohort study enrolled children with any signs of infection and is referred to as the fever cohort.

fuel use and this impacted the air quality in the Kathmandu Valley.

The second cohort study recruited participants with all causes of infection, and is referred to as the "fever cohort" in this thesis.

Enrolment to the fever cohort began in September 2021 and the study is ongoing as of May 2024. When the fever cohort began recruitment, there were still high levels of COVID-19 cases being reported in Nepal.[145, 146]

There was also a large outbreak of dengue fever in the latter half of 2022 throughout Nepal, with high levels of cases in the Kathmandu Valley.[51]

1.11 Importance of the topic

Infections in children are common. While most children with infections do not require any medical intervention, a small but important minority will have negative outcomes without prompt and appropriate medical care. Identifying

these children early allows medical teams to prevent adverse outcomes.[147]

Unfortunately, using current diagnostic methods, serious infections continue to be identified later than we would like.[148] Improvements to diagnostics are needed to identify serious infections earlier. Improved understanding of infections is an important step in developing new diagnostic tests that can be used in clinical practice. The results presented in this thesis aim to add to our understanding of paediatric infection and highlight potential avenues to improve diagnosis of paediatric infectious diseases in children.

1.12 Hypotheses, aims and objectives

The hypotheses presented here are tested in two cohort studies set in Nepal. The first cohort consists of children admitted to hospital with LRTIs. Different platforms were used to measure RNA and protein levels in samples from these children.

The second cohort study recruited children with any sign of infection. The participants recruited were classified into different groups using routine testing and additional molecular diagnostic panels.

The major hypotheses tested in this thesis are:

1. Differences in whole blood RNA expression and plasma protein abundance can be used to differentiate between different causes of LRTIs in children.

2. In children with different causes of LRTIs, whole blood RNA expression and plasma protein abundance levels can be used to improve our understanding

of the pathophysiology of disease.

3. The addition of molecular diagnostics to standard-of-care testing provides useful diagnostic information, when describing the causes of fever in a Nepali context.

Aims and objectives

Linked to hypothesis one – Differences in whole blood RNA expression and plasma protein abundance can be used to differentiate between different causes of LRTIs in children. I aimed to measure the different gene expression and protein abundance levels across three different platforms, RNA-seq, MS proteomics and an MSD cytokine panel. I examined the differences in RNA and proteins levels across different LRTI classification groups and tried to identify RNA and protein signatures to differentiate between bacterial and viral LRTIs.

Linked to hypothesis two – In children with different causes of LRTIs, whole blood RNA expression and plasma protein abundance levels can be used to improve our understanding of the pathophysiology of disease. Using the differences in gene count and protein abundance levels between different LRTI classification groups, I performed pathway analysis with the aim of finding immune changes of interest in different LRTI pathogens of interest.

The addition of molecular diagnostics to standard-of-care testing provides useful diagnostic information, when describing the causes of fever in a Nepali context. I aimed to describe the causes of fever presenting to a hospital in the Kathmandu Valley. I aimed to investigate if additional molecular testing could have added diagnostic value in this cohort of children with LRTIs.

Each chapter has specific hypotheses, aims and objectives, linked to one or more of these main hypotheses.

Chapter 2

Analysis of the blood transcriptome of Nepali children with lower respiratory tract infections

2.1 Introduction

2.1.1 Chapter in context

In Chapter 1, I outlined the importance of LRTIs in children. I highlighted that in LMICs like Nepal, the outcomes for LRTIs are worse than in high-income countries.[88] I also outlined the difficulties with assigning an LRTI aetiology using currently available diagnostics and the importance of developing new diagnostic tests.[17, 60, 61]

In this chapter, I present analyses of blood transcriptomic results from Nepali children with LRTIs. The results are from a prospective cohort study of children who presented to hospital with signs of pneumonia in an urban Nepali setting. Research samples were obtained near the time of hospital admission and again at convalescence when the child was well.

I have used the data from the transcriptomic analysis presented in this chapter to differentiate between different LRTI aetiologies and to identify immune mechanisms in different LRTI aetiologies.

Different, novel, host immune biomarkers have been suggested to improve diagnostics in infectious diseases.[82, 83, 84] Platforms which can evaluate many potential biomarkers simultaneously, such as RNA-seq of the blood transcriptome, can speed up biomarker discovery. The following sections describe common approaches to measuring multiple RNA transcripts and evaluating their use as biomarkers for infectious diseases.

2.1.2 Transcriptomics

Transcriptomics involves studying the set of all RNA transcripts produced by an organism in a certain biological state, such as LRTI.[102] Different levels of gene or transcript expression are observed in different disease processes. Measuring these differences and analysing the results allows us to better understand disease processes. This allows for biomarker discovery and an improved understanding of the pathophysiology of disease.[149]

RNA-seq is a commonly used tool to measure the quantity of different RNA transcripts produced. Illumina short-read/next-generation sequencing techniques have provided most of the published data up to now, however, long-read and direct-RNA sequencing are likely to become more common.[150, 149] The results in this chapter were produced using a short-read approach and there are three general steps for this process.

First, prepare an RNA-seq library. This involves extracting RNA from the samples followed by fragmentation of RNA to allow for sequencing. Then RNA is converted to complementary deoxyribonucleic acid (DNA). DNA is more stable than RNA and it is easier to amplify DNA. Adaptors are added to the sequencing machine can identify the fragments, and short sequences called barcodes allow different samples to be tested simultaneously. A library of DNA fragments is then created by amplification using PCR.

Second, sequence the library results. The DNA fragments are inputted into the sequencing machine. The sequencer uses fluorescent probes to identify the nucleotides in each DNA fragment. Identification of all the nucleotides in a

fragment creates a sequencing read. The sequencing machine creates millions of reads which are then stored in a digital file. Low-quality reads are filtered out and the reads are aligned to a genome by matching the read fragments to genome fragments. The number of reads that align to each gene are quantified and the data is normalised based on the number of reads per sample.

The third step involves analysing the results. The data is plotted on quality control plots looking for outlying samples. differential expression analysis (DEA) is performed and pathway analyses are used to aid in understanding the results.

Differential expression analysis

DEA allows researchers to look for important differences in gene expression between different disease groups, e.g. between bacterial and viral infections. There are several methods for DEA, but the basic principles are the same for each method. The levels of gene expression are measured between the different disease groups, then a test statistic is used to identify genes which are statistically differentially expressed. Because of the large number of genes in RNA-seq results, the results of the statistical tests need to be corrected for multiple testing to account for potential false positives.[151]

edgeR and DESeq2 are among the most commonly used tools for DE analysis. Both of these tools incorporate models with a negative binomial distribution. A method adjusted from microarray methods called 'limma voom' has been suggested as an alternative approach to the negative binomial-based methods. With limma voom a linear model is fitted instead of a negative binomial distribution and Bayesian methods are used to estimate a standard error. The potential

advantages of limma voom include improved error control when sample sizes are small, and the increased range of statistical tools available for limma voom through the utilisation of microarray methods. In simulation tests, limma voom performed at least as well as edgeR or DESeq2 in terms of power and error rate control.[151, 152] DEA and false discovery rate adjustments are discussed further in the methods section.

After generating a list of differential expression (DE) genes, pathway analysis can help to understand the role of these genes in disease processes. More than 70 pathway analysis methods have been proposed, each with advantages and disadvantages. Generally, the list of DE genes from an experiment is compared with a database with information on known biological pathways. A common approach is known as over-representation analysis (ORA) which takes the list of DE genes and identifies pathways where these genes are over- or under-represented.[153]

Other approaches consider all of the genes in a dataset, not just the DE genes, such as functional class scoring methods. The hypothesis for these approaches is that small changes in functionally related genes can be important, even if the differences in individual genes are not statistically significant. gene-set enrichment analysis (GSEA) is a popular functional class scoring method.[153, 154]

2.1.3 RNA transcripts in infectious disease biomarker studies

Various studies have reported that differences in RNA transcript levels could be useful in the diagnosis of conditions which typically cause diagnostic difficulties, as discussed in Section 1.7.1. Avoiding missing serious bacterial infections is important clinically. Changes in RNA transcript levels have been used to differentiate between paediatric viral and bacterial infections with high sensitivity and specificity. In a prospective cohort study of febrile children presenting to hospital, a two-gene signature had a sensitivity of 100%, and specificity of 96.4% in differentiating between bacterial and viral infections.[103] Mahajan *et al.* proposed a ten-transcript signature with 94% sensitivity and 95% specificity for identifying bacterial infections in infants less than 60 days of age.[100]

RNA signatures have shown promising results in Kawasaki disease, tuberculosis and multisystem inflammatory syndrome in children (MIS-C).[104, 105, 155, 156]

Ideally, RNA-based signatures should improve on currently available diagnostic tests. Suarez *et al.* compared RNA transcripts in adults with bacterial and viral LRTIs to procalcitonin; the authors reported that a 10-gene signature differentiated between bacterial and viral LRTIs with a 95% (CI 77%–100%) sensitivity and 92% specificity (CI 77%–98%), compared with a sensitivity of 38% (18%–62%) and specificity of 91% (76%–98%) for procalcitonin.[157]

In this chapter, I look for RNA transcripts that can differentiate between different LRTI aetiologies.

2.1.4 Hypotheses

Two hypotheses related to the analysis of the blood transcriptome are tested in this chapter:

1. Blood RNA results can be used to differentiate between different LRTI aetiologies.
2. Changes in blood RNA transcripts can be used to describe immune mechanisms in different LRTI groups.

Aims and objectives

Linked to hypothesis one – Blood RNA results can be used to differentiate between different LRTI aetiologies. I aim to show that certain transcripts will be significantly different depending on the aetiology of LRTI. To address this aim the following comparisons will be made across the different LRTI groupings in the study:

1. Samples from the acute bacterial LRTI group compared with convalescent samples from the same participants.
2. Samples from the acute viral LRTI group compared with convalescent samples from the same participants.
3. Directly compare viral and bacterial LRTI groups.
4. Using DEA, create and test models to differentiate between bacterial and viral LRTIs

5. LRTIs with co-infections compared with mono-infections

By first comparing acute with convalescent samples from the same individuals, bias due to age, sex or other unknown factors can be accounted for as the convalescent samples act as age- and sex-matched healthy controls.

Linked to hypothesis two – Changes in blood RNA transcripts can be used to describe immune mechanisms in different LRTI groups. To address this hypothesis the following comparisons will be made:

1. LRTIs associated with particular pathogens of interest compared with other acute LRTIs.

   (a) Pathogens of interest include RSVs, influenza viruses and *S. pneumoniae*.

2. Pathway analysis in different LRTIs aetiologies.

2.2 Methods

2.2.1 Clinical study

As this is the first chapter related to this LRTI cohort, I have provided a detailed description of the clinical methods related to the study delivery, and refer to the following methods sections in other chapters related to this cohort study.

Study design

A prospective cohort study was undertaken at Patan Hospital, Lalitpur, Nepal. In this hospital setting, children were eligible for enrolment if they presented to hospital and were assigned an admission diagnosis of pneumonia, as decided by the clinical team. Blood samples for RNA and protein analyses, as well as nasopharyngeal samples for molecular diagnostics, were taken on presentation to hospital. Demographics and medical record data were recorded, to assist in correctly classifying the different LRTIs for further analysis. Participants were asked to return six to eight weeks after admission to provide a convalescent blood sample when they had recovered, see Figure 2.1.

Ethical approval

The study was approved by the Oxford Tropical Research Ethics Committee (OxTREC) (reference, 05-14) and the Nepal Health Research Council (NHRC) (registration number, 04/2014). The International Conference on Harmonisation-Good Clinical Practice (ICH-GCP) guidance was followed throughout the study. Participants were free to withdraw from the study at any point. Informed consent was obtained by trained research staff.

Setting

Patan Hospital, Lalitpur, Nepal is a 640-bed teaching hospital. Lalitpur is a largely urban area in the Kathmandu Valley. Patan Hospital serves a large population in the Kathmandu Valley and is also a referral centre for complex

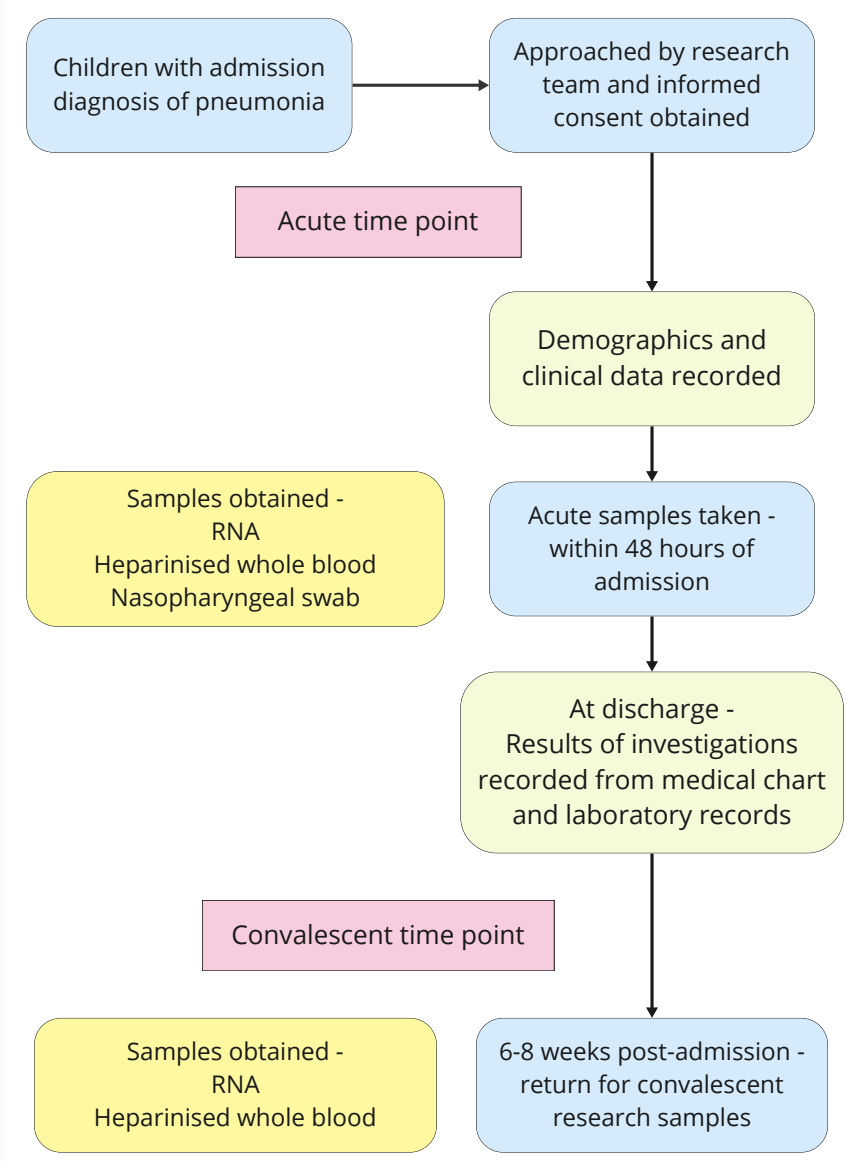


Figure 2.1: Study design with steps from presentation of potential participant to hospital to enrolment in the cohort study. Data and research samples recorded at acute and convalescent time points also shown.

cases from other parts of Nepal.

Enrolment period and population

Children were enrolled between March 2015 and December 2017. The enrolment period for the study allowed for recruitment of participants before and after the introduction of the PCV10 vaccine into the routine immunisation programme in Nepal. Recruitment of children with LRTIs was part of a vaccine impact study being undertaken in Patan Hospital at the time.[143]

Children between 2 months (over 60 days) and 14 years of age, who were admitted to Patan Hospital, and who had a diagnosis of pneumonia on admission were eligible for enrolment in the study. The diagnosis of pneumonia at admission was made by the clinical team. Usually, this diagnosis was made before the results of blood, or radiological investigations were available.

Participants were excluded from the study if they had a significant medical co-morbidity which, in the researcher's opinion, would put the participant at risk by taking part in the study.

Study procedures at each time point

Acute time point - Enrolment and during admission If the clinician assessing the child decided that the admission diagnosis was pneumonia, then the participant was approached by the research team. Informed consent was obtained from the parent or guardian before enrolling the participant in the study. Research samples were taken as early in the admission as possible, and within

a maximum of 48 hours after hospital admission. Demographics were recorded. Research staff obtained permission to access the participant's medical records during their hospital stay. At discharge or shortly after admission, medical notes were reviewed to record the results of investigations while in hospital, and the outcome, for each participant.

Convalescent time point Six to eight weeks after enrolment in the study, the parents/guardians were asked to bring their child back to Patan Hospital. At this follow-up visit consent was reconfirmed verbally. The participant's clinical course since enrolment was documented. A second set of research blood samples were taken to act as comparator samples. These convalescent samples can act as age- and sex-matched controls for the acute participant samples.

Healthy controls Well children, taking part in a separate vaccine study in the hospital were recruited to act as healthy controls. These children were already having samples taken for research purposes. After informed consent was obtained nasopharyngeal and blood samples were taken on one occasion only from these participants. These samples were used as a comparator group to the acute cases.

Research samples

Acute time point A nasopharyngeal sample was obtained as soon after enrolment as was feasible. The swab was immediately placed into skim milk, tryptone, glucose and glycerol (STGG) media and transported to the microbiology laboratory at Patan Hospital. Blood samples were obtained, if possible, when

clinically indicated blood draws were being undertaken. Approximately, 3 ml of blood was taken from each participant and divided as follows:

1. RNA-stabilising tube, 1 ml - Tempus Blood RNA Tube, ThermoFisher, MA, USA

2. Heparinised centrifuge tube, 2 ml

Blood was transferred to the laboratory at Patan Hospital within two hours.

Convalescent time point. The same set of blood samples was taken from each participant when they presented for their convalescent visit. A nasopharyngeal sample was not obtained.

2.2.2 Laboratory methods

Nasopharyngeal samples

In the microbiology laboratory at Patan Hospital, the nasopharyngeal samples were cultured for pneumococci. A 200 µl aliquot of the STGG media was transferred to a Columbia agar plate containing 5% sheep blood and incubated at 37°C overnight. The swab and remaining media were stored at -80°C. If a colony of *S. pneumoniae* was identified from the agar plate a single colony was taken for pneumococcal serotyping, using the Quellung method.[141] All colonies were typed by experienced research laboratory staff at Patan Hospital.

An aliquot of STGG media from each nasopharyngeal sample was transported to Micropathology Ltd., University of Warwick Science Park, UK. Nucleic acid was

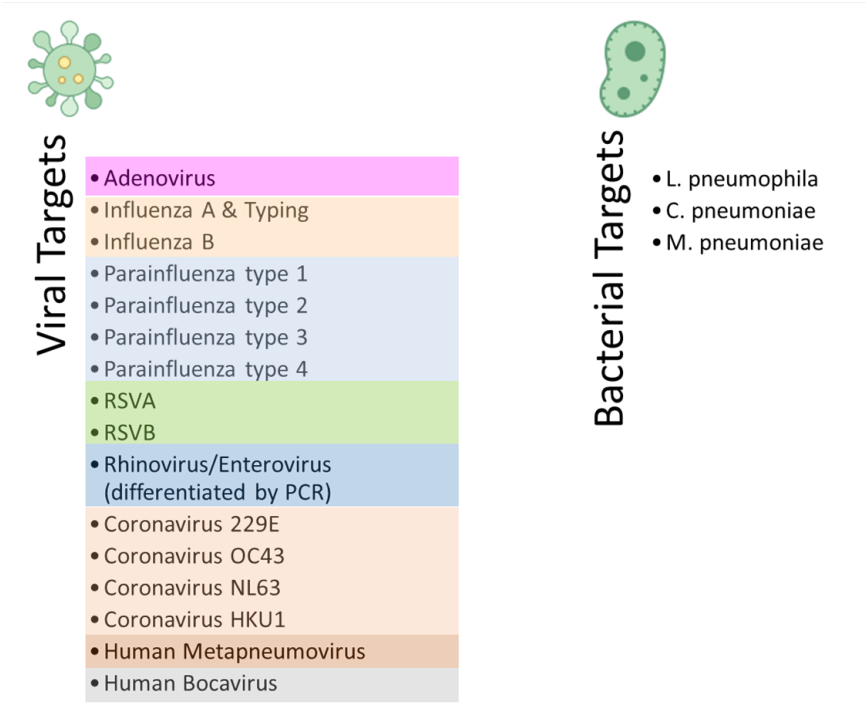


Figure 2.2: Nasopharyngeal swabs obtained from participants in the LRTI cohort study at admission to hospital underwent molecular testing using the NxTAG™ Respiratory Pathogen Panel (RPP) assay (Luminex® Corporation), pathogen targets listed here. These tests were carried out at the laboratory of Micropathology Ltd., University of Warwick Science Park, United Kingdom. RSV, respiratory syncytial virus; PCR, polymerase chain reaction; *L. pneumophilia*, *Legionella pneumophila*; *C. pneumoniae*, *Chlamydia pneumoniae*; *M. pneumoniae*, *Mycoplasma pneumoniae*. Image used with permission from Micropathology Ltd.

extracted from 200 µl of STGG using the QIAGEN DNeasy 96 kit (QIAGEN, UK). Forty microlitres of extracted nucleic acid were analysed by the NxTAG™ Respiratory Pathogen Panel (RPP) assay (Luminex® Corporation) according to the manufacturer's instructions. This panel allows for the detection of nucleic acid from 16 viral and 3 bacterial targets, see Figure 2.2.

Blood samples

RNA-stabilising tubes Samples added to the Tempus Blood RNA Tubes were frozen at -80°C, within four hours of sampling, at Patan Hospital.

These samples were then transported to the laboratory of the Wellcome Trust Centre for Human Genetics (WTCHG), University of Oxford, where nucleic acid was extracted. Total RNA was extracted from these samples using the Tempus Spin Isolation Kit. See Section 2.2.4 for how samples were selected for the RNA-seq experiment.

Libraries were prepared using the Illumina TruSeq Stranded Total RNA Library Prep Kit. Following depletion of ribosomal and globin RNA, sequencing was undertaken using IlluminaHiSeq4000 sequencing to generate 75 base pair (bp) paired end reads. All samples were sequenced to a depth of at least 28 million reads. Samples from this study were sequenced in two batches at the same lab, the first in 2017 and a second sequencing experiment in 2019.

Heparinised blood tube. The fresh whole blood in the heparinised centrifuge tube was processed within four hours of sampling. The blood was centrifuged and separated into plasma, peripheral blood mononuclear cells (PBMCs), and a mixture of red cells and polymorphonuclear cells. The plasma component was used for the protein analyses in Chapters 3 and 4, and the processing of the plasma samples is discussed further in those chapters. The other components were used for experiments which are outside of the scope of this thesis and will not be discussed further here.

2.2.3 Classification of cases

Participants were classified into different LRTI groupings based on the likely cause of their illness. The results of routine investigations and the NxTAG respiratory PCR panel were used for classification, see Figure 2.3. It is important to assign participants to the correct groups so that meaningful analyses can be carried out when comparing the different causes of LRTI.

Participants were classified as definite bacterial if they had a culture-positive pathogenic bacteria isolated from a normally sterile site (blood or pleural fluid). The probable bacterial group included participants with high inflammatory markers (CRP >60mg/L) in the absence of a culture-positive bacteria and the absence of an LRTI-associated virus. LRTI-associated viruses were RSV A and B, influenza A and B, human metapneumovirus and parainfluenza 1, 3 and 4. These viruses have been associated with paediatric LRTIs in previous case-control studies. [17, 54]

Participants were classified as definite viral if they had an LRTI-associated virus isolated in their nasopharynx, with low inflammatory markers (CRP <60mg/L and neutrophils $\leq$ 12 x $10^9/L$, and with no pathogenic bacteria isolated from a sterile site. The 'viral syndrome-high or no inflammatory markers' group consisted of participants who had an LRTI-associated virus detected in their nasopharynx and also had a CRP >60mg/L or neutrophils >12 x $10^9/L$ or no result for either inflammatory marker available. Bacterial-viral co-infections were classified as those who met the criteria for the definite bacterial group, and who also had an LRTI-associated virus isolated from their nasopharynx. Participants who failed to meet the criteria for any other grouping were classified as unknown.

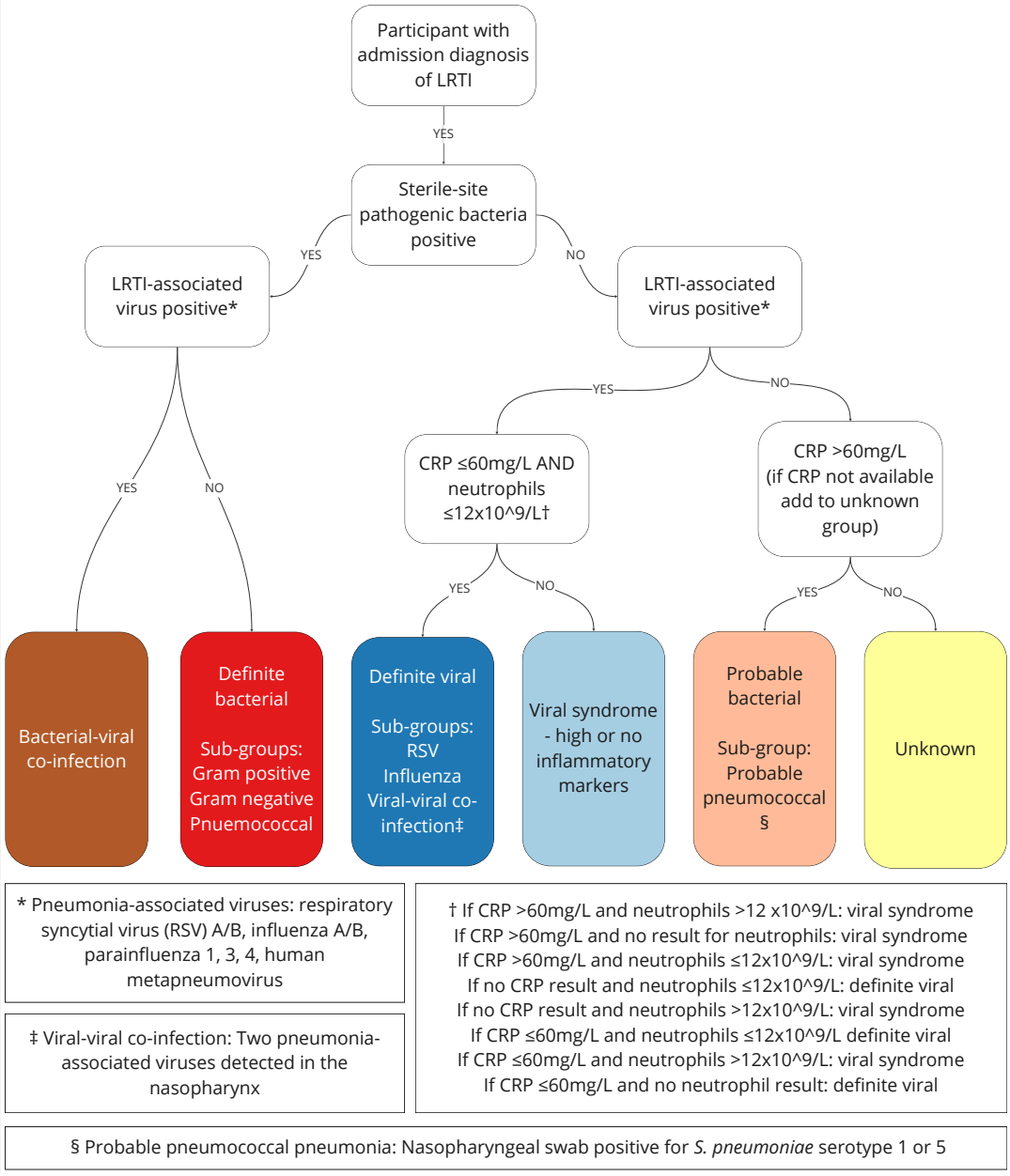


Figure 2.3: System used to classify the participants in the LRTI cohort into different lower respiratory tract infection groups. Classification is based on routine testing at Patan Hospital and molecular testing of nasopharyngeal samples.

Secondary groupings were used to look at individual pathogens and severity. The RSV group were a sub-group of the definite viral group who had RSV isolated from their nasopharynx. The influenza group were a sub-group of the definite viral group who had an influenza virus isolated from their nasopharynx. The pneumococcal group included any participants in the definite bacterial group who had *S. pneumoniae* cultured from a normally sterile site, or participants in the probable bacterial group who had pneumococcal serotypes 1 or 5 cultured from their nasopharynx. Serotypes 1 and 5 are the most prevalent pneumococcal serotypes in children with invasive bacterial disease and are highly associated with case status in case-control studies in Nepal.[56, 57]

2.2.4 RNA-sequencing

A selection of RNA samples were sent for RNA-seq as the resources available for the study did not allow for all samples to be tested. All cases in the definite bacterial and viral groups were included for RNA-seq as these were the cases with the most confident diagnosis. A selection of probable bacterial, viral syndrome and unknown cases were included to give a good representation of the different causes of LRTIs in the cohort. Where a convalescent sample was available for any of the selected cases this sample was also sent for RNA-seq.

Initial processing

The High-Throughput Genomics Group at the WTCHG generated the RNA sequencing data. Raw reads were aligned to the human genome (build GRCh38,

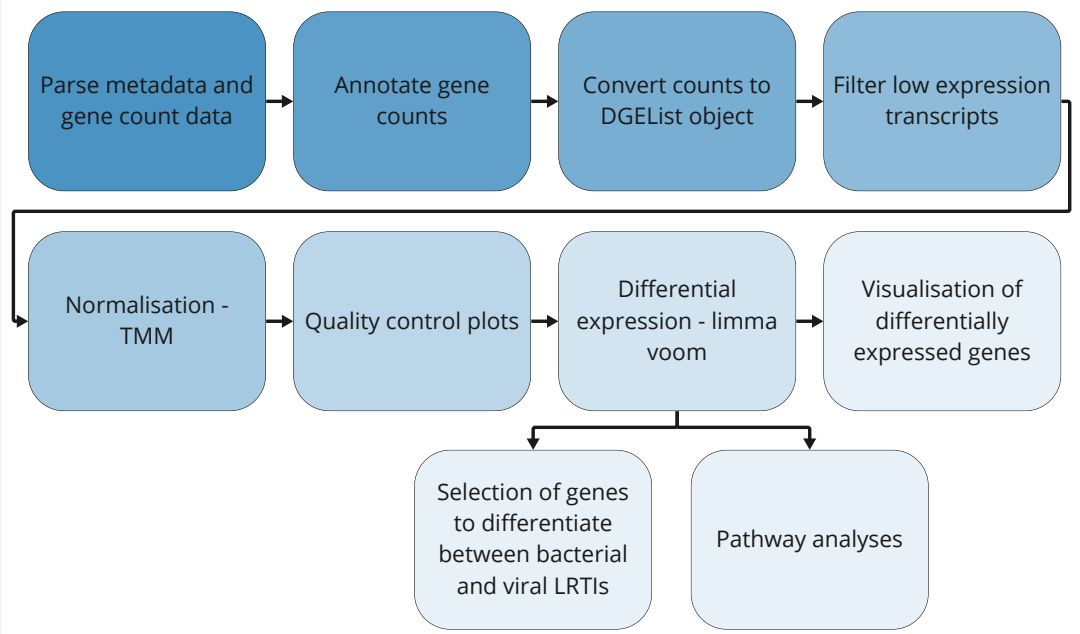


Figure 2.4: Overview of processing steps undertaken when analysing the RNA-seq data from the LRTI cohort study, using R programming language.

available from http://www.gencodegenes.org/releases/21.html) using the Salmon alignment tool.[158] After alignment, duplicate reads were identified and removed using samtools.[159] The HTSeq tool was then used to generate counts of the number of filtered reads.[160] This initial processing work was carried out as part of a previous thesis.[161] Following alignment and filtering of duplicate reads, these data were transferred to me. I continued the analysis using R, as described below.

Processing of read count data

A detailed description of these methods can be found in Appendix B, including R packages used. See Figure 2.4 for an overview of the processing steps described below.

Annotation and filtering. The count data from the two RNA-seq batches were merged. The count data was annotated with Entrez IDs using the org.Hs.eg.db Bioconductor package, and the National Center for Biotechnology Information (NCBI)-Gene database. The clinical metadata was uploaded and parsed for later analyses.[162, 163]

The data were converted to a DGElist object which allows the count data, annotation labels and metadata to be stored together. Genes with low counts across samples were filtered out using the filterByExpr function in the edgeR package which takes into account the average library size and the smallest clinical grouping of samples.[164]

Normalisation and quality control. The weighted trimmed mean of the log-expression ratios, or M values, were calculated and used as normalisation factors for each sample. These trimmed mean of M values (TMM) normalisation factors were included in the model for differential expression.

Bar plots of library sizes, box plots of counts per million and PC biplots were created to look for unusual patterns in the data.

The count data were batch-corrected to account for the two different RNA-seq experiments using the removeBatchEffect function.[150] The PC biplots were repeated post-batch correction to look for outlier samples and assess potential confounding factors.

Differential gene expression analysis

A design matrix was created which included the LRTI classifications and confounding factors.

Before performing differential expression analysis, the data were transformed using the voom function which transforms the data into log2-counts per million (log2CPM) while taking into account the mean-variance relationship.[165] To take into account that a proportion of participants have acute and convalescent samples, the duplicateCorrelation function was used to account for correlation between samples from the same participant.[150] Voom transformation was performed for a second time taking into account the correlation between acute and convalescent samples. A linear model was fit to the voom transformed data to test for differential gene expression using the limma package to create a fit object. A contrast matrix was applied to the fit object to get results for the comparisons of interest. See Table 2.1 for the different comparisons tested. Empirical Bayes shrinkage was performed and adjusted p-values calculated using the BH Procedure.[152, 166]

Volcano plots were created to visualise the DE genes using adjusted p-value thresholds of 0.05, and log2-fold changes of 1.

Table 2.1: Lower respiratory tract infection (LRTI) comparisons of interest to include in contrast matrix, to address chapter hypotheses.

Hypothesis	**LRTI group of interest**	**Contrast**
Plasma RNA results can be used to differentiate between bacterial and viral LRTIs	Acute bacterial	Convalescent bacterial
	Acute viral	Convalescent viral
	Acute bacterial	Acute viral
	Acute bacterial-viral co-infection – pathogenic bacterial and pneumonia-associated virus identified	Convalescent bacterial
	Acute bacterial-viral co-infection – pathogenic bacterial and pneumonia-associated virus identified	Convalescent viral
Plasma RNA results can be used to differentiate between LRTIs due to different pathogens	Acute RSV	Convalescent RSV
	Acute influenza	Convalescent influenza
	Acute pneumococcal	Convalescent pneumococcal
	Acute RSV	Acute other viral
	Acute influenza	Acute other viral
	Acute pneumococcal	Other bacterial

2.2.5 Signatures to differentiate between bacterial and viral LRTIs

Identifying samples to include in bacterial-viral comparison

A subset of samples was selected to train and test a model to differentiate between bacterial and viral infections. As discussed, gold standard diagnostics cannot always reliably classify cases. There were a relatively small number of definite bacterial cases, and I wanted to be more confident that the participants in the probable bacterial group were bacterial cases. I used a semi-supervised approach to increase confidence in the classification of cases into their correct groupings.

The unsupervised machine learning approach, K-means clustering, was used to identify unlabelled clusters in the data. To find the optimal number of clusters to use, the within-cluster sum of squares was measured for different numbers of clusters. Using the elbow method, the optimal number of clusters is the point where an acute angle is seen when the within-cluster sum of squares for each possible number of clusters is plotted.

The previously assigned LRTI classification labels were then added to identify clusters which correlated with the LRTI classification labels. Bacterial and viral clusters were identified. See Figure 2.5 for a graphical explanation of this semi-supervised approach, with a hypothetical ideal outcome.[167]

The samples with a probable bacterial and definite bacterial label that clustered together were selected as the bacterial group for model training and testing. The samples which had a probable bacterial label but clustered with the viral samples were excluded. Similarly, samples with a definite viral label that clustered with

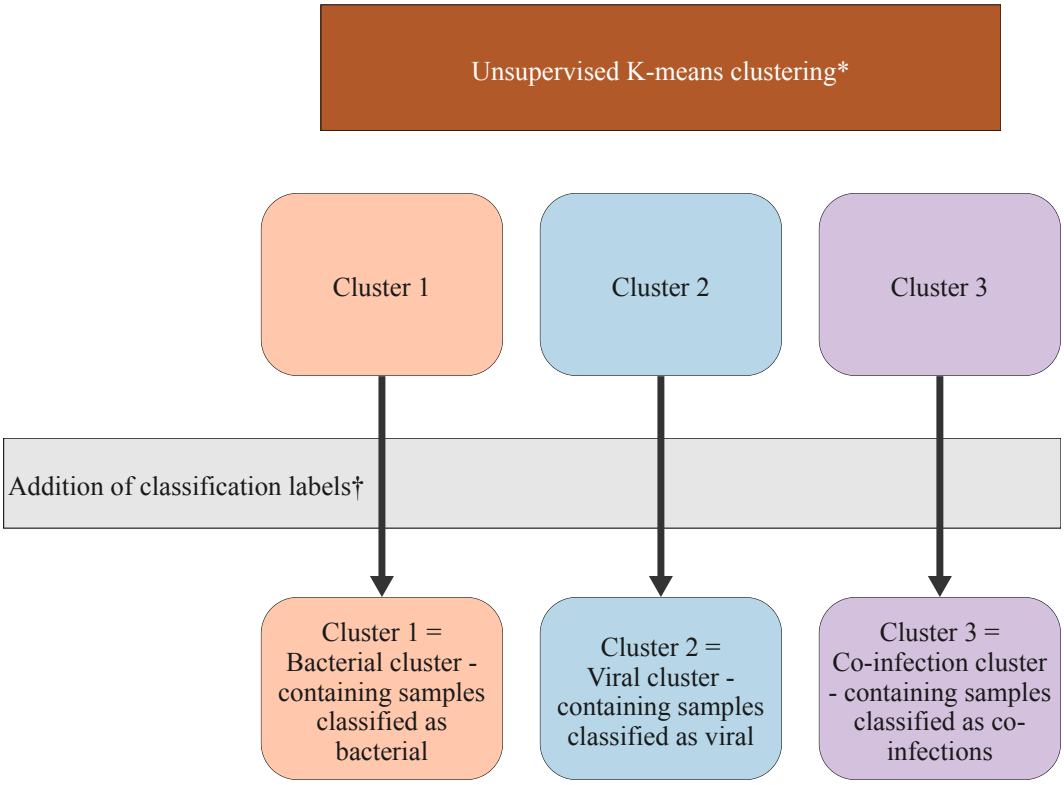


Figure 2.5: Graphical overview of the semi-supervised clustering approach using K-means clustering. *Using an example where three is the optimal number of clusters. †Classification groups based on the classification system in Figure 2.3 involving routine investigations and nasopharyngeal swab results.

the bacterial samples were excluded.

As the co-infection group also met the criteria for definite bacterial infection, culture-positive for a pathogen from a normally sterile site, these samples were included with the bacterial group for comparisons between bacterial and viral cases. For the creation of an RNA signature, it would be of more use clinically if the co-infections were identified as bacterial, as these cases will need treatment with antibiotics.

Identifying genes to include in bacterial-viral model testing

A 70:30 split was used to divide the data randomly into training and test datasets, keeping the ratio of bacterial to viral cases in each group similar. In the following DGE analyses, comparisons with a log2-fold-change >1, and an adjusted p-value <0.05 were considered significant.

Training dataset In the training dataset, DGE analysis was performed. The contrast matrix included comparisons between acute and convalescent samples from the participants in the bacterial LRTI group, and acute and convalescent samples from the participants in the viral LRTI group. Comparing acute and convalescent samples from the same participants should help to account for any age and sex biases in the data.

The list of DE genes when the acute bacterial group were compared with the convalescent group, was combined with the list of DE genes when the acute viral group was compared with the convalescent group. These genes were taken forward to the next phase of analysis.

This subset of genes were then used to perform DGE analysis again, but this time directly comparing the acute bacterial samples with the acute viral samples. The list of DE genes between bacterial and viral samples was then taken forward for feature selection.

Using the caret package in R, the partial least squares approach was used to train a model comparing bacterial and viral groups, with 10-fold cross-validation repeated three times. The differentially expressed genes (DEGs) were ranked in order of importance. The least important genes were excluded. The importance measure used was the weighted sums of the absolute regression coefficients.

A correlation matrix of the most important DEGs was created and highly correlated genes were excluded. The model was re-trained after excluding highly correlated genes. The root mean squared error (RMSE) was calculated for models containing different numbers of genes. The optimal number of genes was selected for use in the bacterial-viral signature. The accuracy of this signature was tested in the training dataset. Different feature selection methods were trialled to create different gene signatures, see Appendix B.2.3 for full details of model creation.

Performance statistics were calculated for the different gene signatures. Sensitivity, specificity, ROC - area under curve (AUC), positive and negative likelihood ratios were calculated for the training set initially. As a test which has a high sensitivity to detect bacterial infections would be more useful clinically, I restricted the sensitivity range to 90% to 100% and re-calculated the AUC. This partial area under the ROC curve (pAUC) showed how well the model was able to predict a correct result when the required true positive rate was high.[168] The top-performing signatures were taken forward for performance analysis in

the test dataset.

Test dataset The normalised gene counts in the test dataset were used to measure the accuracy of the models created using the training data.

2.2.6 Pathway analysis

DGE analysis was repeated with the groups changed to reflect pathogen classifications. The three main pathogen groups are the RSV, influenza and pneumococcal groups. I wanted to look for a set of genes that would differentiate these pathogens from other acute LRTIs.

First, the acute samples from the different pathogen groups were compared with convalescent samples from the same participants. Then acute pathogen groups were compared with other acute samples. For example, the acute pneumococcal group was compared with the convalescent pneumococcal group. Then taking the subset of differentially expressed genes in the pneumococcal acute versus convalescent comparison, the pneumococcal group was compared with a group of samples with other bacterial diagnoses. The aim was to identify a group of genes which were important in pneumococcal infections. Similarly, acute RSV and influenza groups were compared with convalescent samples and then with samples from participants with other viral infections.

GSEA was used to look for pathways in curated databases which were enriched in these pathogen comparisons. GSEA incorporates the full list of genes tested, ranked by their log-2-fold change. Over-enriched gene sets were searched for in two groups of gene sets, C5: ontology gene sets:Gene Ontology (GO)-Biological

Process (BP) and C7: immunologic signature gene sets, in the Molecular Signature Database (MSigDB).[154]

The gene sets were ranked by their normalised enrichment score (NES). Redundant gene sets, gene sets with very similar genes described, were removed from the final list. In GSEA analysis, genes are identified which contribute most to the detected enrichment signal of a pathway, these are called 'leading edge' genes.[169] To compare multiple genes from the same pathway on the same plots, the DGE levels were scaled and centred to produce z-scores. Z-scores were calculated as the normalised gene count values minus the mean value for each gene, divided by the standard deviation.

2.3 Results

2.3.1 Description of cohort

Sample numbers

RNA-seq was carried out on blood samples in two different batches in the same laboratory, 192 samples were processed in 2017 and 213 samples in 2019. The results from the combined total of 405 samples are presented here. The 405 samples are made up of 258 samples from children with acute LRTIs and 112 samples from the same children 6-8 weeks later during convalescence. Samples from 35 healthy controls were also included. The healthy controls were included to help adjust for batch effects, and won't be included in further analyses.

Table 2.2: In the RNA-seq results, the number of genes included for analysis after annotation and filtering steps. The number of genes excluded at each step and the reason for exclusion is shown.

	Gene counts	
	Exclude	**Include**
Raw count data		56,852
Annotation using Entrez IDs and gene symbols	19,491	37,464
Exclude duplicate genes*	69	37,385
Filtering (filterByExpr)	15167	22,149
Gene ID discontinued†	17	22,132
Number of genes used for analyses		22,132

* Where two rows with the same gene Entrez ID were present, the row with the lower summed total counts was excluded.
† As per the National Center for Biotechnology Information database, accessed on 10/Apr/2024

Processing and quality control

The raw count data included 56,852 gene rows. Following annotation and filtering, 22,132 genes were included for analysis, see Table 2.2.

Following TMM normalisation, quality control plots were created. PC biplots identified two distinct groups in the results, Figure 2.6 shows that these groups correspond to the two different RNA-seq experiments. A batch correction step was carried out to account for this. Following batch correction, PC biplots were repeated and three outlier samples were identified, two healthy control samples and one convalescent sample. Box plots of log2-counts per million did not identify any additional outlier samples. See Appendix B for further QC plots. These three samples were excluded which meant that 402 samples were included for

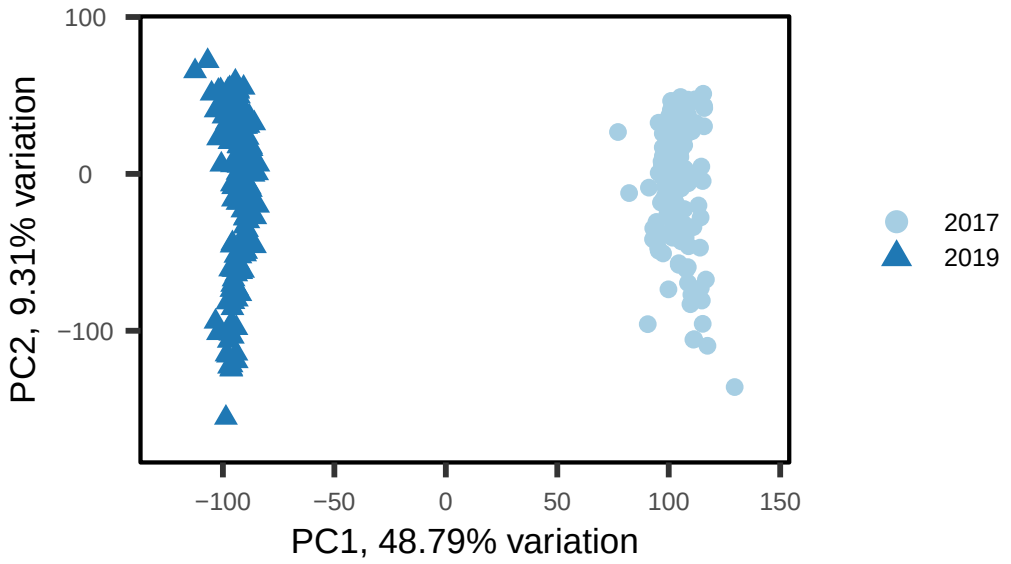


Figure 2.6: Results from PCA of results of gene count data from participants in the LRTI cohort study, prior to batch correction. Principal component (PC1) biplot showing PC1 and PC2, coloured by year of RNA-seq experiment, prior to batch correction.

further analyses.

Demographics and clinical data post-quality control

Table 2.3 shows the 402 samples divided into their LRTI classifications with demographics and investigations at admission also shown. The average age of the cases was 3.1 years, with 38% female. Convalescent samples results were available for 43% of the cases. The duration of hospital stay was longest in the bacterial-viral co-infection group, however, numbers were small in this group, seven cases. The inflammatory markers at admission were higher in the definite

Table 2.3: Cases divided into their lower respiratory tract infection classification and healthy controls. Basic demographics and investigations at time of admission to hospital also presented. CRP, C-reactive protein; WCC, blood white cell count.

	Definite viral	Viral syndrome - high or no inflammatory markers	Unknown	Probable bacterial	Definite bacterial	Co-infection*	Total cases	Healthy controls
Number of cases	77	20	61	77	16	7	258	33
Convalescent sample available	59	6	17	19	8	2	111	NA
Age, years (median)	1.5	1.9	2.4	5	5.9	1.4	3.1	0.8
Female (%)	38	45	34	44	25	29	38	64
Duration of hospital stay, days (median)	6	7	4	7	6	8.5	6	NA
CRP (mg/L)	7	97	10	152	151	163	40	NA
WCC x10^9/L	10.7	14.3	13.6	19.5	19.5	9.4	14.2	NA
Neutrophils x 10^9/L	5.2	11.4	7.7	15.2	14.5	3.8	8.5	NA

bacterial groups, compared with the definite viral group, as expected.

The definite bacterial group consisted of ten cases of *S. pneumoniae*, three *S. aureus* cases, two *S.* Typhi and one *Pseudomonas species*.

2.3.2 Principal component analysis

Following batch correction, PC biplots were repeated. The biplot in Figure 2.7 is coloured by study time point with convalescent samples and healthy controls

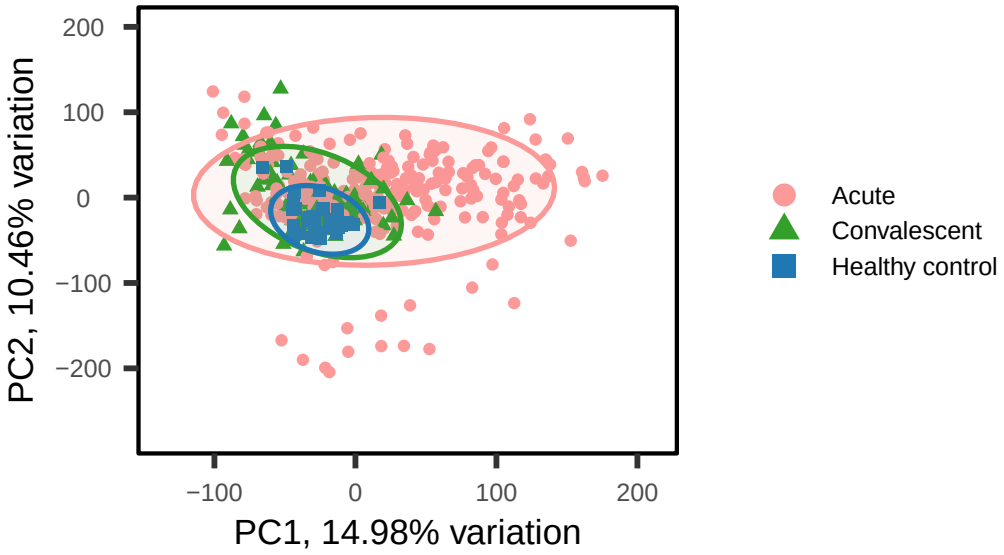


Figure 2.7: Results from PCA of results of gene count data from participants in the LRTI cohort study, post-correcting for the batch effect due to different RNA-seq experiments. Principal component biplot showing PC1 and PC2, coloured by study time point

overlapping. Acute samples also overlap with convalescent and healthy control samples but there is some separation of acute and convalescent samples along PC1.

After excluding convalescent and healthy control samples, PC biplots with just acute samples were examined. The PC biplot in Figure 2.8 is coloured by primary LRTI classification. There is some separation along PC1 between viral samples to the left and bacterial samples to the right of the plot. This is clearer in Figure 2.9 where a more simple classification is used which highlights the bacterial (probable and definite combined) and definite viral samples.

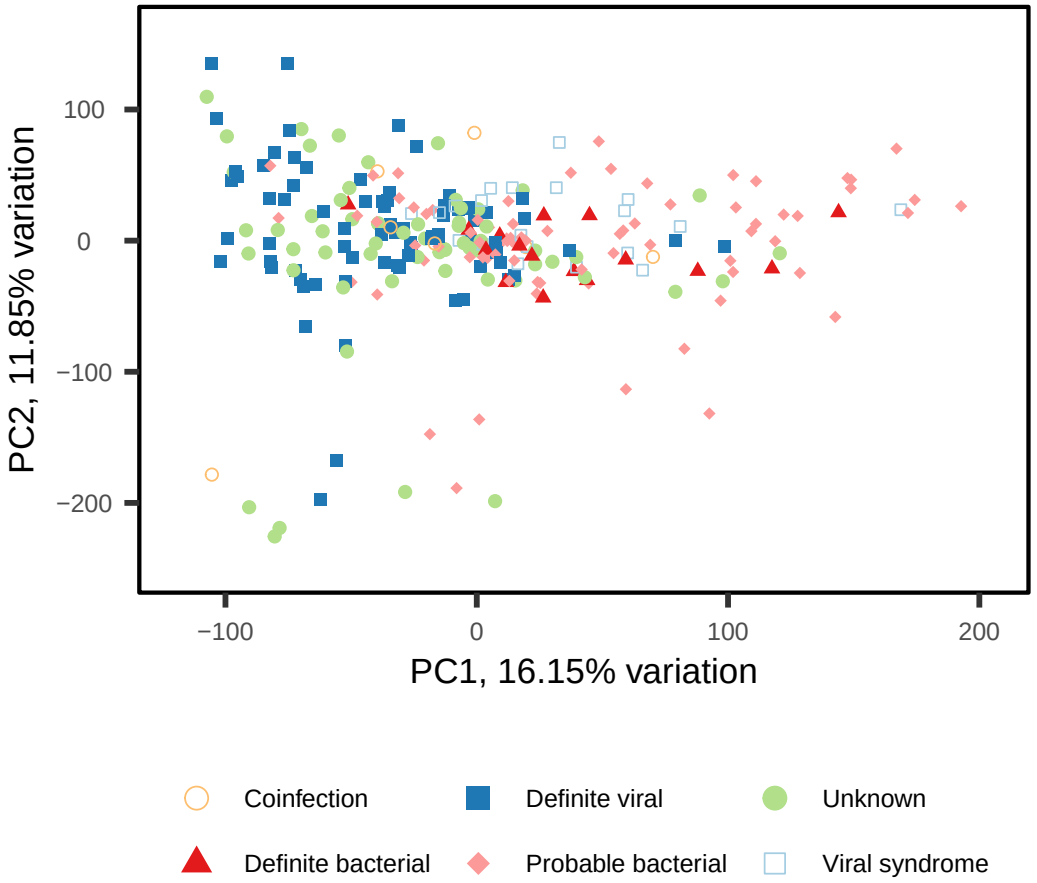


Figure 2.8: Results from PCA of results of gene count data from participants in the LRTI cohort study, post-correcting for the batch effect due to different RNA-seq experiments. Principal component biplot showing PC1 and PC2, coloured by primary lower respiratory tract infection classification, post-correcting for the batch effect due to different RNA-seq experiments.

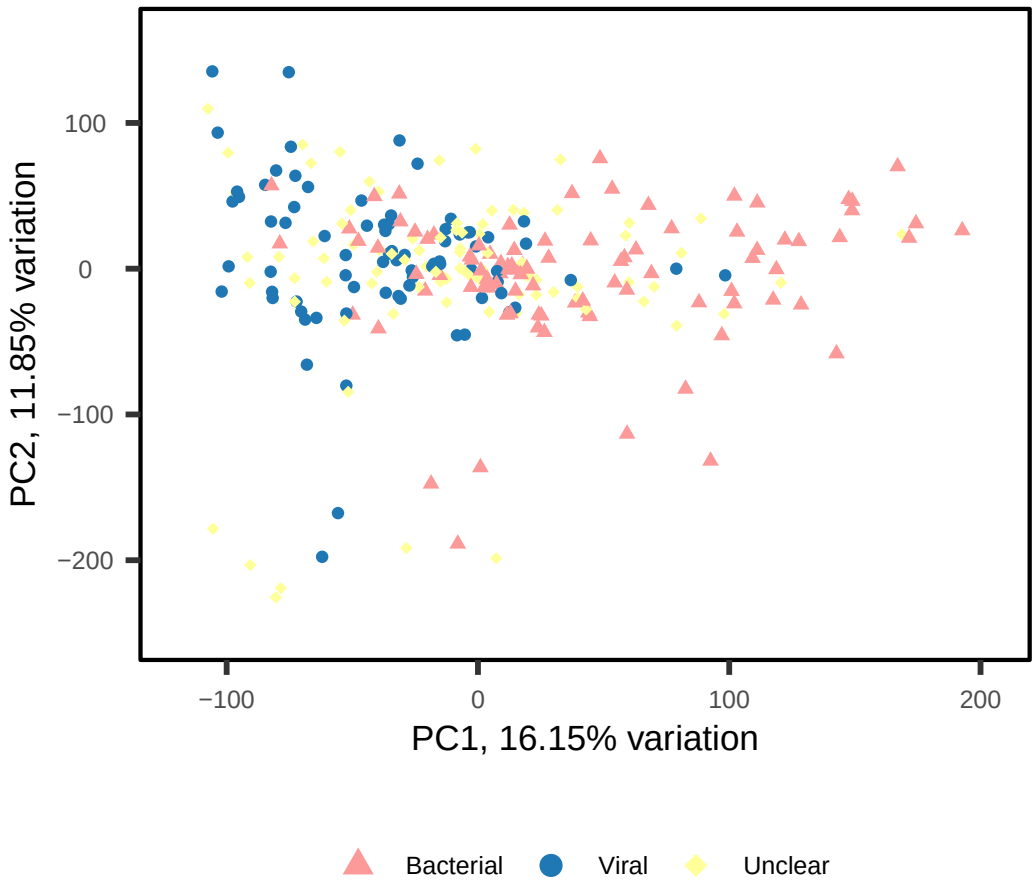


Figure 2.9: Principal component biplot showing PC1 and PC2, coloured by bacterial (probable and definite groups combined), definite viral or unclear groupings, post-correcting for the batch effect due to different RNA-seq experiments. The light colour for the unclear group was deliberately chosen so that the viral and bacterial groups in the LRTI cohort could be highlighted.

2.3.3 Identifying samples for bacterial-viral comparison

It is important to correctly classify bacterial and viral cases. The number of culture-confirmed definite bacterial cases is relatively low, n=16, and most of the bacterial cases are in the probable bacterial group, n=77.

The definite LRTI viral cases, n=77, were identified by testing samples from the nasopharynx. The pathogens found in these upper airway samples are likely to be associated with the LRTI, but we do not have samples from the lower respiratory tract to confirm.

To increase confidence that cases have been assigned correctly, a semi-supervised approach was used.

Semi-supervised classification of bacterial and viral cases

K-means clustering was used to group the acute cases without their labels. Figure 2.10 is a plot of the within-cluster sum of squares, this is a measure of the within-cluster variation. There is an elbow to the plot at cluster five and this would suggest that five is the optimal number of clusters.

Figure 2.11a shows the clusters of cases across the five clusters. Figure 2.11b shows the distribution of cases across the different clusters.

When the labels from known classifications were added, Clusters 2 and 4 were identified as bacterial clusters. The two definite viral cases in Cluster 2 were re-assigned to the unknown group. There were no viral cases in Cluster 4.

Cluster 3 is predominantly viral. Of the 11 bacterial cases in the viral cluster,

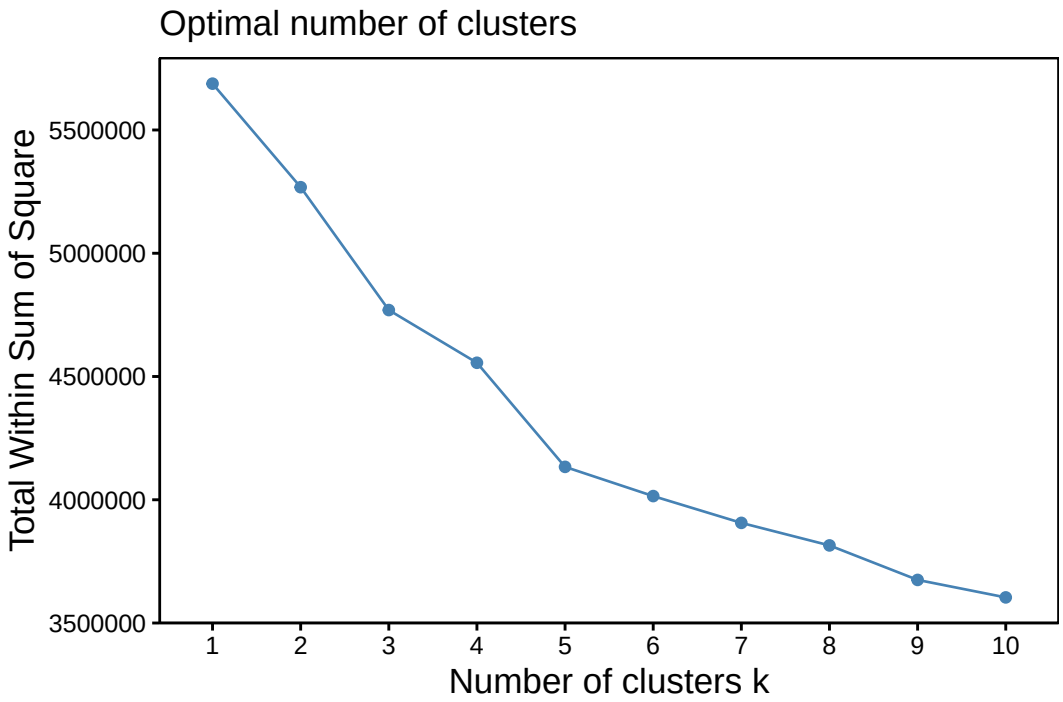


Figure 2.10: When using K-means clustering to assist in classifying cases in the LRTI cohort. Within-cluster variation measured using the within-cluster sum of squares at different numbers of clusters. Using the elbow method, five was taken to be the optimal number of clusters.

nine of these were probable bacterial cases and were re-assigned to the unknown group. The two definite bacterial cases in Cluster 3 were not re-assigned.

Cluster 1 had a mix of bacterial and viral samples. Cluster 5 had more bacterial than viral samples but numbers in Cluster 5 were small. Cases in Clusters 1 and 5 were not reassigned.

Following the incorporation of K-means clustering, there were 65 bacterial and 75 viral cases which underwent DEA and were used to look for differences between bacterial and viral cases at an RNA level.

Co-infections

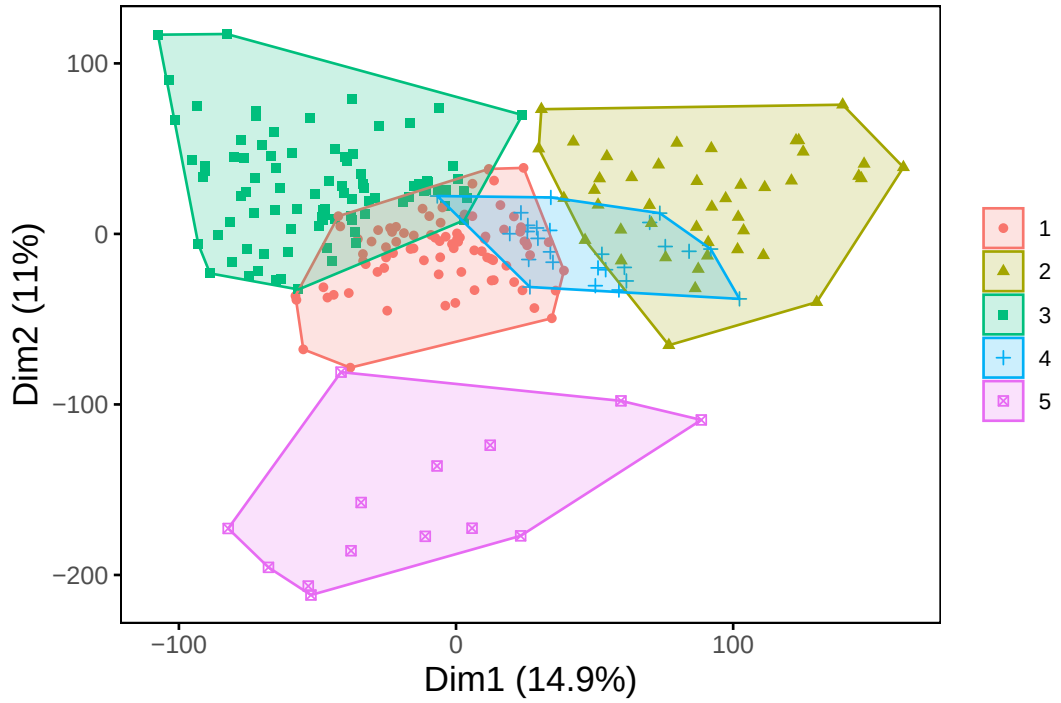
As co-infections may have different RNA patterns, I wanted to look at these samples separately. However, as the number of co-infections was small, n=7, it was difficult to interpret the results in this group, see Appendix B.2.3.

Samples classified as bacterial-viral co-infections met the criteria for inclusion in the definite bacterial group. Clinically, it is more useful for a dichotomous diagnostic test to identify these co-infections as bacterial, and these cases will be grouped with the bacterial cases for creation of a signature to differentiate between bacterial and viral cases.

2.3.4 Differential expression analysis – limma voom

A linear model was fitted to the voom transformed data to test for DGE. Significantly DE genes were examined between the comparisons of interest, see Table 2.1.

(a) Cluster plot



(b) Cases per cluster

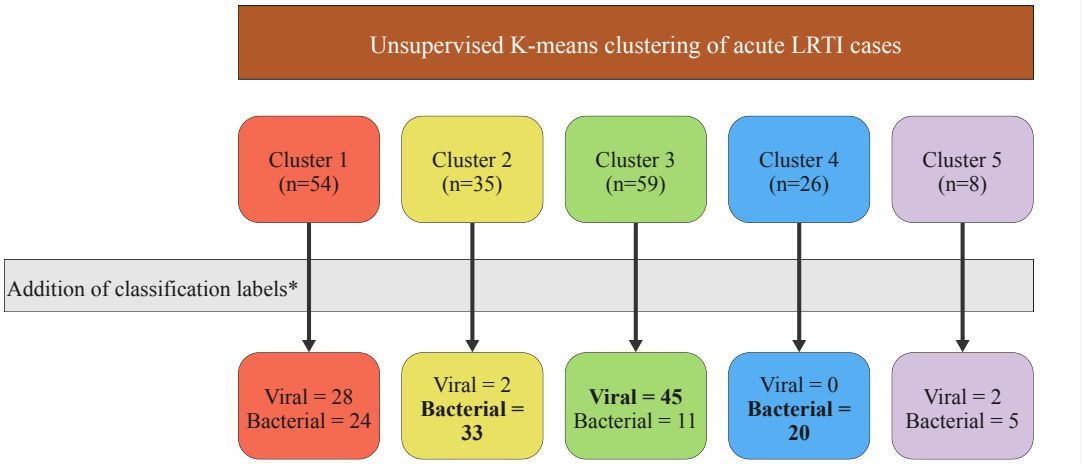


Figure 2.11: (a)Cluster plot showing the distribution of cases when five clusters are taken as optimal. (b) The number of unlabelled cases assigned in each cluster, and the number of bacterial (probable and definite groups) and definite viral cases in each cluster. *Samples with an unclear diagnosis (unknown, viral syndrome and co-infections excluded). Bacterial cases include definite and probable bacterial groupings.

2.3.5 Comparing bacterial and viral lower respiratory tract infections

Initially, acute and convalescent samples were compared. Comparing acute and convalescent samples across the viral and bacterial groups separately, helped to account for differences in age and sex between the bacterial and viral groups.

The volcano plot in Figure 2.12 shows the genes which were significantly differentially expressed between samples from the acute bacterial group (following re-classification using a semi-supervised approach) and convalescent samples from the same participants.

Figure 2.13 shows the DE genes when samples from the acute definite viral group are compared with convalescent samples from the same participants.

The Venn diagram in Figure 2.14 shows the number of genes that are significantly up- and down-regulated when acute bacterial samples are compared with convalescent samples, and when acute viral samples are compared with convalescent samples. There were 1,739 unique genes significantly up- or down-regulated across the two comparisons. This subset of genes, which should exclude DE genes associated with age and sex, were used for the direct comparisons between acute bacterial and acute viral samples.

Directly comparing bacterial and viral LRTIs

The data were split in a 70:30 ratio into training and test datasets. The following model creation was performed on the training dataset which consisted of 100 cases (47 bacterial and 53 viral cases).

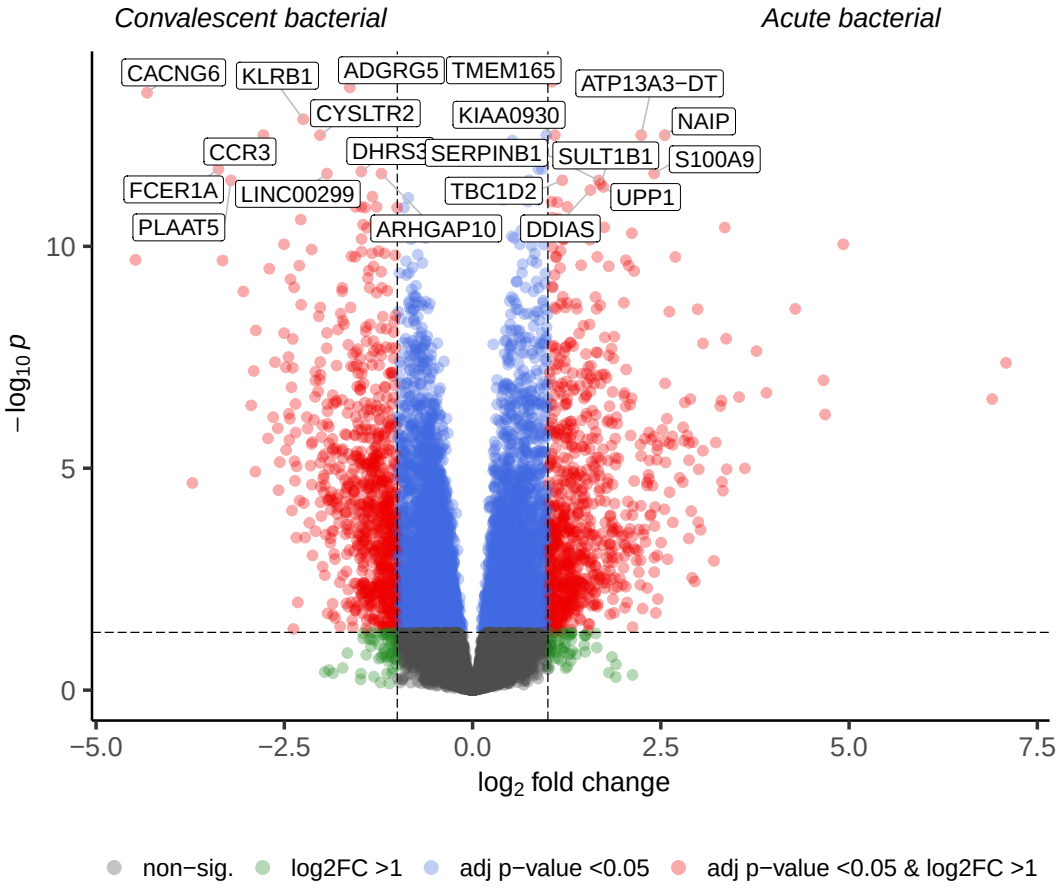


Figure 2.12: Volcano plot showing differentially expressed genes between acute bacterial samples and convalescent samples from participants with bacterial LRTIs. The labelled genes are the top ten up- and down-regulated genes, ranked by Benjamini-Hochberg adjusted p-values.

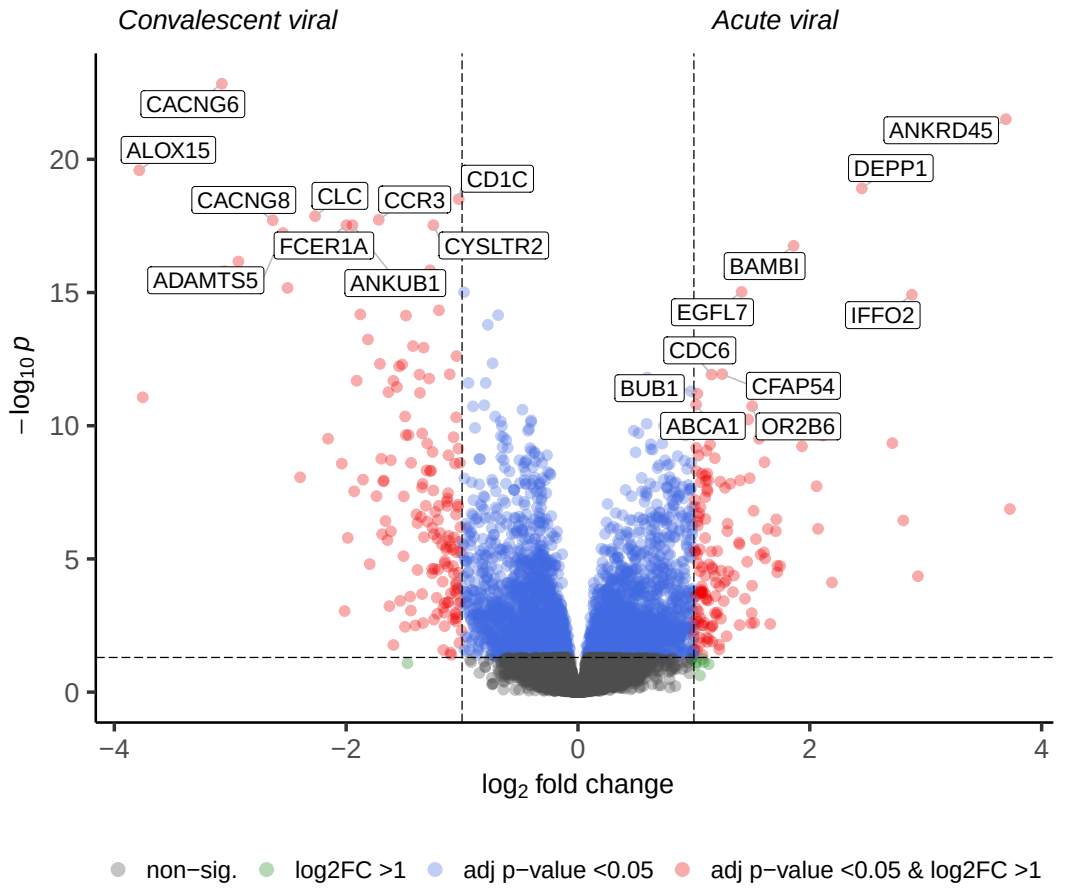


Figure 2.13: Volcano plot showing differentially expressed genes between acute definite viral samples and convalescent samples from participants with viral LRTIs. The labelled genes are the top ten up- and down-regulated genes, ranked by Benjamini-Hochberg adjusted p-values.

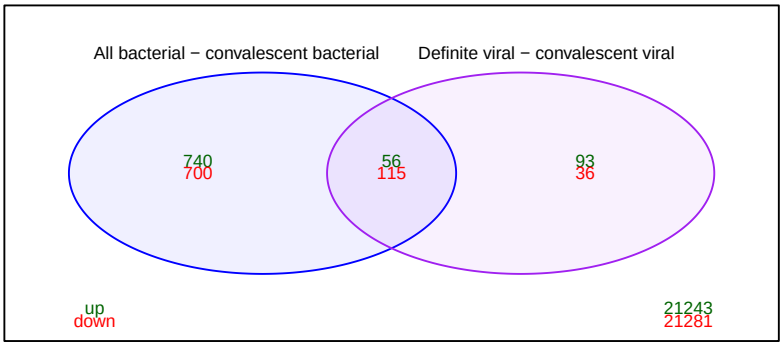


Figure 2.14: The number of significantly differentially expressed genes across two comparisons and the overlapping genes. On the left, the number of up- and down-regulated significantly differentially expressed genes when acute bacterial samples are compared with convalescent samples from the same participants. On the right, the number of up- and down-regulated significantly differentially expressed genes when acute viral samples are compared with convalescent samples from the same participants. Differentially expressed genes were significant if the adjusted p-value was <0.05 and the log2-fold-change was >1.

The subset of 1,739 genes significantly up- or down-regulated when acute samples were compared with convalescent samples was used to directly compare acute viral and acute bacterial groups.

The volcano plot in Figure 2.15 shows the DE genes. In total, 334 genes were significantly relatively increased in the bacterial group, compared with 166 significantly increased in the viral group. This gives a total of 500 DE genes.

These 500 differentially expressed genes were taken forward to use in creating a model to differentiate between bacterial and viral LRTIs.

RNA signatures to differentiate between bacterial and viral LRTIs

A model with bacterial or viral infection as the outcome variable was initially trained using the 500 genes that were differentially expressed between bacterial and viral groups. A partial least squares approach was used, and models were created as per the steps in Section 2.2.5. Slightly different feature selection approaches resulted in four different models potential models being selected, see B for details. Table 2.4 shows the genes included in the different models created with performance statistics. Two eight-gene models were selected, and I have labelled these A and B for differentiation purposes. The two other models contained ten genes and three genes. Figure 2.16 shows the ROC curves for each model. All four models had high AUC. A pAUC (pAUC) was calculated at a sensitivity range of 90% to 100%.

The ten-gene model had the highest pAUC and was taken forward for evaluation in the test dataset. The three-gene model was also taken forward for testing as

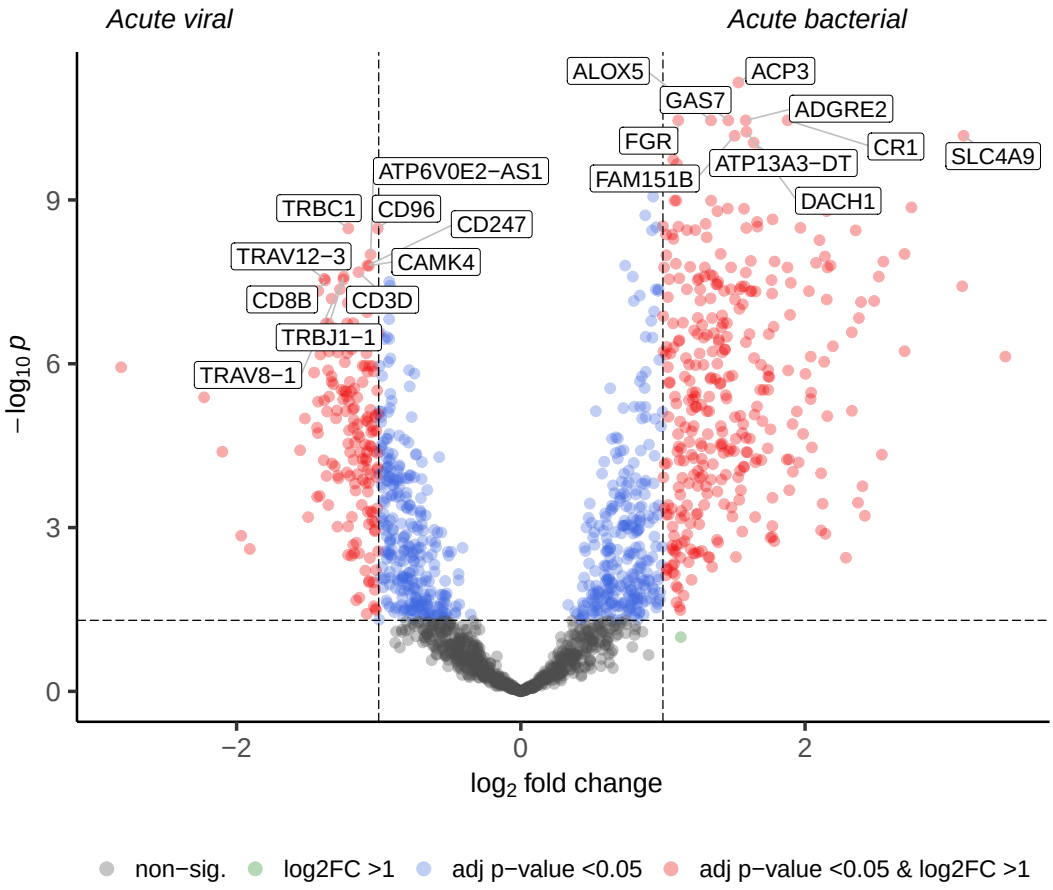


Figure 2.15: Volcano plot showing differentially expressed genes between acute bacterial samples and acute viral samples samples. The top ten up- and down-regulated genes, ranked by adjusted p-value are labelled.

Table 2.4: Different genes selected for inclusion as model parameters when comparing acute bacterial and viral groups. Results of model performance in the training dataset presented.

Performance metric	Ten-gene signature		Eight-gene signature A		Eight-gene signature B		Three-gene signature	
Genes included	ACP3, SLC4A9, ATP6V0E2-AS1, CDCA4P3, LINC00937, CD177P1, DAAM2, ATP2C2, G0S2, FGF13-AS1		ACP3, FAM157B, CCDC13, LST1, CATIP_AS1, SLC49A4, HK3, CAMKK2		ACP3, FAM157B, CCDC13, FAM151B, LINC01002, CREB5, F5, SIRPA		ACP3, ALOX5, B3GNT5	
	Reference - bacterial	Reference - viral	Reference - bacterial	Reference - viral	Reference - bacterial	Reference - viral	Reference - bacterial	Reference - viral
Predicted bacterial	44	0	42	0	44	1	44	2
Predicted viral	3	53	5	53	3	52	3	51
Accuracy (95% CI)	0.97 (0.91, 0.99)		0.95 (0.89, 0.98)		0.96 (0.90, 0.99)		0.95 (0.89, 0.98)	
Sensitivity	0.94 (0.82, 0.99)		0.89 (0.77 to 0.96)		0.94 (0.82 to 0.90)		0.94 (0.82 to 0.99)	
Specificity	1 (0.93, 1)		1 (0.93, 1)		0.98 (0.9, 1)		0.96 (0.87, 1)	
Positive predictive value	1 (0.92, 1)		1 (0.92, 1)		0.98 (0.86, 1)		0.96 (0.85 to 0.99)	
Negative predictive value	0.95 (0.91, 0.99)		0.91 (0.82, 0.96)		0.95 (0.85, 0.98)		0.94 (0.85, 0.98)	

this was a model which performed well with fewer genes. Models which include fewer genes are potentially easier to convert to useful diagnostic tests.

Test data set

The test data set consisted of 18 bacterial and 22 viral cases. The ten-gene and three-gene models were used to identify bacterial and viral cases. Both models identified the same number of cases correctly, 16/18 bacterial and 22/22 viral cases, see Table 2.5 for performance parameters. The pAUC at a sensitivity

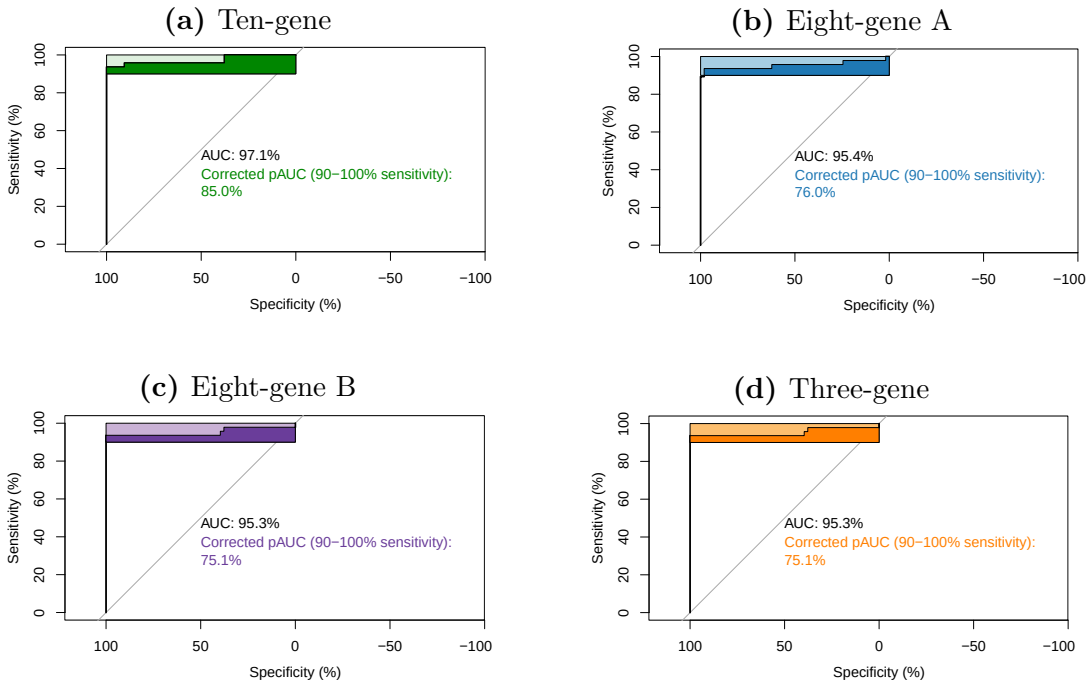


Figure 2.16: Comparing RNA gene signatures from the LRTI cohort study. Receiver operating characteristic (ROC) curves for four different models differentiating between the acute bacterial and acute viral groups, in the training dataset. The area under the curve (AUC) presented for each with a partial AUC at sensitivity range from 90%-100% also shown.

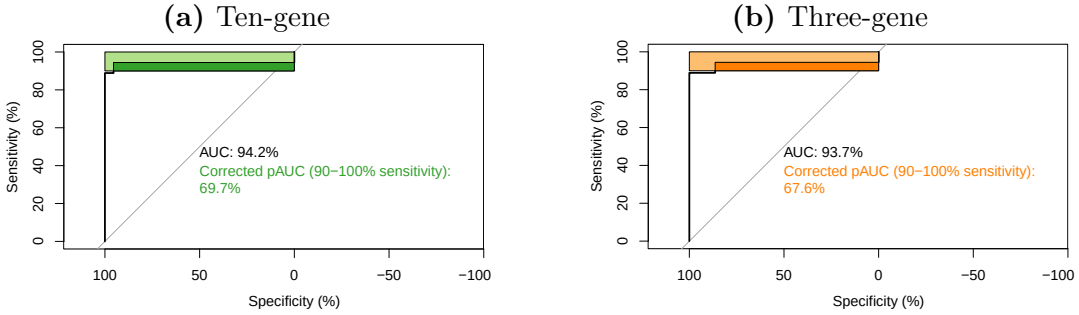


Figure 2.17: Receiver operating characteristic (ROC) curves for four different models differentiating between the acute bacterial and acute viral groups. The area under the curve (AUC) presented for each with a partial AUC at sensitivity range from 90%-100% also shown.

range of 90%-100% for both models were also similar, with pAUC 69.7% for the ten-gene model and 67.6% for the three-gene model, see Figure 2.17.

There were two co-infections included in the test data set bacterial group, one co-infection was identified as bacterial and the other as viral, in both models.

2.3.6 Differential gene expression in pathogen groups

Samples were next grouped based on the likely LRTI pathogen and DGE was repeated. The three pathogen groups with the highest number of cases were the RSV (n=44 acute samples), pneumococcal (n=31 acute samples) and influenza groups (n=18 acute samples). The pneumococcal group was made up of 12 culture-confirmed cases and 19 probable bacterial cases that had pneumococcal serotype 1 or 5 cultured from the nasopharynx. The reasoning for including cases with these serotype results is explained in Section 1.4.1. These three groups were used for DGE analysis.

Table 2.5: The genes selected for the two model parameters used to compare acute bacterial and acute viral groups in the test dataset. Results of model performance in the test dataset presented.

Performance metric	**Ten-gene signature**		**Three-gene signature**	
Genes included	ACP3, SLC4A9, ATP6V0E2-AS1, CDCA4P3, LINC00937, CD177P1, DAAM2, ATP2C2, G0S2, FGF13-AS1		ACP3, ALOX5, B3GNT5	
	Reference - bacterial	**Reference - viral**	**Reference - bacterial**	**Reference - viral**
Predicted bacterial	16	0	16	0
Predicted viral	2	22	2	22
Accuracy (95% CI)	0.95 (0.83, 0.99)		0.95 (0.83, 0.99)	
Sensitivity	0.89 (0.65, 0.99)		0.89 (0.65, 0.99)	
Specificity	1 (0.85, 1)		1 (0.85, 1)	
Positive predictive value	1 (0.79, 1)		1 (0.79, 1)	
Negative predictive value	0.92 (0.75, 0.98)		0.92 (0.75, 0.98)	

Initially, acute samples in each of the three pathogen groups were compared with convalescent samples from the same participants. The Venn diagram in Figure 2.18 shows the numbers of up- and down-regulated genes, and the numbers of genes that overlap between the different pathogen comparisons. The pneumococcal comparison resulted in more DEGs, 2,577 genes, compared with the influenza, 543 genes, and the RSV group, 235 genes. The volcano plots in Figures 2.19, 2.20, 2.21 highlight the top ten up- and down-regulated genes in each comparison.

The DE genes from each of these comparisons were then taken forward to look for genes which are DE when acute samples from pathogen groups were compared with other acute samples.

Directly comparing different pathogen groups

The subset of DE genes when acute and convalescent pathogen groups were compared were used for the following analyses.

To look for genes which are only DE in pneumococcal LRTIs, the acute pneumococcal group was compared with acute samples from participants with other bacterial infections. Of the 2,577 genes compared, 1,030 were differentially expressed between acute pneumococcal and acute other bacterial samples, see Table 2.6 and Figure 2.22. These 1,030 differentially expressed genes will be used for pathway analysis of pneumococcal LRTIs.

To look for genes which are DE in RSV and not in other viruses, the acute RSV group was compared with acute samples from participants in the other

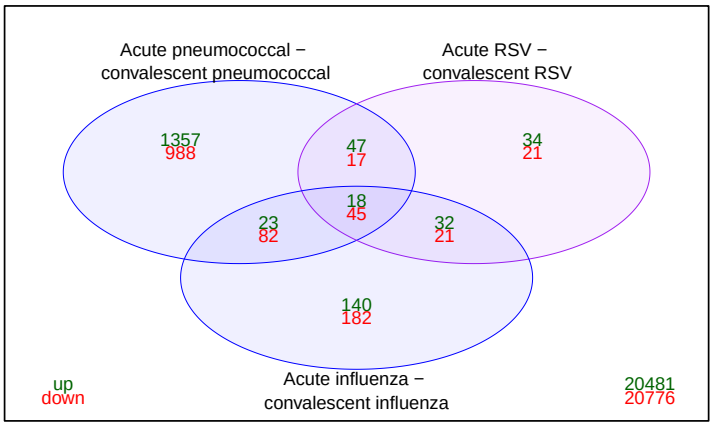


Figure 2.18: The number of significantly differentially expressed genes across three comparisons and the overlapping genes. On the left, the number of up- and down-regulated significantly differentially expressed genes when the acute pneumococcal group are compared with the convalescent pneumococcal group. On the right, the acute RSV group is compared with the convalescent RSV group. Below, the acute influenza group is compared with the convalescent influenza group. Differentially expressed genes were significant if the Benjamini-Hochberg adjusted p-value was <0.05 and the log2-fold-change was >1.

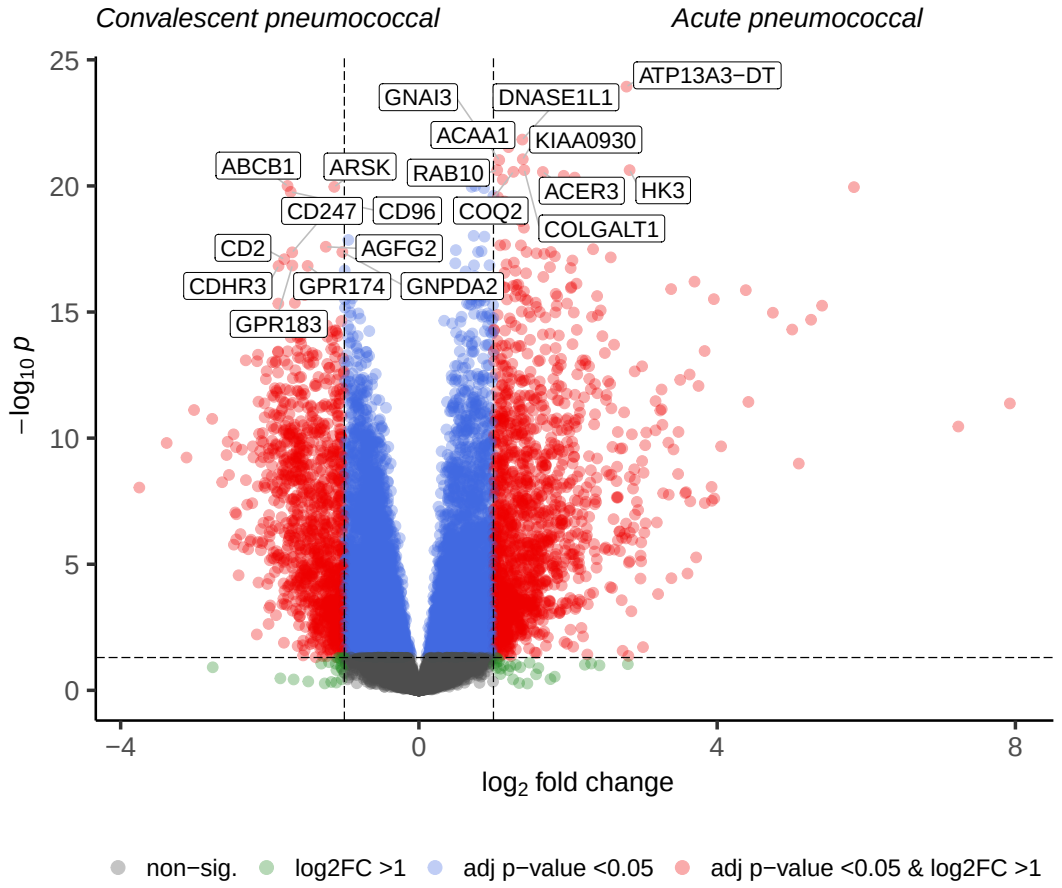


Figure 2.19: Volcano plot showing differentially expressed genes between acute pneumococcal samples and convalescent samples from the same individuals. The top ten up- and down-regulated genes, ranked by Benjamini-Hochberg adjusted p-value are labelled.

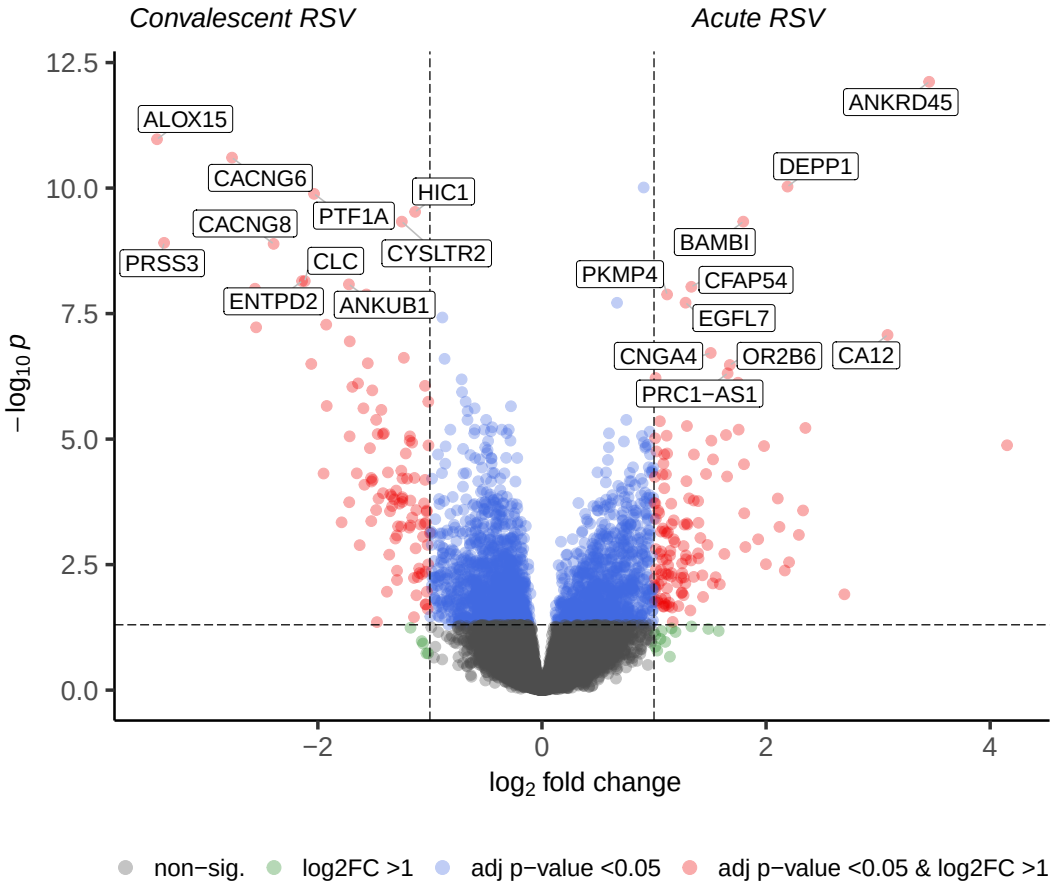


Figure 2.20: Volcano plot showing differentially expressed genes between acute RSV samples and convalescent samples from the same individuals. The top ten up- and down-regulated genes, ranked by Benjamini-Hochberg adjusted p-value are labelled.

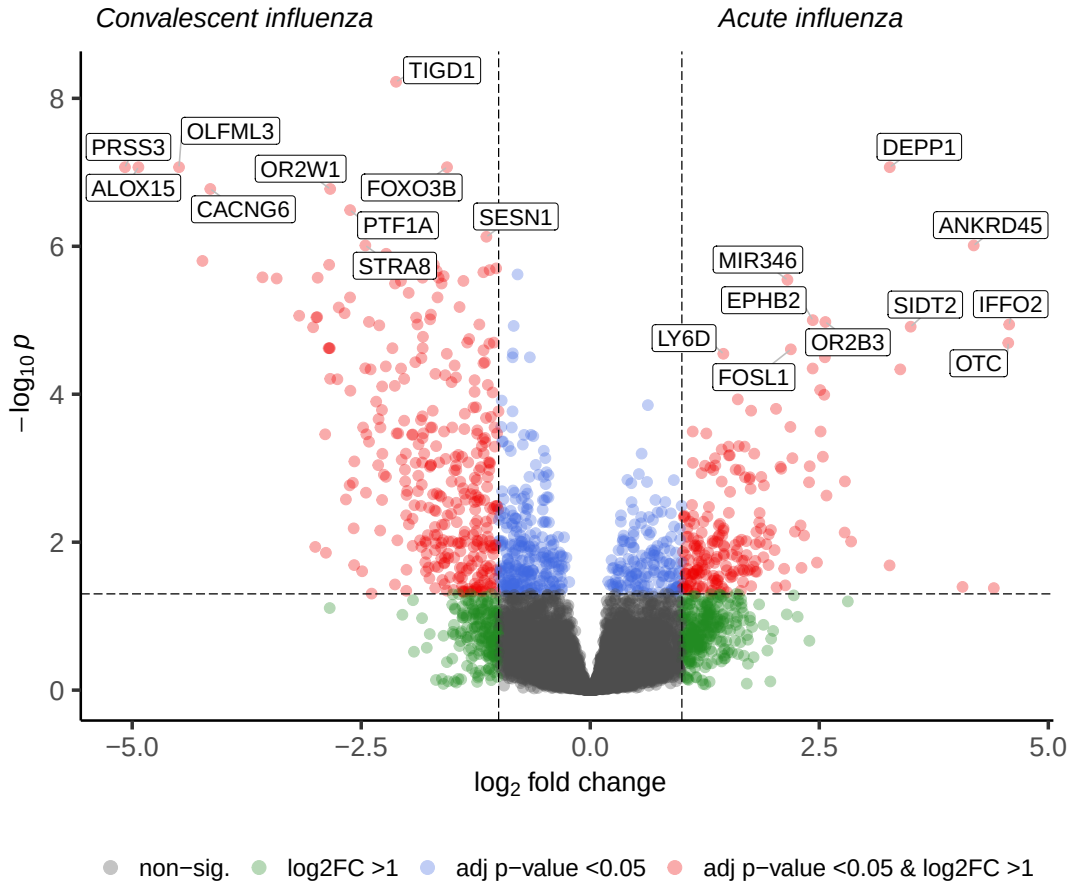


Figure 2.21: Volcano plot showing differentially expressed genes between acute influenza samples and convalescent samples from the same individuals. The top ten up- and down-regulated genes, ranked by Benjamini-Hochberg adjusted p-value are labelled.

Table 2.6: The number of differentially expressed genes when the acute pneumococcal group are compared with the acute other bacterial group, the acute RSV group are compared with other acute viral groups, and the acute influenza group are compared with other acute viral groups.

	Acute pneumococcal compared with acute other bacterial	**Acute RSV compared with acute other viral**	**Acute influenza compared compared with acute other viral**
Adjusted p-value <0.05* and log2-fold-change >1			
Up	588	0	10
Not signficant	1,547	235	530
Down	442	0	3
Total significantly DE genes	**1,030**	**0**	**13**

* p-value adjusted using the Benjamini-Hochberg method.

viral groups. Of the 235 genes which were DE between acute RSV samples and convalescent samples from the same individuals, no genes were DE when acute RSV samples were compared with acute samples from other viral LRTI groups.

The acute influenza group was compared with acute samples from participants in the other viral groups. Of the 543 genes that were DE between acute influenza and convalescent samples from the same individuals, 13 genes were DE when the acute influenza group was compared with acute samples from the other viral LRTI groups, see Table 2.6 and Figure 2.23.

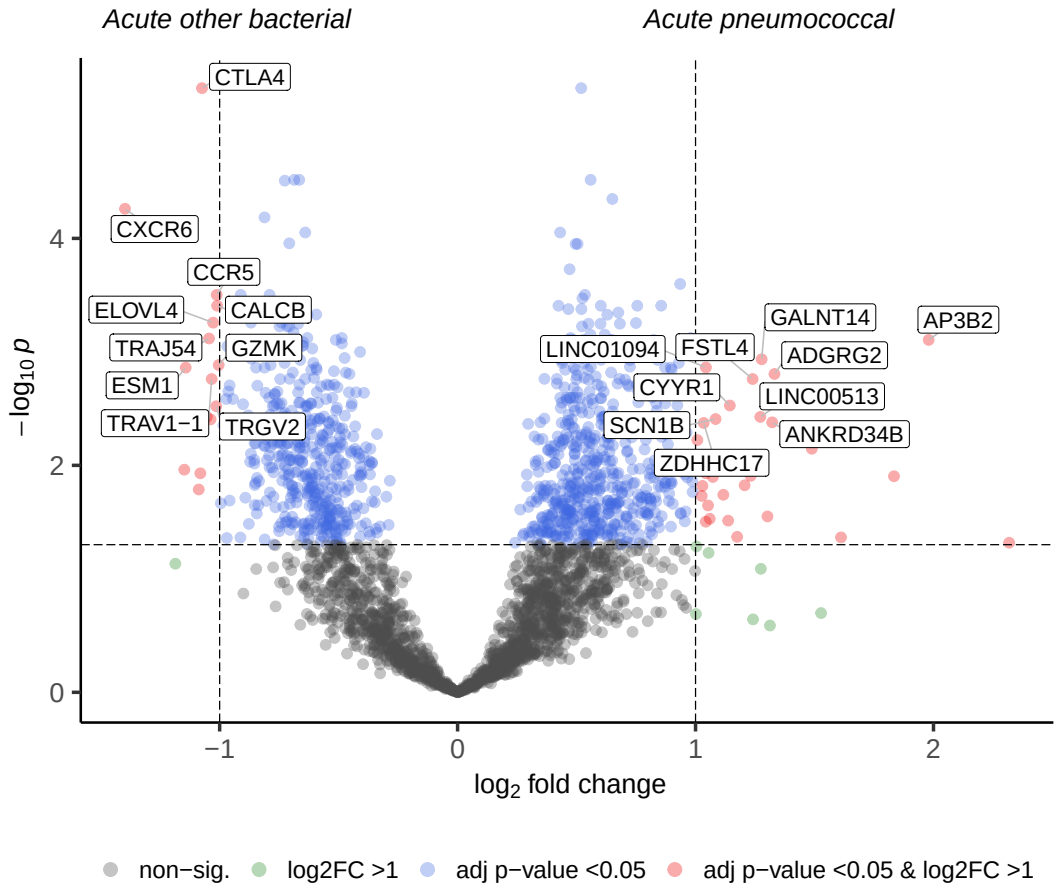


Figure 2.22: Volcano plot showing differentially expressed genes between the acute pneumococcal group and the acute, non-pneumococcal bacterial group. The top ten up- and down-regulated genes, ranked by Benjamini-Hochberg adjusted p-value are labelled.

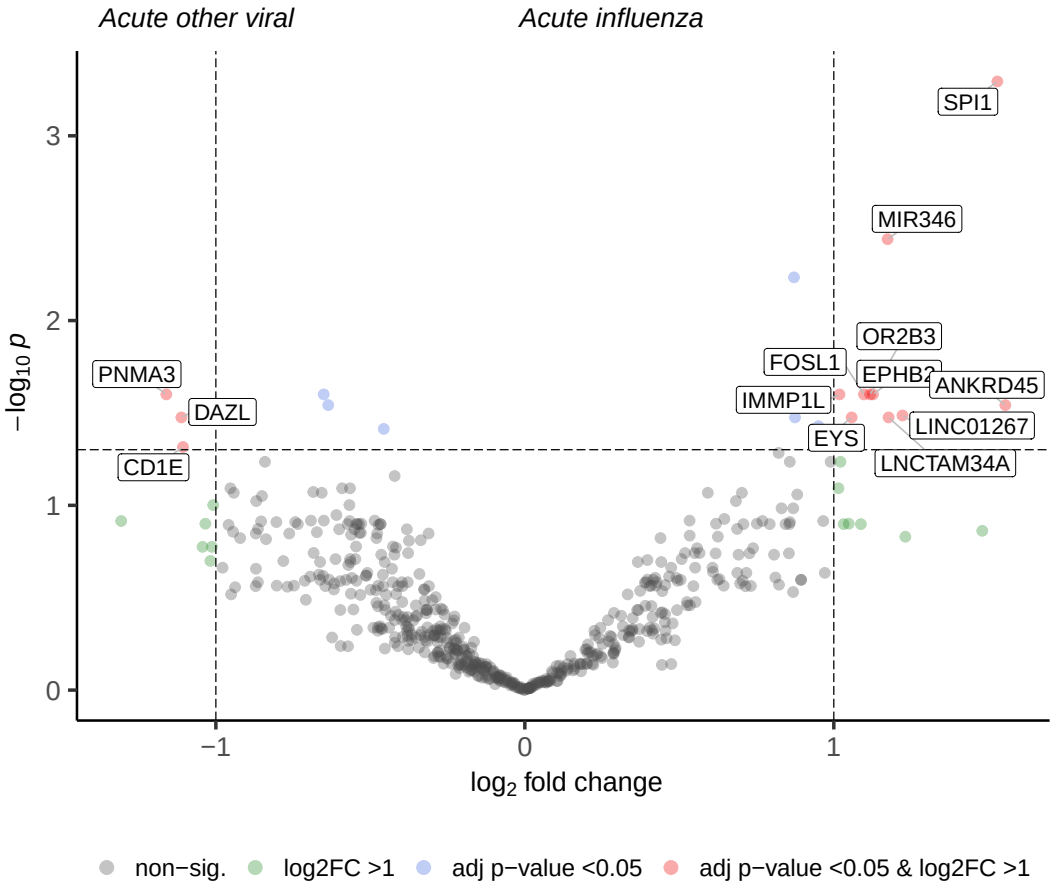


Figure 2.23: Volcano plot showing differentially expressed genes between the acute influenza group and the acute, non-influenza, viral group. The top ten up- and down-regulated genes, ranked by Benjamini-Hochberg adjusted p-value are labelled.

2.3.7 Pathway analysis

Pathway analysis was carried out on the results of gene comparisons across the pneumococcal, RSV and influenza groups. GSEA was performed where the gene lists ranked by log-2-fold-change results for each DGE analysis were used to look for enriched pathways in gene set databases. Two gene sets were used, the GO-BP and the immunologic signature gene sets from MSigDB.[170]

GSEA was carried out in all three pathogen groups looking for enriched pathways in acute infections when compared with convalescent samples. These comparisons used the full list of genes 22,132, post quality control steps.

GSEA was also carried out using only the DGE results when the acute pneumococcal group was compared with other bacterial LRTI. This was to look for pathways which were different between pneumococcal and other bacterial infections. These results are presented below.

Due to the smaller number of genes used when comparing RSV and influenza with other viral infections, GSEA was less useful in these comparisons, and very few significant pathways were found. These results are included in Appendix B.

Pneumococcal - acute compared with convalescent samples

Gene Ontology-Biological Processes The full list of 22,132 genes was ranked based on the log-2-fold-change values when the acute pneumococcal group was compared with convalescent samples from the same individuals. This ranked list was used for GSEA against pathways in GO-BP MSigDB. Using an adjusted p-value threshold of 0.05, 446 were significantly enriched, and 166

gene sets were included in the results after removal of overlapping, redundant pathways. The pathways with the highest and lowest enrichment scores are shown in Figure 2.24.

The pathway with the highest NES represented genes up-regulated in response to Gram-positive bacteria. Pathways related to IL-6 production and regulation were both in the highest-ranked gene sets. The IL-1-alpha production gene set was also enriched. Two pathways with highly enriched scores were related to the generation of reactive oxygen species. Gene sets related to the acute inflammatory response in general, phagocytosis and activation of myeloid leukocyte cells were also over-enriched.

The two pathways with the lowest enrichment scores are related to ribosomal biosynthesis. Other gene sets with significantly low enrichment scores related to T-cell selection, signalling and activation.

When comparing acute pneumococcal and convalescent groups, GSEA was also performed using the immunologic signature gene sets. Immunologic pathways over-enriched included genes over-expressed in monocytes and neutrophils compared with other inflammatory cells, see Appendix B.

Pneumococcal acute compared with other bacterial samples. Next, the subset of genes that were included in the DEA comparing acute pneumococcal LRTIs with other bacterial LRTIs were used for GSEA. This list of 2,577 genes was used to look for pathways that were different between pneumococcal and other bacterial infections. The gene list was ranked by log-2-fold-changes and the GO-BP MSigDB was used to look for enriched pathways. Forty-three gene

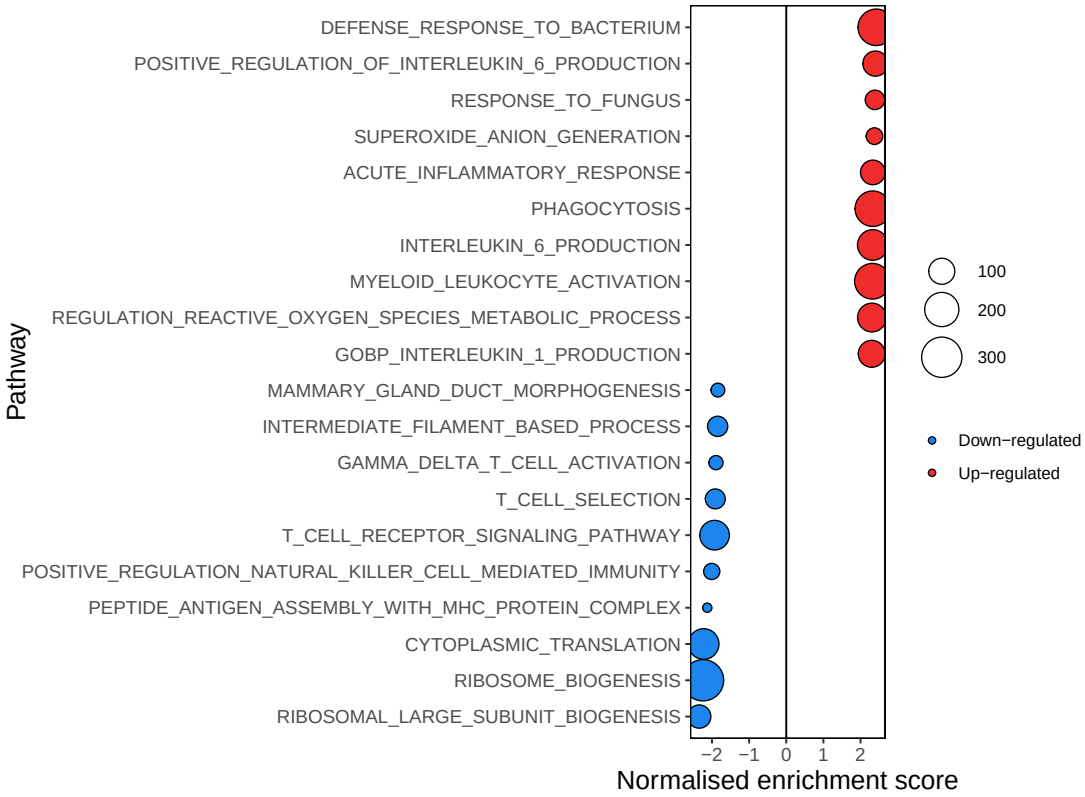


Figure 2.24: Gene-set enrichment analysis results when acute pneumococcal samples were compared with convalescent samples. Using the Gene Ontology Biological Processes gene sets, the ten pathways with the highest and lowest normalised enrichment scores are shown. The size of each circle represents the number of genes in the pathway.

sets were significantly enriched, and following the removal of redundant gene sets 19 gene sets remained, see Figure 2.25.

The top-ranked pathway involved genes involved in lipid movement into cells. Other over-enriched pathways included IL-1 production, IL-6 regulation, IL-8 production, myeloid leukocyte production, chemokine production and genes related to the activation of nuclear factor-kappa-B (NF-kappaB).

Only two pathways had significantly negative enrichment scores, adaptive immune response and T cell selection pathways.

To examine the genes contributing to these pathways, the leading edge genes, which contribute most to the enrichment score, were examined. The leading edge genes for each pathway were compared with the list of DE genes when pneumococcal samples were compared with other bacterial samples. Genes which were common to the leading edge list and DEG list were included in box plots. Figure 2.26 shows the differentially expressed leading edge genes for the IL-1 production pathway. Similarly, Figure 2.27 shows the differentially expressed leading edge genes for the IL-6 production pathway. Figure 2.28 shows the differentially expressed leading edge genes for the IL-8 production pathway. leukocyte immunoglobulin-like receptor A2 (LILRA2) and toll-like receptor 6 (TLR6) are common to the three pathways, and have higher expression in the pneumococcal group compared with other bacterial infections.

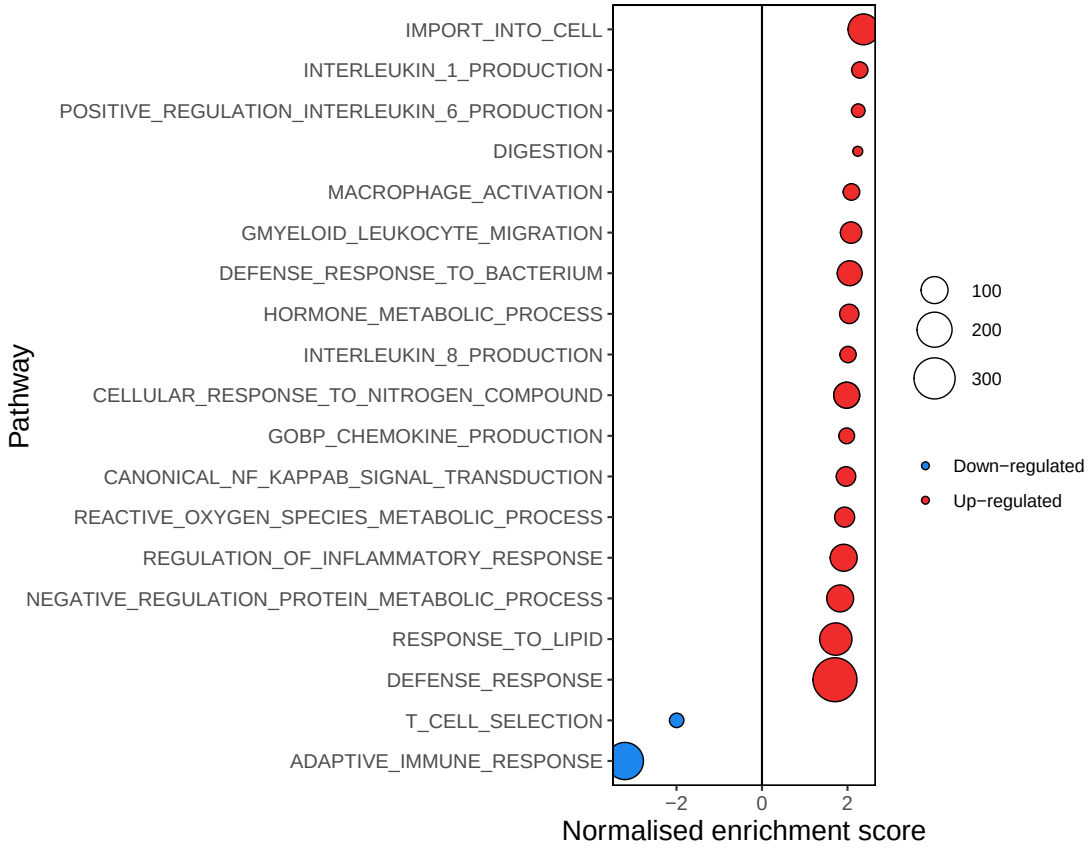


Figure 2.25: Gene-set enrichment analysis results when acute pneumococcal samples were compared with other bacterial samples. Using the Gene Ontology-Biological Process gene sets, all of the significantly enriched pathways are shown. The size of each circle represents the number of genes in the pathway.

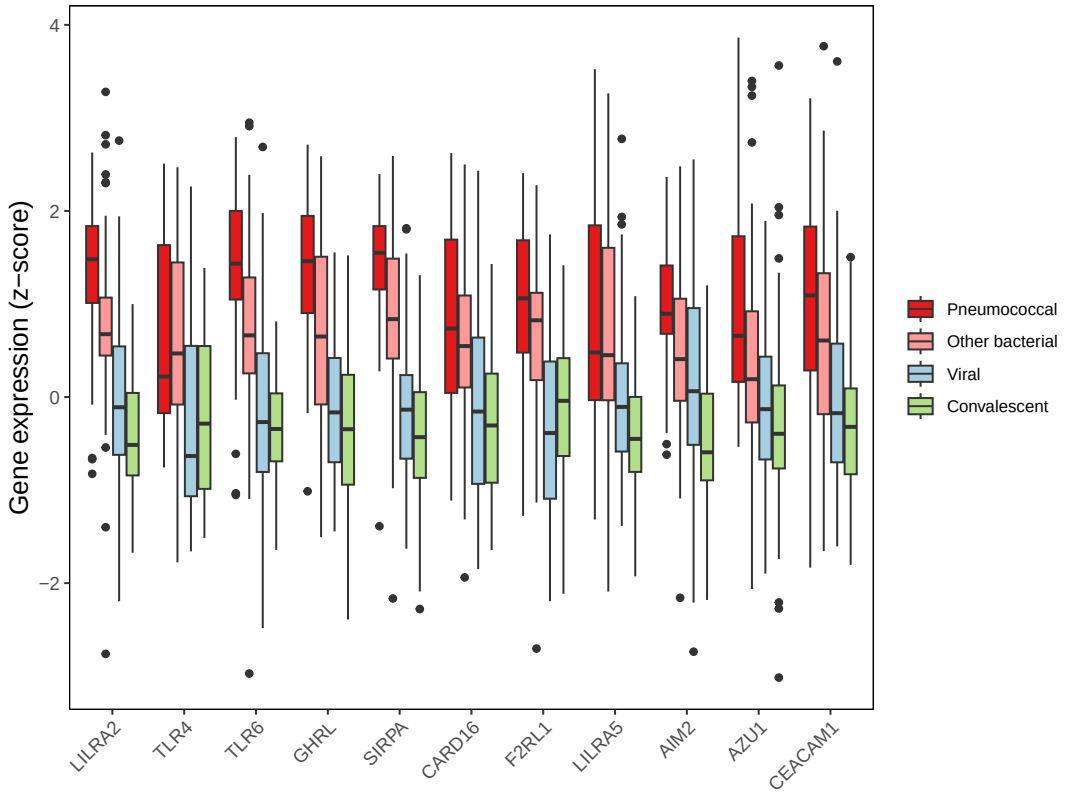


Figure 2.26: Box plots of the common genes between the leading edge of the gene set for IL-1 production, GO-BP INTERLEUKIN 1 PRODUCTION, and DE genes when acute pneumococcal samples were compared with convalescent samples.

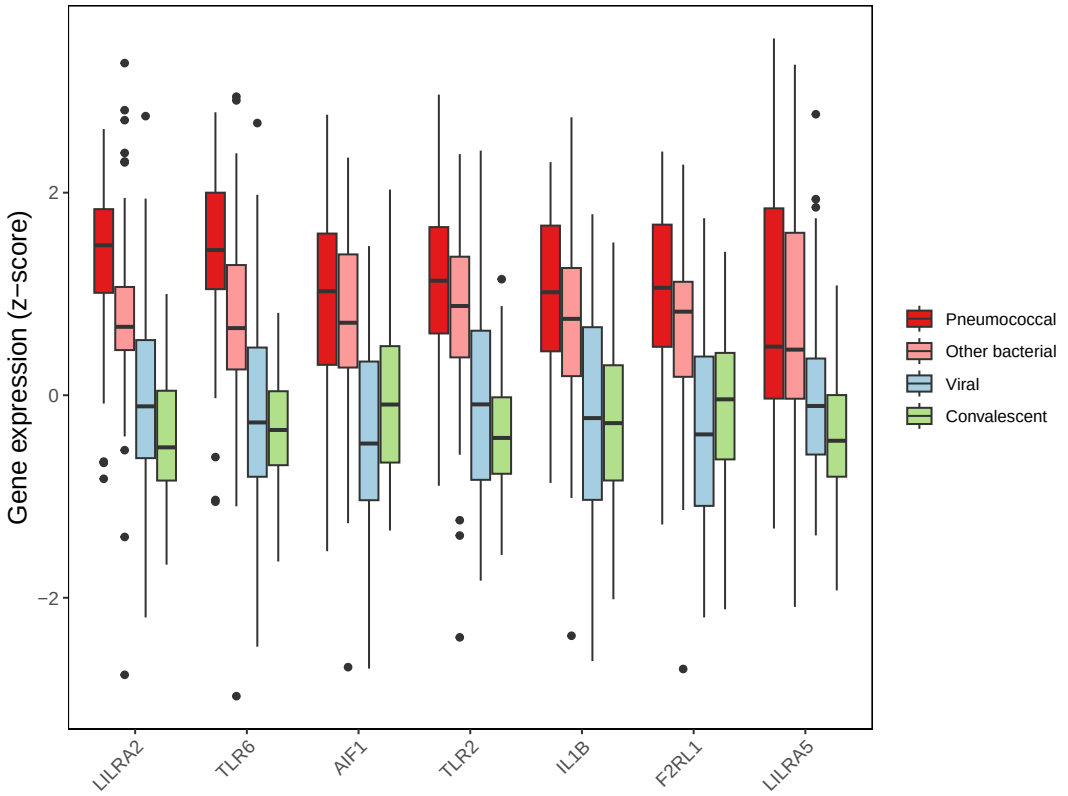


Figure 2.27: Box plots of the common genes between the leading edge of the gene set for positive regulation of IL-6 production, GO-BP POSITIVE REGULATION OF INTERLEUKIN 6 PRODUCTION, and DE genes when acute pneumococcal samples were compared with convalescent samples.

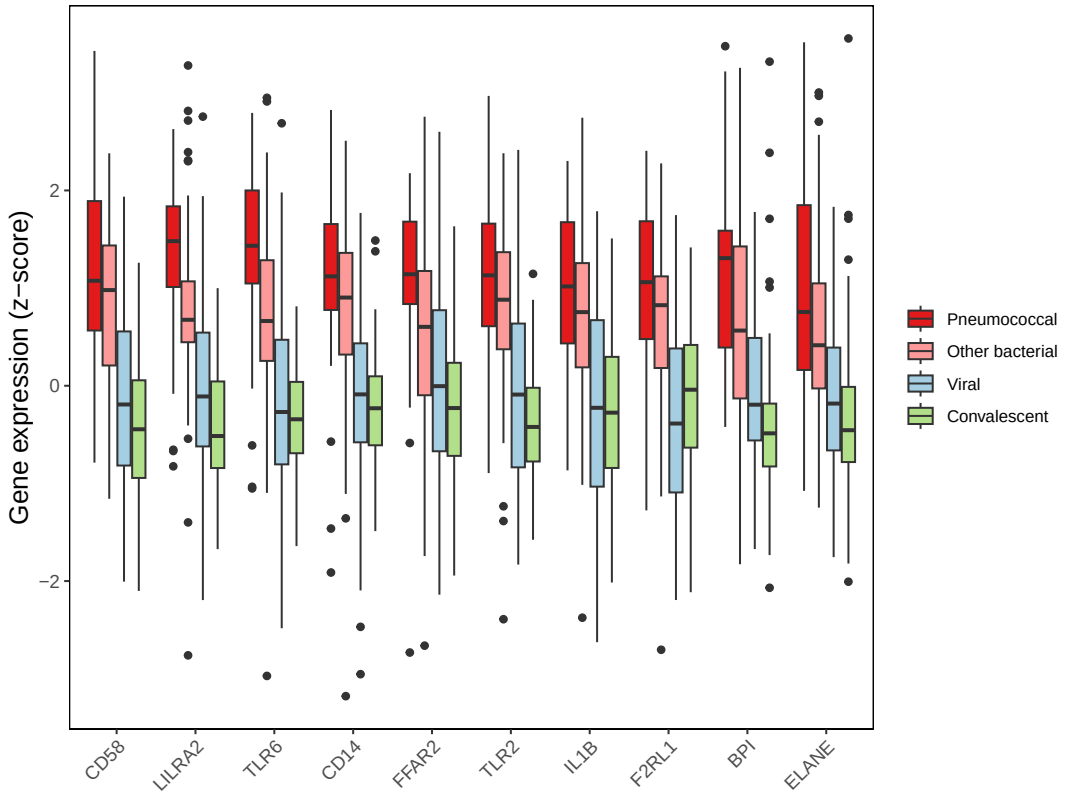


Figure 2.28: Box plots of the common genes between the leading edge of the gene set for IL-8 production, GO-BP INTERLEUKIN 8 PRODUCTION, and DE genes when acute pneumococcal samples were compared with convalescent samples.

RSV

The full list of 22,132 genes were ranked based on the log-2-fold-change values when the acute RSV group was compared with convalescent sample from the same individuals. This ranked list was used for GSEA of pathways in GO-BP MSigDB, 115 significantly enriched pathways were found. After removing redundant gene sets 60 significant gene sets remained.

The highest-ranked pathways based on enrichment scores are linked to mitosis and chromosome separation, see Figure 2.29. Other highly enriched pathways related to varying biological processes including protein-DNA complex assembly, nucleosome organisation and bile acid transport. The pathways with the lowest enrichment scores included genes linked to ribosomal activity and gamma-delta T cell activation.

This GSEA in the RSV group was repeated using the immunologic signature gene sets. After removing overlapping pathways, 447 significant gene sets were found. Pathways over-enriched included several pathways related to vaccines including an influenza vaccine. Figure 2.30 shows the top-ten up- and down-regulated gene sets. Two pathways, including the top-ranked pathway, were from the same study related to genes activated in CD8 effector T cells in wild-type mice, compared with those deficient in TNF receptor-associated factor 6 (TRAF6). Figure 2.31 shows the genes that were differentially expressed in the data and also present in the TRAF6-related pathway. The box plots for these genes show that they are up-regulated in RSV compared with bacterial or convalescent samples. However, the RSV group had similar gene expression levels to samples from those with other viral classifications.

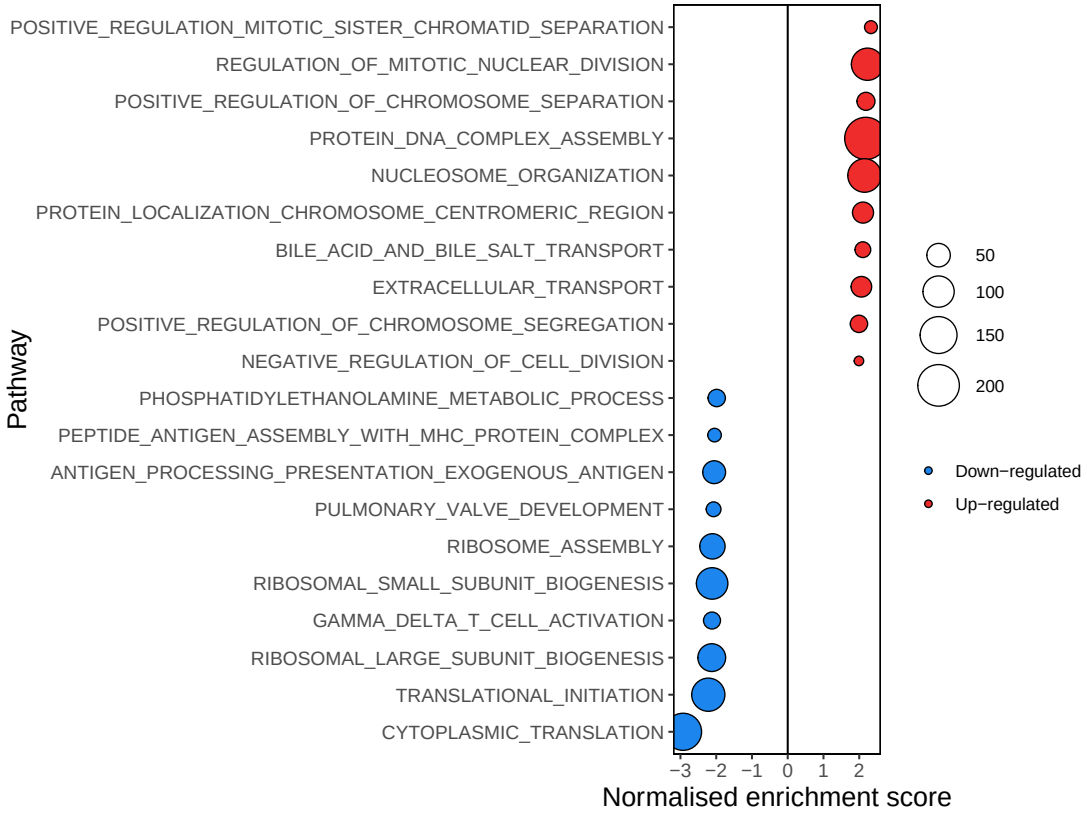


Figure 2.29: Gene-set enrichment analysis results when acute RSV samples were compared with convalescent samples. Using the Gene Ontology Biological Process gene sets, the ten pathways with the highest and lowest normalised enrichment scores are shown. The size of each circle represents the number of genes in the pathway.

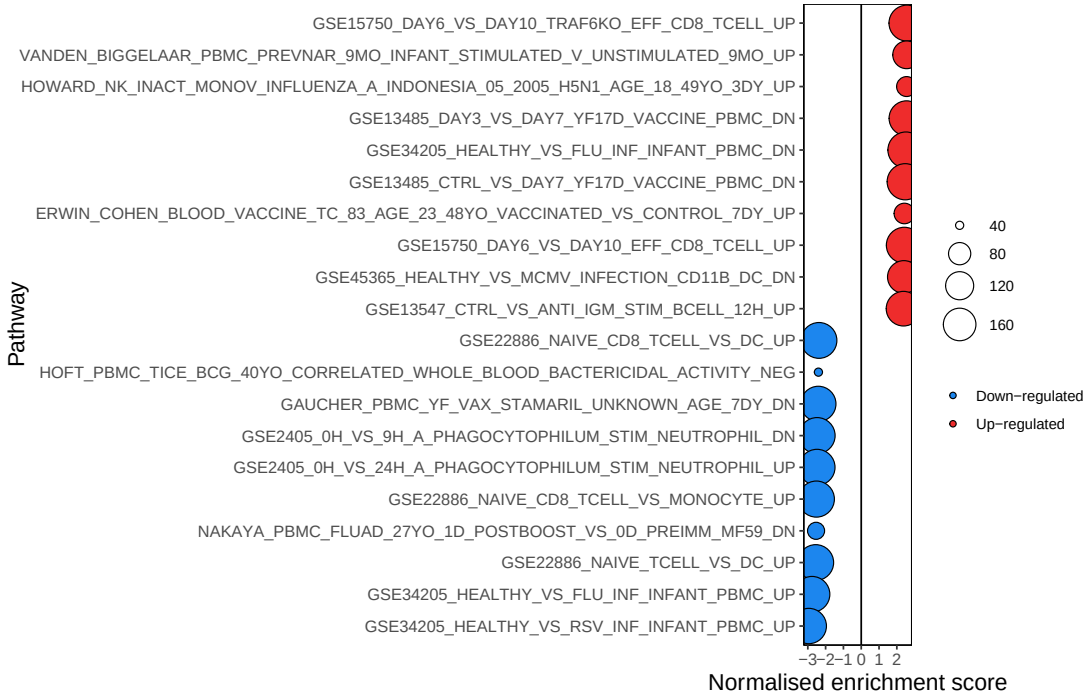


Figure 2.30: Gene-set enrichment analysis results when acute RSV samples were compared with convalescent samples. Using the immunologic signature gene sets, the ten pathways with the highest and lowest normalised enrichment scores are shown. The size of each circle represents the number of genes in the pathway.

The pathway with the lowest NES score includes genes up-regulated in healthy donors when compared with genes in patients with acute RSV infection. Other pathways with low NES values relate to CD8 T cells and genes negatively correlated with bactericidal activity.

Influenza

The full gene list was ranked according to log-2-fold-change when DEA was carried out between the acute influenza group and convalescent samples from the same participants. GSEA was repeated with this ranked list, searching for enriched gene sets in the GO-BP MSigDB. Sixty-nine gene sets were significantly

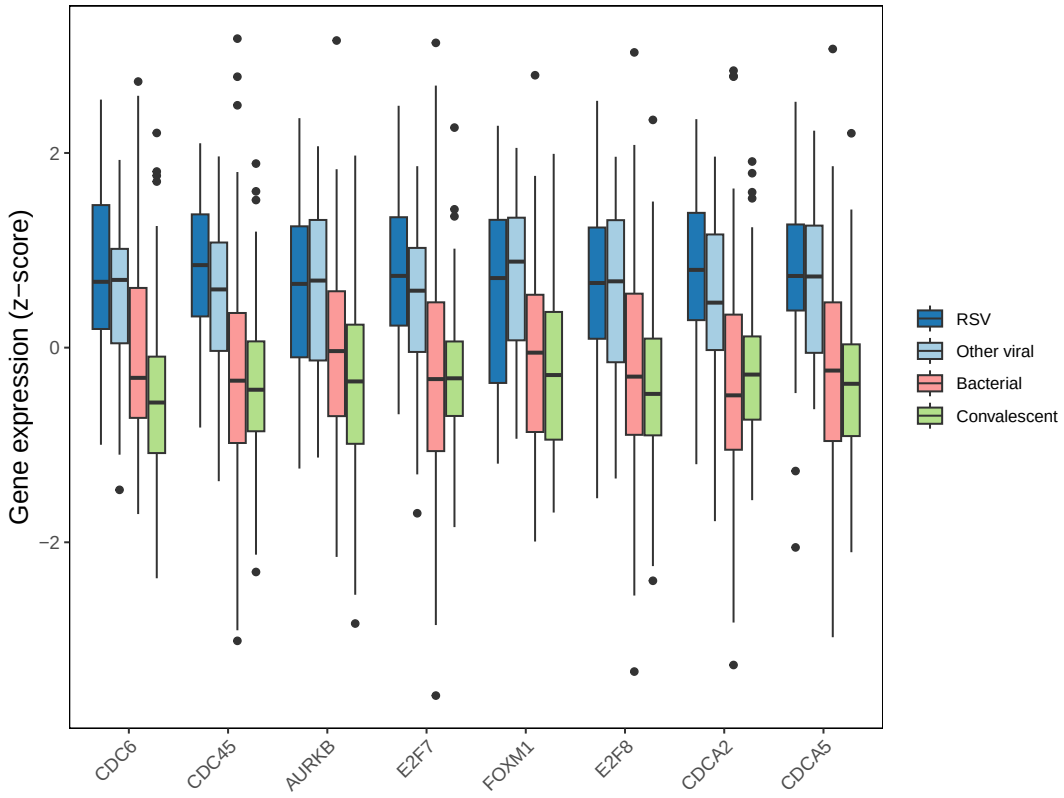


Figure 2.31: Genes from the top-ranked gene-set enrichment analysis when acute RSV samples were compared with convalescent samples, GSE15750 DAY6 VS DAY10 TRAF6KO EFF CD8 TCELL UP. These genes were differentially expressed when the acute RSV group was compared with convalescent samples. Box plots show the scaled levels of gene expression (z-scores) across the acute RSV group, other viral group, bacterial group and convalescent groups. NES, normalised enrichment score.

enriched, and following removal of redundant gene sets, 32 gene sets remained. Figure 2.32 shows the gene sets with the highest and lowest enrichment scores.

Several of the pathways with the highest enrichment scores in the RSV comparison also were found in the influenza comparison.

Immunoglobulin production was the pathway with the highest enrichment score. B cell mediated immunity was the other immune-related pathway with a high enrichment score. Only five pathways had significant negative enrichment scores including genes linked to the breakdown of hydrogen peroxide and cellular response to stimuli with copper ions.

GSEA was then repeated for the influenza group using the immunologic signature gene sets, 317 significant pathways were found. When redundant pathways were removed, 239 pathways remained. Figure B.14 shows the ten pathways with the highest and lowest NES.

The pathways with the highest NES again include pathways related to vaccination including two influenza vaccines. The top-ranked pathway related to gene changes following influenza A vaccination. One pathway was related to genes up-regulated in acute influenza infection. One pathway was related to *E. coli* and *S. pneumoniae* infections. The same TNF receptor-associated factor (TRAF)6 knockout study that was highlighted in the RSV GSEA was also in the top-ten influenza pathways, see Figure 2.31.

The two pathways with the most negative NES were related to genes which were elevated in acute RSV infection compared with influenza infection.

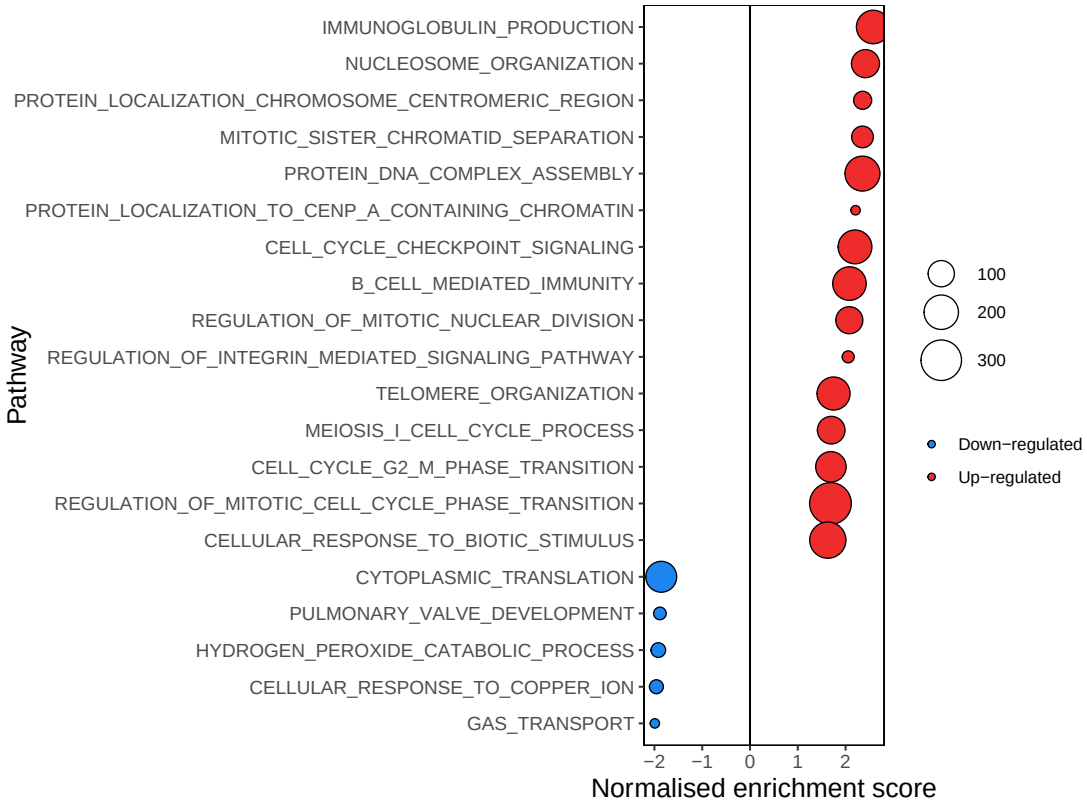


Figure 2.32: Gene-set enrichment analysis results when acute influenza samples were compared with convalescent samples. Using the Gene Ontology Biological Process gene sets, the ten pathways with the highest and lowest normalised enrichment scores are shown. The size of each circle represents the number of genes in the pathway.

2.4 Discussion

In this chapter, I describe the analysis of the plasma transcriptome of 258 Nepali children with acute LRTIs. DEA comparing bacterial and viral LRTIs identified 500 DE genes. These 500 genes were used to train models that could differentiate between bacterial and viral infections. A three-gene and a ten-gene signature both performed well when differentiating between bacterial and viral cases in the test data set.

As expected, there were relatively few culture-confirmed bacterial cases in the cohort, *S. pneumoniae* was the most commonly isolated pathogen. RSV and influenza were the most commonly detected viruses. GSEA identified several pathways which were enriched in pneumococcal infections compared with other bacterial infections. Pathway analysis for RSV and influenza identified several pathways that were common to both infections.

2.4.1 Differentiating between bacterial and viral infections

A three-gene signature performed reasonably with high sensitivity and specificity to differentiate between bacterial and viral LRTIs. The three-gene signature consists of acid phosphatase 3 (ACP3), arachidonate 5-lipoxygenase (ALOX5) and
UDP-GlcNAc:betaGal beta-1, 3-N-acetylglucosaminyltransferase 5 (B3GNT5). All three transcripts had higher relative expression in the bacterial group. These three genes have been linked to infectious processes in previous reports.

ACP3 encodes a phosphatase protein. Phosphatases act as enzymes that hydrolyse phosphate esters, which are involved in a variety of processes including cellular signal transduction and energy metabolism.[171] These enzymes are grouped based on their optimum pH. Acid phosphatases are found in different organisms including plants, animals and bacteria. Acid phosphatases are secreted by both Gram-positive and Gram-negative bacteria.[172, 173, 174] While ACP3 in humans encodes for prostatic acid phosphatase and has not been reported to be a marker of acute infection, the increased ACP3 expression could be due to presence of bacteria in the blood.[175]

ALOX5 catabolises arachidonic acid and initiates the synthesis of leukotrienes.[176] These proinflammatory leukotrienes act as neutrophil and macrophage chemoattractants. Activation of various leukotrienes can lead to tissue oedema, mucus secretion and bronchoconstriction.[177] One study has suggested that the activation of ALOX5 and the subsequent release of pro-inflammatory leukotrienes could contribute to severity in COVID-19 disease. Another study which reported leukotriene B4 as a possible marker for COVID-19 linked this to the activation of ALOX5 in COVID-19 patients.[177, 178] Outside of infections, ALOX5 has been suggested as a prognostic marker for glioma, linked to immune cell infiltration.[179]

B3GNT5 encodes a glycosyltransferase protein which is an important protein in the synthesis of carbohydrate chains on glycolipids.[180] In acute *Helicobacter pylori* infections, up-regulation of B3GNT5 was reported to be associated with TNF-induced activation of the NF-kappaB pathway.[181] B3GNT5 has been reported to be up-regulated in patients with COVID-19 in one study.[182]

However, B3GNT5 has not widely been reported as a biomarker for acute infection.

The combination of these three transcripts differentiated well between bacterial and viral infections in this cohort. It is more important for any bacterial-viral differentiation to avoid missing serious bacterial infections, and for this reason a higher sensitivity is desirable. When the ROC is restricted to a sensitivity range of 90%-100%, the AUC falls from 93% to 68%. This pAUC is a more useful measure of how the test would perform if used clinically.

As outlined in Section 1.7.1, RNA signatures to differentiate between bacterial and viral infections have previously been published.[103, 100] I would have liked to examine RNA signatures generated in previous work and test them in this cohort. However, the final list of filtered gene counts in this study did not contain all of the genes required for this.

When comparing bacterial and viral LRTIs using DGE analysis in this chapter, ALOX5 was one of the top ten differentially expressed genes in the bacterial group, when ranked by adjusted p-value. A closely related gene, arachidonate 5-lipoxygenase activating protein (ALOX5AP) was included in the ten-gene signature reported by Mahajan *et al.* which had a high sensitivity for identifying bacterial infections in young infants.[100] It is encouraging that similar genes are found in a study of American infants, Mahajan *et al.* and the data from Nepali children with LRTIs. This would highlight that these genes related to leukotriene synthesis could be useful biomarkers in diverse populations.

Several genes included in the signatures in this chapter were not included in the RNA signatures in the previously published large cohort studies differentiating

bacterial and viral infections in children.[103, 100] There could be several reasons for this including the different profiles of the populations in these studies, the different infections in different areas or technical differences across the studies. It is known that there are genetic differences in RNA expression, and the background genetic differences in this Nepali cohort likely played a role in which genes were identified as important.[183, 184] This will make it important to validate any new RNA biomarkers in diverse populations, to ensure that RNA signatures provide robust results in different populations.

2.4.2 Clinical use of RNA signatures

The signatures reported in the chapter, and some of the previously reported signatures, perform well at differentiating between bacterial and viral infections. Taking any RNA signatures and incorporating it into clinical practice will require high confidence in the test results. Missed bacterial infections can have catastrophic consequences for patients and reduce user confidence in new tests.[3] While both the three-gene and ten-gene signatures reported in this chapter differentiated between bacterial and viral infections with a high level of accuracy, both signatures misclassified 2 out of 18 bacterial cases in the test dataset. This would be a concern if these tests were to be used clinically. However, no diagnostic test should be used in isolation and RNA signatures, incorporated with clinician experience and currently available diagnostic test, could improve identification of children with serious infections.

2.4.3 Pathway analysis

The cases enrolled in the cohort allowed for pathway analysis of pneumococcal, RSV and influenza groups. There were more DE genes in the pneumococcal comparisons compared with RSV and influenza. This allowed for more detailed analyses in the pneumococcal group.

When acute pneumococcal cases were compared with acute cases with other bacterial diagnoses there were several enriched pathways of interest. Pathways related to IL-1, IL-6 an IL-8 production pathways were enriched in the pneumococcal group. The canonical NF-kappaB pathway was also enriched in the pneumococcal group, this pathway is stimulated by pro-inflammatory cytokines like IL-1-beta and TNF-alpha.[185] In Chapter 4, the plasma levels of IL-6, IL-8 and TNF-alpha are measured in the same individuals, and there I discuss cytokines measurements further.

When looking at the individual genes contributing to the enrichment of the pathways in the pneumococcal group compared with other bacterial cases, a small number of genes were increased in the pneumococcal group in several pathway, including TLR6 and LILRA2. Toll-like receptors (TLRs) are an important component of the innate immune system and different TLRs recognise different microbial patterns. TLR6 is a cell surface receptor present on respiratory epithelial cells, and responds to bacterial and fungal infections. In mouse studies, decreased levels of TLR6, among other TLRs, have been associated with increased susceptibility to pneumococcal infection.[186, 187] However, to my knowledge there is no study which suggests that TLR6 is more important in *S. pneumoniae* compared with other bacterial infections. Leukocyte

immunoglobulin-like receptors (LILRs) are a group of proteins that play a role in regulating the immune response. LILRA2 can help activate neutrophils and monocytes in response to bacteria. LILRA2 recognises antibodies to bacteria and fungi, including *S. pneumoniae* and *Candida species.*[188, 189] From reviewing the literature, LILRA2 has not been reported to be expressed at higher levels in *S. pneumoniae* when compared with other bacterial infections. The suggestion in this study that LILRA2 is important in *S. pneumoniae* would need to be further validated. Examining the different RNA and proteins with increased levels in pneumococcal infections can aid our understanding of pneumococcal disease processes, and this is discussed further when these results are interpreted together in Chapter 5.

When the acute RSV group was compared with convalescent samples, several of the pathways with significant enrichment scores in the GO-BP gene set were not directly related to the immune response but due to the effect of the viral infections on normal cellular function. For example, several ribosomal biogenesis pathways were down-regulated in RSV-infected children compared with convalescence. RNA viruses, such as RSV, are known to decrease host protein production by inhibition of ribosome assembly. This disruption of ribosomes can interrupt normal cellular function and lead to cell death.[190]

Immune-related pathways highlighted in RSV included down-regulation of genes related to gamma-delta T cell subsets. Gamma-delta T cells are a minor proportion of circulating lymphocytes, however, they have been reported to contribute to anti-cancer immunity, autoimmune disease and defence against infection.[191] Circulating levels of gamma-delta T cells have been reported to be low in different

viral infections including dengue fever and COVID-19 disease.[192, 193, 194] In a small study of children with severe RSV infection, de Weerd *et al.* report that gamma-delta T cell levels were lower in ventilated children with RSV than in non-ventilated children.[195] It is not entirely clear in the literature whether the low gamma-delta T cells in viral infections are due to T cell exhaustion, movement of cells to tissues or decreased production. The results reported in this chapter would suggest that at the RNA level, RSV infection leads to decreased activation of gamma-delta T cells and that this would contribute to decreased circulating levels.

GSEA of the RSV and influenza groups resulted in some common gene sets having the highest enrichment scores. For example, genes up-regulated in a study of wild-type mice compared with TRAF6 deficient mice was in the top-ten enriched pathways for both RSV and influenza.

TRAF proteins are key signalling molecules in the host response to immunity.[196] The absence of TRAF6 has been associated with an impaired immune response to RNA virus infections, including a reduction in IL-6 and type 1 interferons, with enhanced viral replication.[197] The genes contributing to this TRAF6 enrichment pathway in our study were increased in both RSV and influenza groups, compared with the bacterial group. These findings would suggest that TRAF6 is an important signalling molecule in viral LRTIs infections. TRAF6 is important in TLR activation in response to infection. Different TLR agonists have been used as vaccine adjuvants due to their ability to trigger an innate and adaptive immune response. TRAF6 may be another potentially useful TLR agonist in future viral vaccines.

2.4.4 Limitations

The certainty of the classification of cases is key to the validity of the reported results. The classification system is limited because the gold standard, culture-based diagnostic tests often do not confidently diagnose cases of LRTI. As expected, culture-confirmed bacterial cases in this study were relatively rare. To increase the number of cases that could be confidently classified as bacterial I used a semi-supervised approach. While this method has likely correctly classified the majority of bacterial cases, it would have been preferable to have a higher number of culture-confirmed bacterial cases in the cohort.

The signatures identified in this chapter work well in the context of results in this particular setting. Like with any newly identified biomarker, the RNA signatures presented in this chapter would need to be validated in an independent cohort. It may be difficult to repeat these results in the same Nepali setting. The PCV10 vaccine was introduced to the Nepali routine immunisation programme during this study and high levels of PCV10 vaccination may have changed the profile of bacterial LRTI infections. Other potential challenges in repeating these results include biases that may have been introduced due to the timing of the research samples relative to treatment. Changes in gene expression can occur due to treatments given or the stage of disease. Samples were taken as close to admission as possible but clinical care was, rightly, not delayed so research samples could be taken and treatments given before obtaining research samples may have influenced the gene expression levels.

The number of co-infections in the cohort was small, and any potential future test should be able to identify co-infections, or at least classify co-infections with

the bacterial group. Co-infections complicate tests which aim to classify cases as bacterial or viral, as I have tried to do in this chapter.

2.4.5 Further work

The RNA signatures derived from these data should be validated in other cohorts. A follow-up cohort study is being conducted at the same site, and enrolling children with a broader range of infectious diagnoses. RNA-seq will be performed on samples from this follow-up cohort. This will provide the opportunity to validate the signatures presented in this chapter. If an RNA signature is validated in this follow-up cohort, the next step would be to test the signature in a non-Nepali setting.

In Chapter 5, the transcripts identified in this chapter are combined with the protein results from Chapters 3 and 4 to see if a combination of proteins and RNA transcripts can better identify bacterial infections.

2.4.6 Conclusions

Differences between gene expression levels can be used to differentiate between LRTIs aetiology. Pathway analyses have highlighted some interesting immune mechanisms involved in LRTIs caused by different pathogens.

RNA signatures which differentiated well between bacterial and viral infection have been presented in this chapter. RNA signatures may be used clinically in the future, but this would require the production of a reliable, high-sensitivity test with results validated across different populations.

Chapter 3

Analysis of the plasma proteome of Nepali children with lower respiratory tract infections

3.1 Introduction

3.1.1 Chapter in context

This chapter describes the analysis of the plasma proteome in children with LRTIs. These samples are from the same cohort of Nepali children that was described in Chapter 2. Plasma proteins were measured in samples taken close to the time of admission in children who presented with signs of pneumonia, and again at convalescence. The setting for this study was a tertiary referral hospital in the Kathmandu Valley.

As previously described in Chapter 1, LRTIs are an important cause of morbidity and mortality in children.[20] Improved diagnostic tests are needed to accurately identify children with potentially severe bacterial LRTIs earlier in their disease course. This would help physicians and public health teams.

In this chapter, I analyse the plasma proteome of children with different causes of LRTI. Measuring many protein changes in different LRTI aetiologies can help to develop diagnostic tests by discovering new biomarkers for LRTIs. Analysing a large set of protein changes can aid in our understanding of disease processes.

This chapter, together with Chapters 2, 4 and 5 summarise the RNA and protein results from a prospective cohort study set in urban Nepal.

In the following sections, I describe how LC-MS can be used to measure plasma proteins and how these results can be used to improve our understanding of LRTIs and develop novel diagnostic tests.

3.1.2 Mass spectrometry proteomics

MS proteomics allows hundreds of proteins to be measured simultaneously. This is useful for biomarker discovery as many biomarkers can be examined in one experiment.[107] MS is a core technology for describing the proteome. Mass spectrometers measure mass-to-charge values and signal intensities to estimate which proteins are present in a sample and in what quantity.

MS proteomics has been used in many experiments to measure protein abundance in different biological states.[107, 109]

LC is combined with MS to increase the sensitivity and precision when separating different components of a sample.

In LC, different components in a sample are separated based on their interactions with a mobile and a stationary phase. As a sample moves through the LC column the different components of the sample move through the column at different rates. Molecules of the same compound will generally move in groups and will be separated into bands within the column. After passing through the LC column, the sample components are ionised to allow MS analysis to take place. Ions are separated based on their mass-to-charge ratio.[198, 199, 200]

In a typical untargeted bottom-up LC-MS experiment, proteins in a clinical sample are initially digested into peptides, using a protease. This peptide solution is then passed through an LC column to separate the peptides. The mass spectrometer then measures the peptide mass-to-charge ratios to calculate the intensities of each peptide present. These intensity measures provide an estimate of protein abundance. Each of the peptides is isolated and fragmented

inside the mass spectrometer and the fragmented mass-to-charge values are used to identify the protein from the amino acid sequence.[198]

Analysing a full proteome using LC-MS produces a large amount of complicated data. To aid the interpretation of these data, traditional proteomic experiments followed a data-dependent acquisition (DDA) approach. In DDA a selection of peptides is taken forward for fragmentation. These fragmented peptides are then matched to a predefined database. This reduced the complexity but at a cost to the depth of data, where peptide results may be missed.[201]

Data-independent acquisition (DIA) was developed as a method to improve the detection of peptides in samples. In DIA the mass spectrometer is configured to cycle through a set of predefined isolation windows, to fragment all the peptides in a sample. This allows for more proteins to be detected. This makes DIA a potentially better method for biomarker discovery. (30, 31)

By fragmenting all the peptides in a sample, DIA produces more complex data than DDA. To deal with this, DIA-NN was developed. DIA-NNis a software suite which uses machine learning neural networks to process these complex data in real-time, to distinguish real signals from noise. DIA-NN also performs some automatic quality control including mass correction and helps to reduce the effort involved in optimising processing workflows.[201]

Several MS proteomics experiments have identified possible biomarkers for different disease states, as outlined in Section 1.4.[113, 101, 111, 112] Of relevance to the results in this chapter, Jackson *et al.*, using a combination of protein-measurement methods, reported that a panel of six proteins could reliably differentiate between bacterial and viral infections in children.[101]

3.1.3 Hypotheses

In this chapter, the plasma proteome results from children with LRTIs of different aetiology were used to test the following hypotheses:

1. Changes in the plasma proteome can be used to distinguish between different LRTI aetiologies.
2. The plasma proteome can be used to identify immune pathways in LRTIs.

3.1.4 Aims and objectives

Linked to hypothesis one – Changes in the plasma proteome can be used to distinguish between different LRTI aetiologies. This study aims to show that differences exist between protein levels depending on the aetiology of LRTI. To investigate this hypothesis, protein results were compared between the following groups:

1. Samples taken during acute viral LRTIs were compared with convalescent samples from the same individuals.
2. Samples taken during acute bacterial LRTIs were compared with convalescent samples from the same individuals.
3. Acute bacterial LRTIs directly were compared with acute viral LRTIs.
    (a) Models were trained and tested, using differentially abundant proteins, to identify protein signatures that can distinguish between bacterial and viral infections.

4. LRTI co-infections will be compared with mono-infections.

Linked to hypothesis two – The plasma proteome can be used to identify immune pathways in LRTIs. This study aims to examine the different immune pathways are involved in different LRTIs. Pathway analysis tools were used to highlight the different biological pathways involved in LRTIs in children. Protein levels in RSV, influenza and *S. pneumoniae* LRTI groups were compared with protein levels in other LRTIs.

3.2 Methods

3.2.1 Clinical study

The clinical methods are outlined in detail in Section 2.2.1, and summarised briefly below.

Study design

A prospective cohort study was undertaken at Patan Hospital, Lalitpur, Nepal. In this hospital setting, children between 2 months and 14 years of age were eligible for enrolment if they were assigned an admission diagnosis of pneumonia, as decided by the clinical team. Children were enrolled between March 2015 and December 2017. Blood samples for RNA and protein analyses, as well as nasopharyngeal samples for molecular diagnostics, were taken on presentation to hospital. Demographics and medical record data were recorded, including the

results of clinical investigations, to assist in correctly classifying the different LRTIs for further analysis. Participants were asked to return six to eight weeks after enrolment to provide a convalescent blood sample when they had recovered.

Study procedures

If the clinical team decided that the admission diagnosis was pneumonia, the research team approached the participant's parent/guardian and sought informed consent. Research samples were taken within the first 48 hours of the admission, and demographics were recorded. See Figure 2.1.

At time of discharge, data on investigations undertaken during the admission were recorded. Six to eight weeks after admission the parent/guardian were asked to return with the participant for repeat research blood samples. These convalescent samples were used as control samples.

Research samples

At enrolment a nasopharyngeal sample and a blood sample were taken from each participant. A 3 ml blood sample was taken and 1 ml was transferred to an RNA-stabilising tube (Tempus Blood RNA Tube) and 2 ml transferred to a heparinised centrifuge tube. Research samples were transferred to the laboratory within two hours.

At convalescence the same research blood samples were taken.

3.2.2 Laboratory methods

Nasopharyngeal samples

An aliquot of the transport media from each nasopharyngeal sample was transferred to an agar plate for pneumococcal culture. If *S. pneumoniae* was found, a colony underwent serotype testing, as outlined in Section 2.2.2.

Another aliquot of transport media from each nasopharyngeal sample was frozen at -80°C and later transferred to Micropathology Ltd., University of Warwick Science Park, UK. At this UK laboratory, nucleic acid was extracted and analysed using the NxTAG™ Respiratory Pathogen Panel. This panel detects 16 viral and 3 bacterial targets, see Figure 6.1.

Blood samples

The processing of the RNA-stabilising tubes was discussed in Section 2.2.2. Plasma samples obtained from the heparinised centrifuge tube were used for the analyses presented in this chapter. The processing of these samples is discussed below.

Within four hours of blood sampling, the fresh whole blood in the centrifuge tube was separated into plasma, PBMCs and a mixture of red cells and polymorphonuclear cells. The use of the PBMCs, red cells and polymorphonuclear cells is outside of the scope of this thesis and will not be discussed further.

The plasma was frozen at -80°C and initially stored at Patan Hospital. Later the plasma was transferred to the laboratories at the University of Oxford,

United Kingdom. At the Centre for Clinical Vaccinology and Tropical Medicine (CCVTM), University of Oxford, 25 µl of plasma from each participant were used to measure cytokine levels using the Meso Scale Discovery platform, as outlined in Chapter 4. At the Discovery Proteomics Facility of the Target Discovery Institute (TDI), University of Oxford, 3 µl of plasma from each participant were used to measure the plasma proteome using liquid chromatography–mass spectrometry. These methods are discussed in Section 3.2.4.

3.2.3 Classification of cases

Participants were classified into different LRTI groupings based on the likely cause of their illness. The results of routine investigations and the NxTAG respiratory PCR panel were used for classification, see Figure 2.3.

The most important groups for the analyses in this chapter are the definite viral and bacterial groups, these are described here. For a full description of case classification, see Section 2.2.3.

Participants were classified as definite bacterial if they had a culture-positive pathogenic bacteria isolated from a normally sterile site (blood or pleural fluid). The probable bacterial group included participants with high inflammatory markers (CRP >60mg/L) in the absence of a culture-positive bacteria and the absence of an LRTI-associated virus.

Participants were classified as definite viral if they had an LRTI-associated virus isolated in their nasopharynx, with low inflammatory markers (CRP <60mg/L and neutrophils $\leq$ 12 x $10^9/L$, and with no pathogenic bacteria isolated from a

sterile site. LRTI-associated viruses were RSV, influenza, parainfluenza viruses and human metapneumovirus.

Bacterial-viral co-infections were classified as those who met the criteria for the definite bacterial group, and who also had an LRTI-associated virus isolated from their nasopharynx. Secondary groupings were used to look at individual pathogens and severity. Participants who had RSV or influenza detected were grouped together. Participants in the pneumococcal group had Streptococcus pneumoniae cultured from a normally sterile site or had pneumococcal serotypes 1 or 5 cultured from their nasopharynx. Pneumococcal serotypes 1 and 5 being the most common serotypes associated with invasive bacterial disease.[56, 57]

Where sample volume allowed, samples from the same participants who underwent differential gene expression analysis using RNA-seq were selected for MS proteomics, see Section 2.2.4 for more details on sample selection. This was to allow for comparisons between the RNA and protein results, see Chapter 5 for details of comparisons across platforms.

3.2.4 Mass spectrometry

The plasma samples from this study were sent to the Discovery Proteomics Facility of the Target Discovery Institute, University of Oxford to undergo analysis using LC-MS. The main steps in this process are outlined below.

Sample processing

Plasma samples were defrosted and 3 microlitre aliquots were transferred to 96 well plates. Plasma samples were trypsin digested. Suspension trapping digestion was also carried out to aid processing as digests were sub-optimal. The samples were not depleted for abundant plasma proteins as the data processing methods the laboratory used could account for highly abundant proteins.

Liquid chromatography-mass spectrometry

Using DIA mode, the digested samples were subjected to LC-MS on the timsTOF fleX Pro instrument for data acquisition. Two pooled samples were made, one was a pool of all the acute samples and the other was a pool of convalescent samples. These samples were used as a quality control check. A pooled sample was run every 12 samples to identify, and account for, instrument drift. This pooled sample was also fractionated, and each fractioned sample was separately analysed using LC-MS. This step results in more proteins being identified in the fractionated samples. The extra information provided by identifying more proteins in the pooled sample can help to identify ambiguous proteins in the other plasma samples. The MS spectra of protein intensities produced from the mass spectrometer were searched against a spectral library using the DIA-NN software which deconvoluted the data. This allowed for peptide identification. Following this, a matrix of the relative quantities of different protein groups was generated. The units of protein abundance are ion counts.

3.2.5 Protein matrix results - pre-processing and quality control

The protein matrix results were received from the team at the Discovery Proteomics Facility. The following sections outline the steps carried out to prepare these data for analysis. Appendix C provides a more detailed version of the steps carried out in R. These analyses were carried out using R version 4.3.1.

The protein matrix results contained protein IDs as labels. These IDs were mapped to protein names for ease of interpretation of results, using the UniProt database.[202]

Filtering

Filtering is required to remove proteins that are undetected in a high proportion of samples across the cohort. Including these proteins could introduce a high level of bias as the majority of the results for these proteins would be estimated. To ensure that proteins detectable in one biological group of interest are not excluded, the results were divided into the following filtering groups: acute bacterial (definite bacterial and probable bacterial groups combined), acute viral (definite viral group), acute unknown (all other acute samples) and convalescent groups. The unknown group was not used for filtering as this is likely to be a heterogeneous group of samples. Samples were also grouped by pathogen groups of interest: RSV, influenza and pneumococcus. This gives a total of six groups for filtering, proteins were retained if they were detected in at least 50% of samples in at least one of the filtering groups.

Remove low-quality results

Sample quality was examined by measuring the number of protein results present for each sample. Samples with low numbers of protein results may represent problems during the LC-MS processing steps. By plotting the number of protein results present in each sample a threshold for excluding low-quality samples was chosen. Samples with protein results present in less than 70% of the mean number of results for the cohort were excluded.

Proteins which were marked as contaminants against the original spectral library search in DIA-NN were assessed for removal.

Normalisation

Normalisation of MS proteomic data is important to correct for technical variation, natural biological variation and to allow for suitable quantitative comparisons between samples. The NormalyzerDE online tool was used to compare different normalisation methods.[203] After inputting the protein matrix data, NormalyzerDE provides a report which helps to choose the most appropriate normalisation method for this study. Variance stabilising normalisation (VSN) was chosen as it was one of the best-performing methods in the NormalyzerDE report, and VSN has been used successfully in previous proteomics studies.[204, 205] VSN brings values to the same scale by first performing a transformation to remove variance caused by experimental factors, and then performing a log2 transformation. [204] Normalisation was performed before imputation. The steps were performed in this order as the randomness introduced by imputing values prior

to normalisation has the potential to mask true biological trends, or mask biases that could be accounted for with normalisation.[205]

Impute missing values

Missing values are common in proteomic datasets. Missing values may be missing because the peptide was present but not detected, or the peptide was present but below the limit of detection for the mass spectrometer or because the peptide is truly absent. It is impossible to know for sure the reason for a missing value so a suitable imputation method must be applied.[205] If a protein result is present for most samples in the cohort, and missing in a small number of samples then it is more likely that this peptide was present, but not detected by the mass spectrometer. The k-nearest neighbour approach was used to impute missing values for these proteins with results in >60% of samples.

If a protein result is absent from most samples, then it is more likely that this protein is missing because it was truly absent, or below the limit of detection of the mass spectrometer. A down-shifted normal distribution was used to impute these proteins with absent results in most samples (proteins present in <60% of samples). The k-nearest neighbour approach is not suitable in the proteins with many absent results as the neighbour proteins cannot be accurately located.[110]

Quality control plots

PC biplots and box plots were examined looking for outlying samples and potential bias introduced by technical factors or other possible confounding

factors.

Batch effect correction

A batch effect due to differing results on the different plates used in the experiment was identified. In proteomic datasets two types of batch effects are encountered, continuous and discrete. The identified plate issue is a discrete batch effect, mean centering methods work well to correct for this when biological factors are not completely confounded by technical batch issues.[206] ComBat is an algorithm to adjust for batch effects using an advanced modification of mean centering.[207]

3.2.6 Differential protein abundance analyses

Differential protein abundance was analysed using the limma package in R.[150] A linear mixed effects model was created. As well as experimental plate, age was also found to be a confounding variable due to age differences across the LRTI classification groups. Age, plate and a combined variable of LRTI classification and time point were the fixed effects included. Several of the samples were paired acute and convalescent samples from the same individual. The contribution of this partial pairing was measured using the average estimated inter-duplicate correlation. Participant identification was included as a random effect in the model, on the basis of this inter-duplicate correlation.[208]

The lmFit function was used to fit a linear model for each protein. A contrast matrix was made to answer the clinical questions of interest. Table 3.1 shows the full contrast matrix. The contrast matrix for each comparison was applied to

the linear model for each protein. Finally, empirical Bayes shrinkage was applied to the standard errors, and adjusted p-values, using BH method and log2 fold changes were calculated for each protein in the different contrasts.

Differential abundance tables were produced to allow comparisons between the different bacterial and viral pneumonia groups including examining individual pathogen groups, severity, and co-infections. Minimum thresholds for differential abundance were adjusted p-value of <0.05, using the BH false discovery rate adjustment, and log-2-fold-change (FC) >0.5 were used. These cut-offs resulted in high numbers of significant proteins in comparisons between acute and convalescent samples, and more stringent thresholds of adjusted p-value <0.001 and log2FC >1 were used in these comparisons.

3.2.7 Pathway analysis

After obtaining lists of differentially abundant proteins in when comparing acute samples from pathogen groups of interest and convalescent samples, pathway analyses were undertaken to aid in the interpretation of the biological significance of these protein groups. Comparisons were also made between protein abundance in acute RSV and other viral samples to look for proteins that were specific to RSV infection. Similarly, protein abundance levels in the acute influenza group were compared with levels in other viral groups. Protein abundance levels in the acute pneumococcal group were compared with levels in acute samples from participants with other bacterial infections. ORA is a gene set analysis method which identifies pathways where the differentially abundant proteins occur more than would be expected by chance. (47) The clusterProfiler package in R was

Table 3.1: Lower respiratory tract infection groups of interest and comparisons used to answer the chapter hypotheses in this MS proteomics experiment.

Hypothesis	**LRTI Group of Interest**	**Contrast**
Changes in the plasma proteome can be used to distinguish between different LRTI aetiologies	Acute definite viral	Convalescent samples from the same participants (viral convalescent)
	Acute definite bacterial	Convalescent samples from the same participants (bacterial convalescent)
	Acute definite bacterial	Acute definite viral
	Acute co-infections	Acute definite viral
	Acute co-infections	Acute definite bacterial
The plasma proteome can be used to identify important immune pathways in LRTIs	Acute pneumococcal	Convalescent samples from the same participants (pneumococcal convalescent)
	Acute pneumococcal	Acute other non-pneumococcal bacterial samples
	Acute influenza	Convalescent samples from the same participants (influenza convalescent)
	Acute influenza	Acute non-influenza viral samples
	Acute RSV	Convalescent samples from the same participants (RSV convalescent)
	Acute RSV	Acute non-RSV viral samples

used for ORA. ORA was performed using pathways in The GO Gene Ontology Resource (GO pathways) as the reference database.[209] Enrichment analysis was carried out using the GO-BP ontology, with BH-adjusted p-value threshold of 0.05, and q-value <0.1.

3.2.8 Protein signature to differentiate between bacterial and viral lower respiratory tract infections

The following steps were carried out to look for a small number of proteins which could be used as a potential diagnostic signature, to accurately differentiate between bacterial and viral LRTIs.

A dataset containing only the bacterial and viral cases based on results from the semi-supervised approach described in Section 2.3.3 was created. Data were randomly split into train and test datasets, using a 70:30 split between train and test.

Differential abundance analysis was performed on the training dataset. The same steps were followed as per Section 3.2.6.

Proteins, and confounding factors, were ranked in order of importance using the partial least squares (PLS) approach. Using PLS the variable importance measure is based on the weighted sums of the absolute regression coefficients.[210] Proteins were also ranked using the learning vector quantisation approach, a neural network algorithm. This approach gave similar results to the PLS approach.

Using the caret function in R a correlation matrix was created and highly correlated proteins were identified, Pearson's correlation coefficient >0.75. Where

two or more proteins were highly correlated the lower-ranked protein(s) were removed.

Following removal of highly correlated proteins, recursive feature elimination was performed using 10-fold cross-validation, to remove potentially redundant features.[211] Following exclusion of redundant proteins, different protein signatures for differentiating bacterial from viral LRTIs were identified.

Performance statistics were calculated for the different protein signatures. Sensitivity, specificity, ROC AUC, positive and negative likelihood ratios were calculated for the training set initially. As a test which has a high sensitivity to detect bacterial infections would be more useful clinically, I restricted the sensitivity range to 90% to 100% and re-calculated the AUC. This pAUC showed how well the model was able to predict a correct result when the required true positive rate was high.[168]

The performance of the selected protein signatures was then evaluated in the test dataset.

3.3 Results

3.3.1 Description of cohort

MS proteomic results were available for 335 plasma samples, prior to QC. Following filtering for low-quality samples, 24 samples were removed. Two further samples were removed following review of QC plots for outliers, 309 samples included for further analyses, see Figure 3.1.

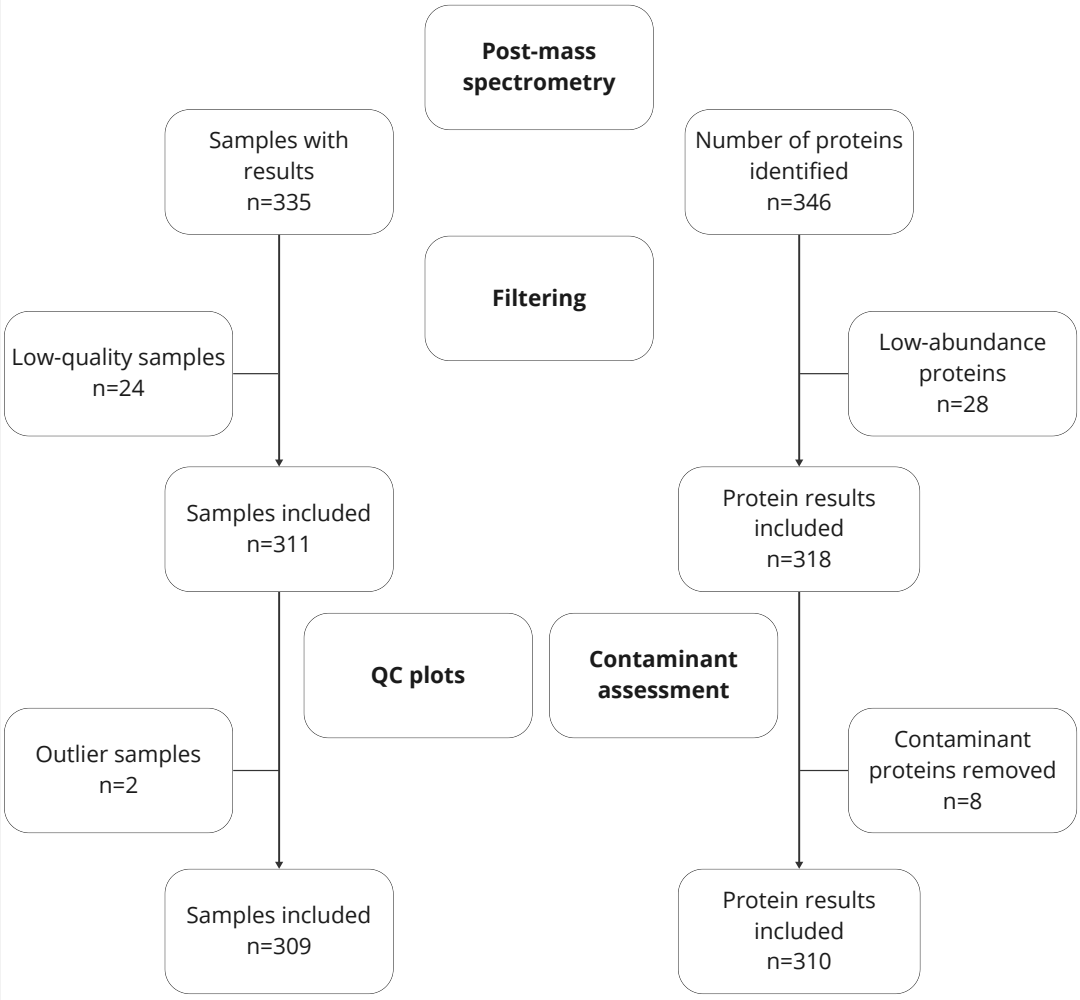


Figure 3.1: Numbers of samples and proteins retained after each quality control step. The final numbers of samples and protein results used in the analyses are shown.

Table 3.2: Demographic and key clinical/laboratory data for the participants with protein results, following exclusion of samples that failed QC, grouped by LRTI classification.

	Definite viral	Viral syndrome with high inflammatory markers	Unknown	Co-infection (bacterial-viral)	Probable bacterial	Definite bacterial	Convalescent	Total
Number of cases	71	18	53	4	69	16	78	309
Age (years)	1.52	2.04	2.48	1.20	5.16	5.87	2.83	3.08
Female (%)	35%	50%	36%	50%	43%	25%	41%	39%
Duration of hospital stay	6	7	4	18	7	6	NA	6
CRP (mg/l)	7.1	97.5	10.0	128.9	158.3	151.0	NA	29.6
WCC x10^9/L	10.7	13.6	13.6	8.8	19.5	19.5	NA	13.8
Neutrophils x 10^9/L	4.9	10.5	7.7	3.0	15.6	14.5	NA	8.4

Samples from 231 participants taken near the time of admission were included, and 78 samples from obtained during convalescence when participants were well. Table 3.2 shows the demographics and key clinical data for the participants with protein results available, grouped by LRTI classification. Of note, the definite viral group consisted of younger children than the bacterial groups and 61% of participants were male. The small number of participants in the co-infection group made meaningful analysis in this group difficult.

3.3.2 Protein results included post-quality control

The protein group matrix results from the Discovery Proteomics Facility included results for 346 proteins, prior to filtering and QC. After retaining proteins that were detected in at least 50% of samples in at least one of the LRTI filtering groups, 318 proteins were retained. Following removal of proteins highlighted as likely contaminants during the DIA process, 310 proteins were retained, see Figure 3.1. Following these QC steps, 8.8% of values across the entire cohort were missing values to be imputed.

3.3.3 Quality control

PC biplots showed that the samples separated into two groups. Figure 3.2a shows that these groupings are not explained by the study time point. Other biplots showed that this separation was not explained by LRTI classification, age or sex.

A PC biplot coloured by the different plates used in the MS experiment showed that the results on Plate 5 grouped separately from the samples on Plates 4a and Plate 4c, Figure 3.2b. There was no known biological reason for the samples on Plate 5 to be different to the rest of the cohort. The separation into two groups on PC plots was likely due to a technical reason during the MS experiment.

The Eigencor plot shows that the mass spectrometry experiment plate has the highest correlation with PC1, Figure 3.3. This helps to confirm that experimental plate is causing a batch effect.

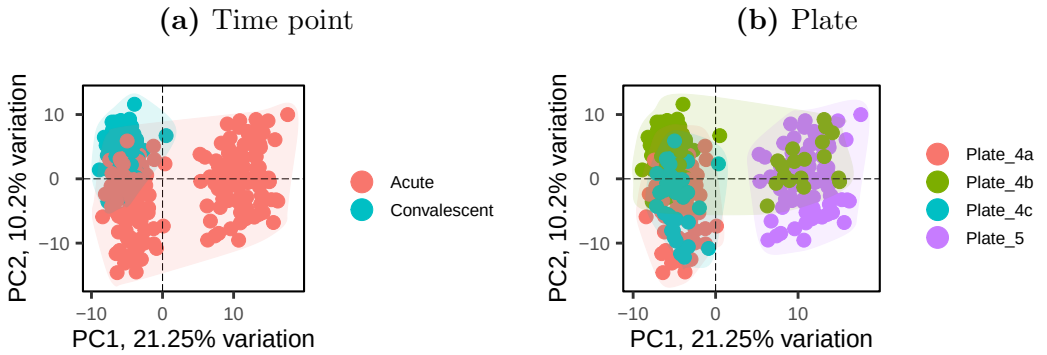


Figure 3.2: PCA analysis of MS proteomic data. Biplot comparing PC1 and PC2 coloured by time point, panel a, and experimental plate, panel b, pre-batch correction.

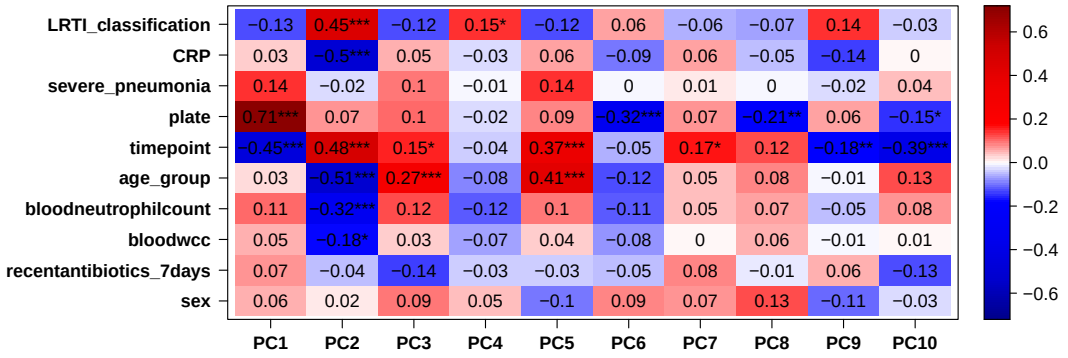


Figure 3.3: Eigencor plot, correlates principal components to metadata variables of interest and tests the significance of these, pre-batch correction. Significance tests adjusted for multiple testing using false discovery rate, Benjamini-Hochberg method. Significance test cut-offs used: * <0.05, ** <0.01, *** <0.001.

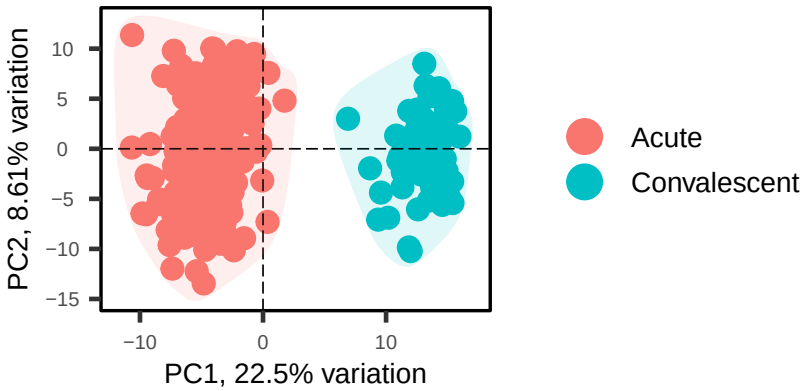


Figure 3.4: PCA of MS proteomic results. Biplot comparing PC1 and PC2 coloured by time point, acute versus convalescent, post-batch correction.

3.3.4 Principal component analysis

PC plots were repeated following batch correction, to ensure an acceptable correction had been achieved. The acute and convalescent samples cluster in two distinct groups post-batch correction, Figure 3.4. The Eigencor plot, Figure 3.5, shows that study time point is the main contributor to PC1. Age is significantly negatively correlated with PC2 and is known to be different between LRTI classification groups. Age was adjusted for when undertaking differential abundance analyses. Further PCA did not show any other obvious confounding variables that needed to be adjusted for in the model.

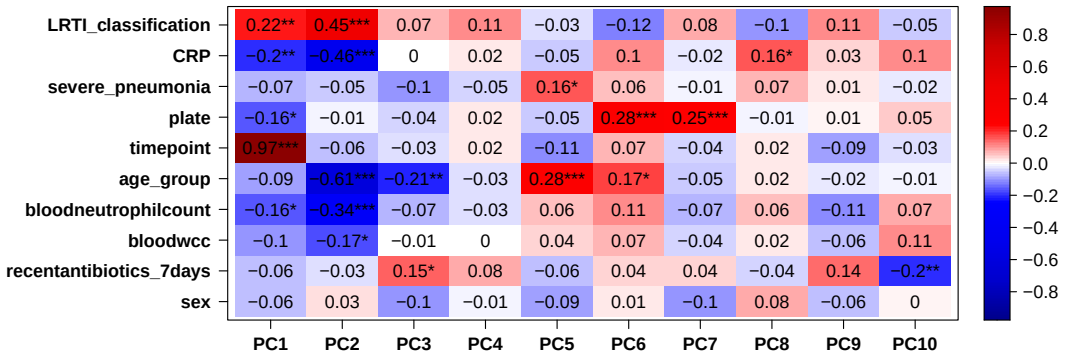


Figure 3.5: PCA of MS proteomic results, Eigencor plot, correlating principal components to metadata variables of interest and tests the significance of these, post-batch correction. Significance tests adjusted for multiple testing using false discovery rate, Benjamini-Hochberg method. Significance test cut-offs used: * <0.05, ** <0.01, *** <0.001.

3.3.5 Comparing bacterial with viral lower respiratory tract infections

To compare acute bacterial and acute viral LRTIs, PCA was repeated with acute LRTI samples only to look for differences between different LRTI groups. The biplot in Figure 3.6 shows that while there is a degree of overlap between bacterial and viral cases, there is some differentiation between bacterial and viral cases across PC1.

Protein signature to differentiate between bacterial and viral lower respiratory tract infections

Using the classification of cases based on results from the semi-supervised approach described in Section 2.3.3, cases were selected to use in training and

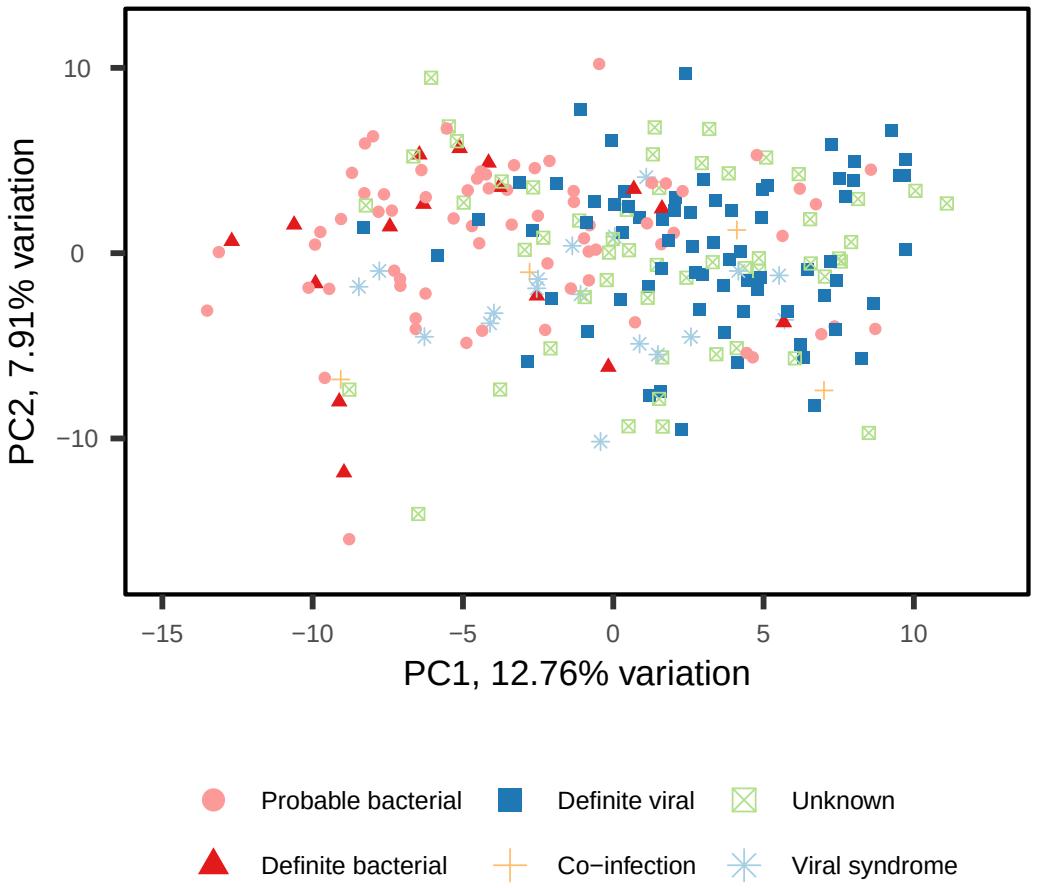


Figure 3.6: PCA of MS proteomic results. Principal component biplot comparing PC1 and PC2 coloured by LRTI classification, post-batch correction.

testing models to differentiate between bacterial and viral infections. The 59 bacterial cases included 16 cases from the definite bacterial group, 39 cases from the probable bacterial group and 4 co-infections. The co-infections were included with the bacterial group as they had a positive culture for bacteria from a normally sterile site, and clinically it would be more useful to identify these cases as bacterial. Differential protein abundance was measured between acute co-infections and the acute bacterial group, and separately differential protein abundance was measured between acute co-infections and the viral group; no significant differentially abundant proteins were identified when comparing co-infections with either bacterial or viral LRTIs, see Appendix C.

To look for a set of useful proteins to include in the model, the data were split into train and test datasets with a 70:30 split. Differential abundance analysis was then performed on the training dataset.

Differential abundance analysis – training dataset

Using an adjusted p-value of <0.05 and a minimum log2 fold change of 0.5, 13 proteins were differentially abundant in the training dataset, when comparing bacterial with viral LRTIs. Figure 3.7 and Table 3.3 shows that these proteins were all relatively more abundant in the bacterial group. CRP was one of the proteins found to be differentially abundant; as CRP was used in the initial classification system, CRP was excluded from further analyses.

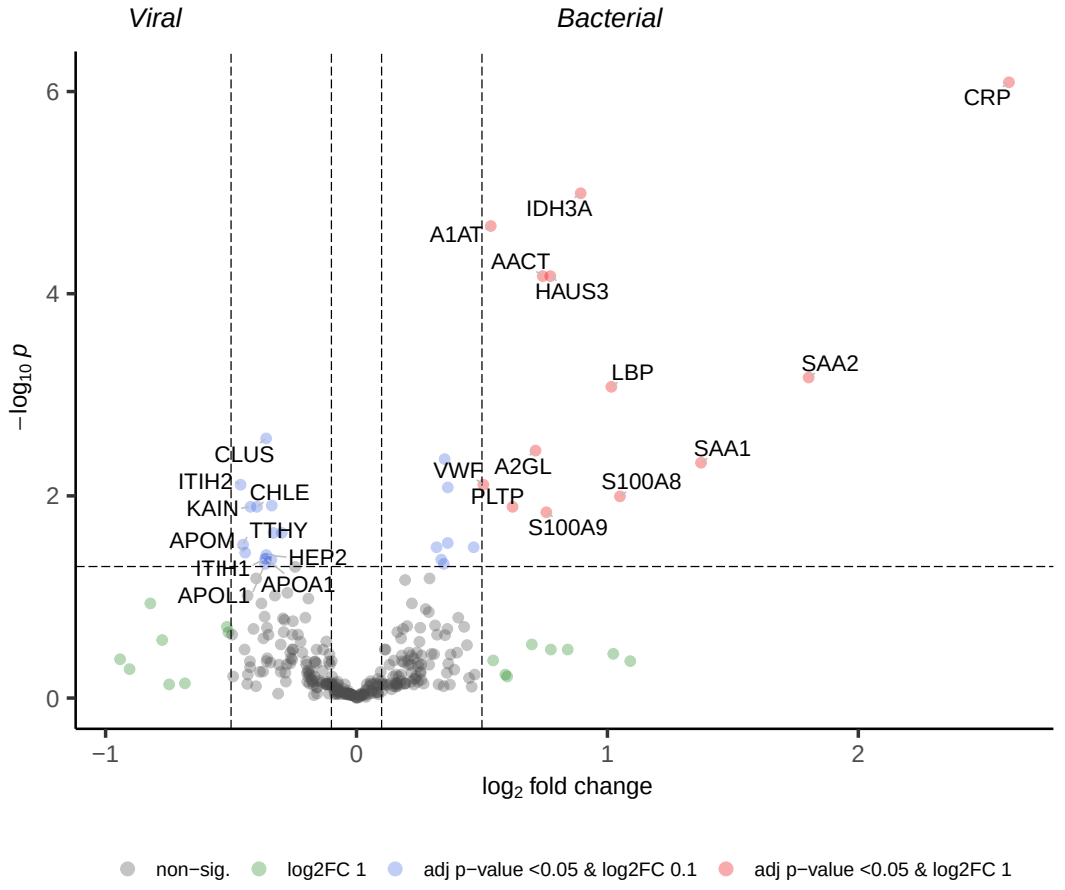


Figure 3.7: Volcano plot showing differentially abundant proteins between participants in the acute bacterial group and the acute viral LRTI group. Participants in the training dataset only included. The labelled proteins are the top ten up- and down-regulated genes, ranked by Benjamini-Hochberg adjusted p-values.

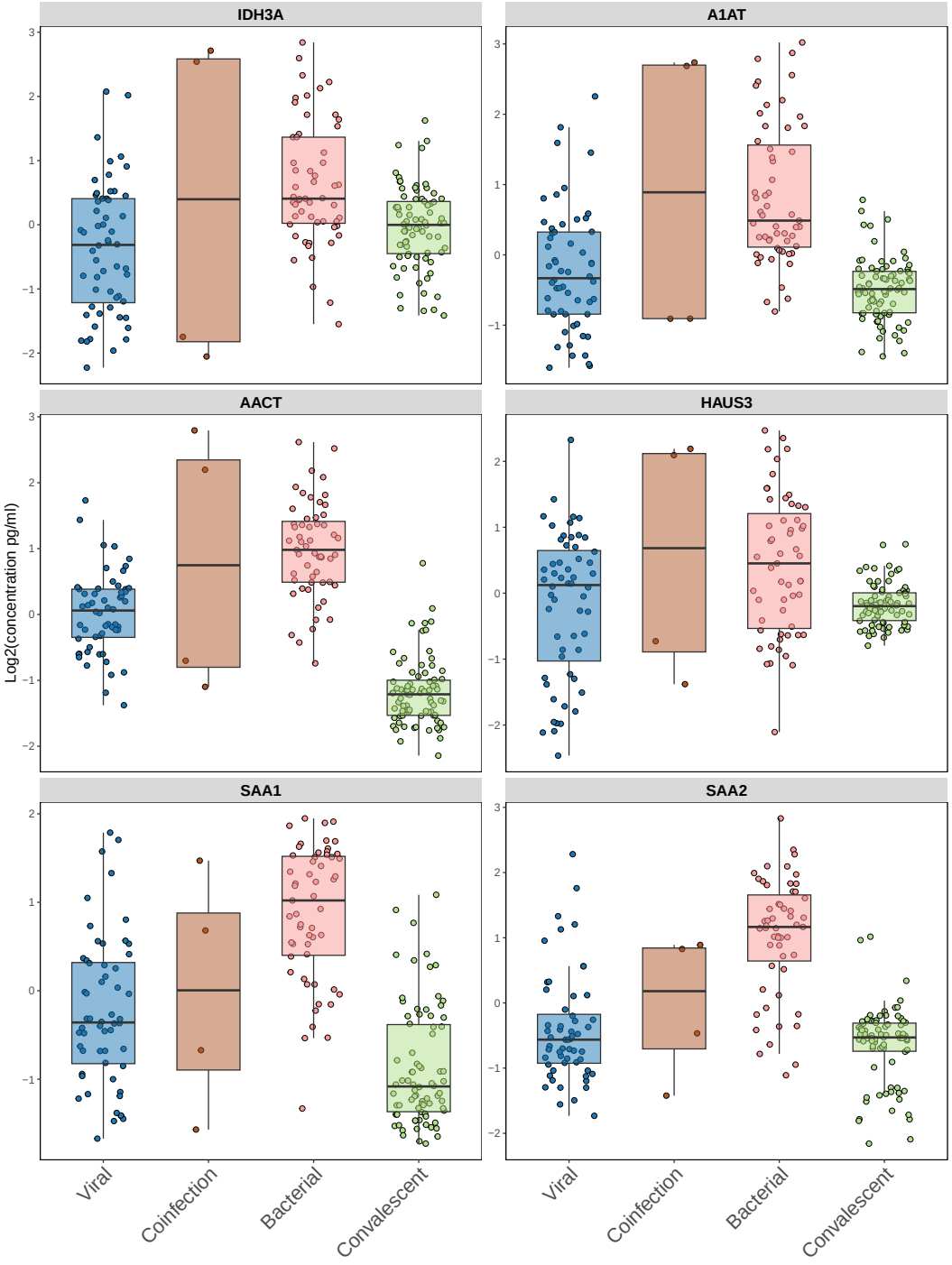


Figure 3.8: Box plots of the first 6 of the 12 differentially abundant proteins when acute bacterial LRTIs are compared with acute viral LRTIs. Samples are grouped by LRTI classification: acute viral samples, acute bacterial-viral co-infections, all other acute bacterial samples and all convalescent samples

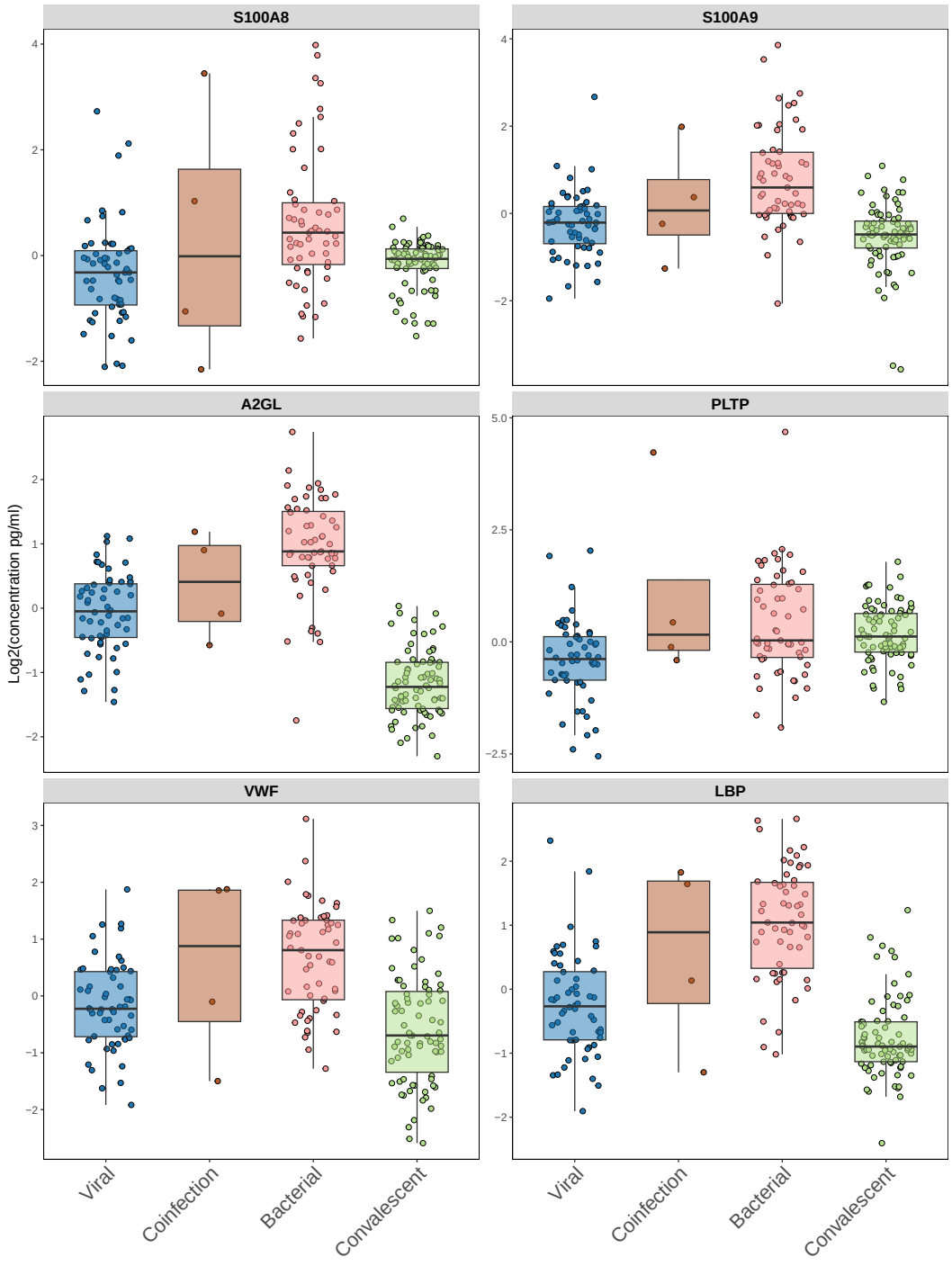


Figure 3.9: Box plots of the second 6 of the 12 differentially abundant proteins when acute bacterial LRTIs are compared with acute viral LRTIs. Samples are grouped by LRTI classification: acute viral samples, acute bacterial-viral co-infections, all other acute bacterial samples and all convalescent samples

Table 3.3: Differentially abundant proteins in the training dataset when acute definite bacterial samples are compared with acute definite viral LRTIs, adjusted p-value <0.05, log2FC >0.5.

Protein	log-2-fold-change	t-statistic	p-value	Adjusted p-value	Protein names
IDH3A	0.89	5.91	6.54355E-08	0.00001	Isocitrate dehydrogenase [NAD] subunit alpha, mitochondrial
A1AT	0.54	5.64	2.07035E-07	0.00002	Alpha-1-antitrypsin
HAUS3	0.77	5.26	1.04647E-06	0.00007	HAUS augmin-like complex subunit 3
AACT	0.74	5.25	1.08084E-06	0.00007	Alpha-1-antichymotrypsin
SAA2	1.80	4.62	0.00001	0.00067	Serum amyloid A-2 protein
LBP	1.02	4.53	0.00002	0.00083	Lipopolysaccharide-binding protein
A2GL	0.71	4.07	0.00010	0.00357	Leucine-rich alpha-2-glycoprotein
SAA1	1.37	3.94	0.00017	0.00470	Serum amyloid A-1 protein
VWF	0.51	3.74	0.00033	0.00775	von Willebrand factor
S100A8	1.05	3.62	0.00049	0.01011	S100 calcium-binding protein A8
PLTP	0.62	3.49	0.00076	0.01283	Phospholipid transfer protein
S100A9	0.76	3.43	0.00094	0.01453	S100 calcium-binding protein A9

Differentially abundant proteins

Feature selection

When all 12 proteins were included in the model, a high correlation was found between eight different protein combinations, see Table 3.4. The proteins were then ranked in order of importance, using the weighted sums of the absolute regression coefficients, see Figure 3.10. The weights calculated are a function of the reduction of the sum of the squares across the PLS components and are calculated separately for each outcome.[211]. Three proteins with high correlation with other proteins, and low importance rankings were removed at this point, A1AT, HAUS3 and SAA1.

The remaining nine protein results were used for recursive feature elimination using 10-fold cross-validation. The highest accuracy score was observed using

Table 3.4: Correlation matrix showing Pearson's correlation coefficient (r) between the differentially abundant proteins considered for the bacterial versus viral protein signature. Proteins with r >0.75 are highlighted in bold.

	IDH3A	A1AT	HAUS3	AACT	SAA2	LBP	A2GL	SAA1	VWF	S10A8	PLTP	S10A9
IDH3A	1	**0.866867**	**0.821586**	0.7219	0.608564	0.646375	0.318651	0.664076	0.508101	0.599133	0.277865	0.256589
A1AT	**0.866867**	1	**0.828149**	**0.799736**	0.666388	**0.773678**	0.40612	0.690479	0.611119	0.627023	0.319678	0.317554
HAUS3	**0.821586**	**0.828149**	1	0.550173	0.49627	0.545358	0.054088	0.568559	0.435764	0.471979	0.217195	-0.00726
AACT	0.7219	**0.799736**	0.550173	1	0.715734	0.749984	0.683872	0.679931	0.631632	0.645936	0.407447	0.594717
SAA2	0.608564	0.666388	0.49627	0.715734	1	**0.809092**	0.649985	**0.899576**	0.678487	0.401281	0.411741	0.365702
LBP	0.646375	**0.773678**	0.545358	0.749984	**0.809092**	1	0.629503	**0.788882**	0.704677	0.432247	0.497866	0.394678
A2GL	0.318651	0.40612	0.054088	0.683872	0.649985	0.629503	1	0.615729	0.54581	0.341493	0.286492	0.646763
SAA1	0.664076	0.690479	0.568559	0.679931	**0.899576**	**0.788882**	0.615729	1	0.633036	0.3774	0.376509	0.30404
VWF	0.508101	0.611119	0.435764	0.631632	0.678487	0.704677	0.54581	0.633036	1	0.401168	0.436443	0.341316
S10A8	0.599133	0.627023	0.471979	0.645936	0.401281	0.432247	0.341493	0.3774	0.401168	1	0.244593	0.612501
PLTP	0.277865	0.319678	0.217195	0.407447	0.411741	0.497866	0.286492	0.376509	0.436443	0.244593	1	0.246237
S10A9	0.256589	0.317554	-0.00726	0.594717	0.365702	0.394678	0.646763	0.30404	0.341316	0.612501	0.246237	1

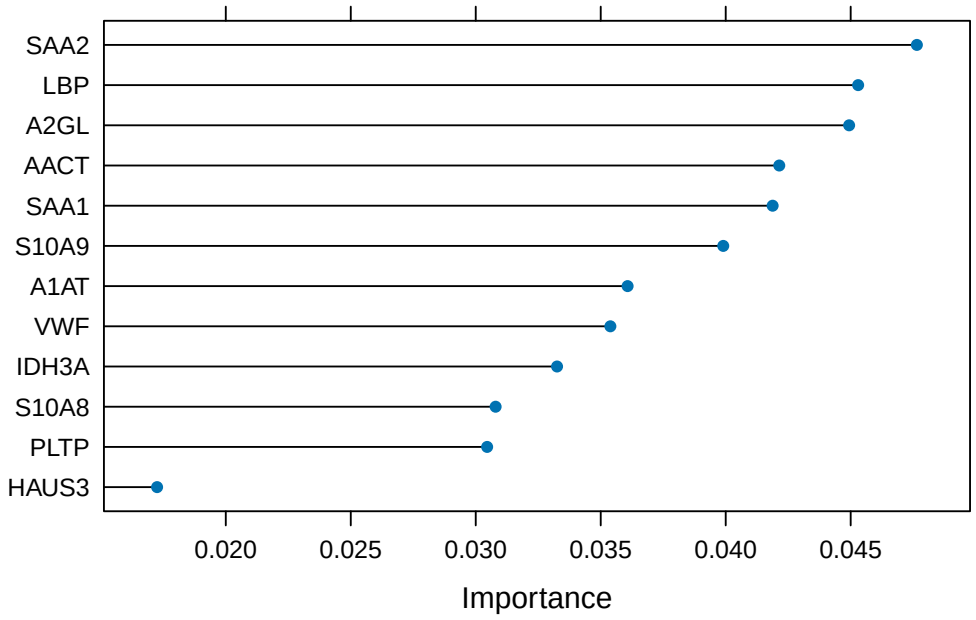


Figure 3.10: Importance of all 12 differentially abundant proteins considered for signature in differentiating between bacterial and viral LRTIs, ranked using partial least squares method. Weighted sums of the absolute regression coefficients presented. Model using partial least square approach used.

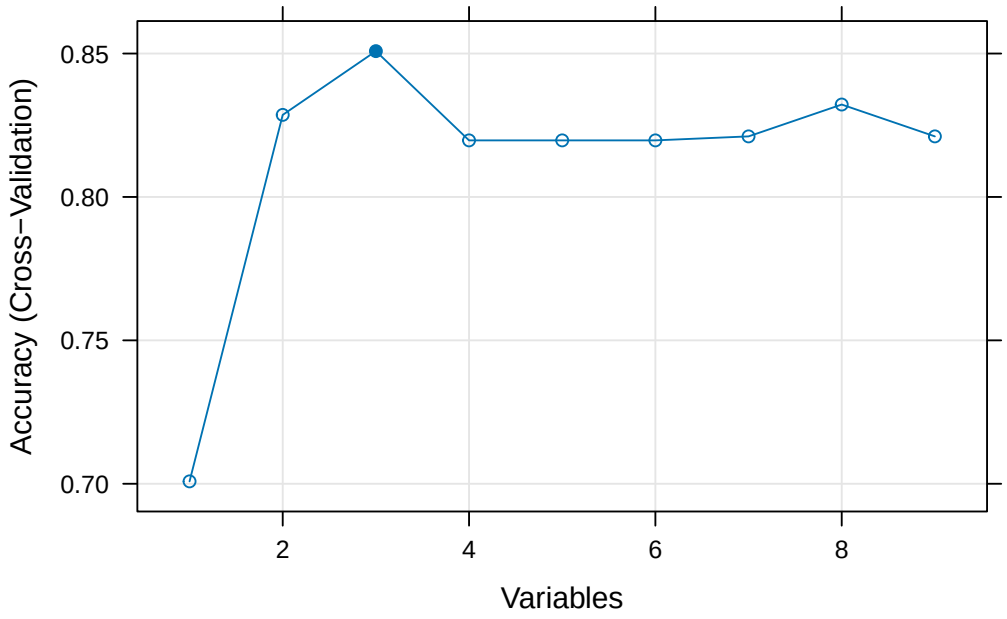


Figure 3.11: Using recursive feature elimination to determine the accuracy of different models and the ideal number of variables to include, considering nine differentially abundant proteins following removal of highly correlated proteins.

three proteins, A2GL, SAA2 and LBP, see Figure 3.10.

The serum amyloid A-2 (SAA2) and LBP results had a Pearson's correlation coefficient above the 0.75 threshold, r=0.81. One from this protein pair was not initially excluded as these were the top two proteins in order of importance. As LBP had lower importance than SAA2 it was excluded and the model was re-trained. Following exclusion of LBP, and repeating recursive feature selection, a seven-protein model had the highest accuracy, see Figure 3.12. The proteins in the seven-protein model were leucine-rich alpha-2-glycoprotein (A2GL), SAA2, S100A9, alpha-1-antichymotrypsin (AACT), phospholipid transfer protein (PLTP), isocitrate dehydrogenase [NAD] subunit alpha, mitochondrial (IDH3A) and von Willebrand

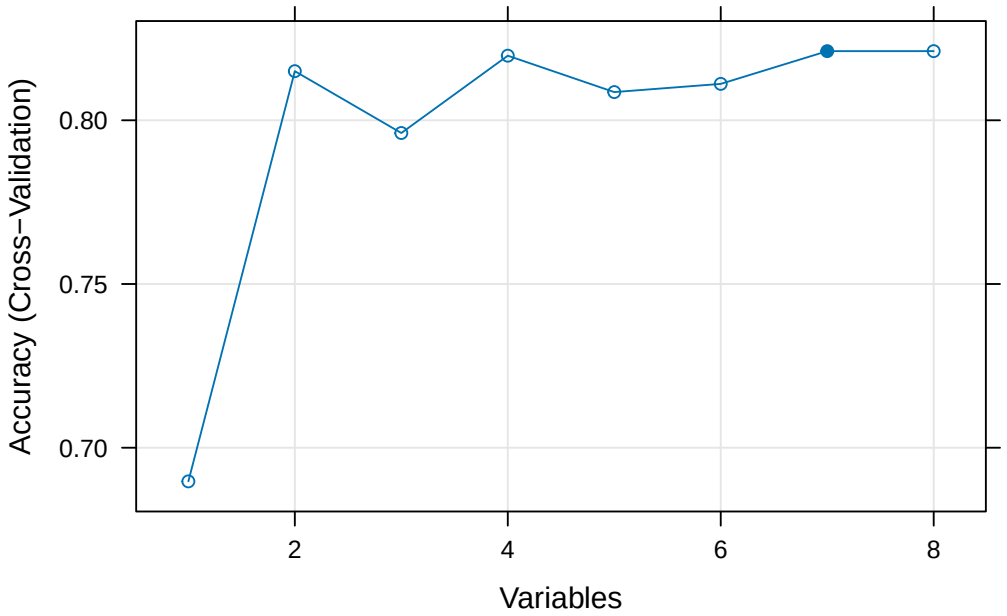


Figure 3.12: Using recursive feature elimination to determine the accuracy of different models and the ideal number of variables to include, considering eight differentially abundant proteins, following the removal of LBP.

factor (VWF). Figure 3.12 also shows that a two-protein signature is almost as accurate as the seven-protein model. This two-protein model includes A2GL and A2GL.

Model performance - Training data

Performance statistics were calculated for the two-protein model, the three-protein model, the seven-protein model and a model containing the 12 differentially abundant proteins. Bacterial LRTIs were classified as the positive class for the protein signature. In the training data, the accuracy was similar across the three models tested. The sensitivity ranged from 79% to 81%, with specificity

Table 3.5: Protein signature performance statistics in the training data with three different protein models compared. The proteins included in each signature are listed. Values for 95% confidence intervals (CI) are provided where appropriate

Performance metric	**Twelve-protein signature**		**Seven-protein signature**		**Three-protein signature**		**Two-protein signature**	
Proteins included	A2GL, A1AT, LBP, SAA1, SAA2, S100A8, S100A9, AACT, PLTP, IDH3A, VWF, HAUS3		A2GL, SAA2, S10A9, AACT, PLTP, IDH3A, VWF		A2GL, SAA2, LBP		A2GL, SAA2	
	Reference - bacterial	**Reference - viral**	**Reference - bacterial**	**Reference - viral**	**Reference - bacterial**	**Reference - viral**	**Reference - bacterial**	**Reference - viral**
Predicted bacterial	35	5	34	5	34	6	34	8
Predicted viral	8	44	9	44	9	43	9	41
Accuracy (95% CI)	0.86 (0.77, 0.93)		0.85 (0.76, 0.91)		0.84 (0.75, 0.91)		0.82 (0.72, 0.89)	
Sensitivity	0.81 (0.67, 0.92)		0.79 (0.64, 0.90)		0.79 (0.64, 0.9)		0.79 (0.64, 0.9)	
Specificity	0.90 (0.78, 0.97)		0.90 (0.78, 0.97)		0.88 (0.75, 0.95)		0.84 (0.7, 0.93)	
Positive predictive value	0.88 (0.75, 0.94)		0.87 (0.75, 0.94)		0.85 (0.73, 0.92)		0.81 (0.69, 0.89)	
Negative predictive value	0.85 (0.75, 0.91)		0.83 (0.73, 0.9)		0.83 (0.73, 0.9)		0.82 (0.72, 0.89)	

ranging from 88% to 90%. The performance statistics for each of the three models are shown in Table 3.5. The ROC curves for each model are shown in Figure 3.13. The AUC ranged from 88.6% to 89.9% across the three models. The pAUC, when sensitivity is restricted to a range between 90% and 100%, was 73% for the twelve-protein model, 76.6% for the seven-protein model and 74.7% for the three-protein model.

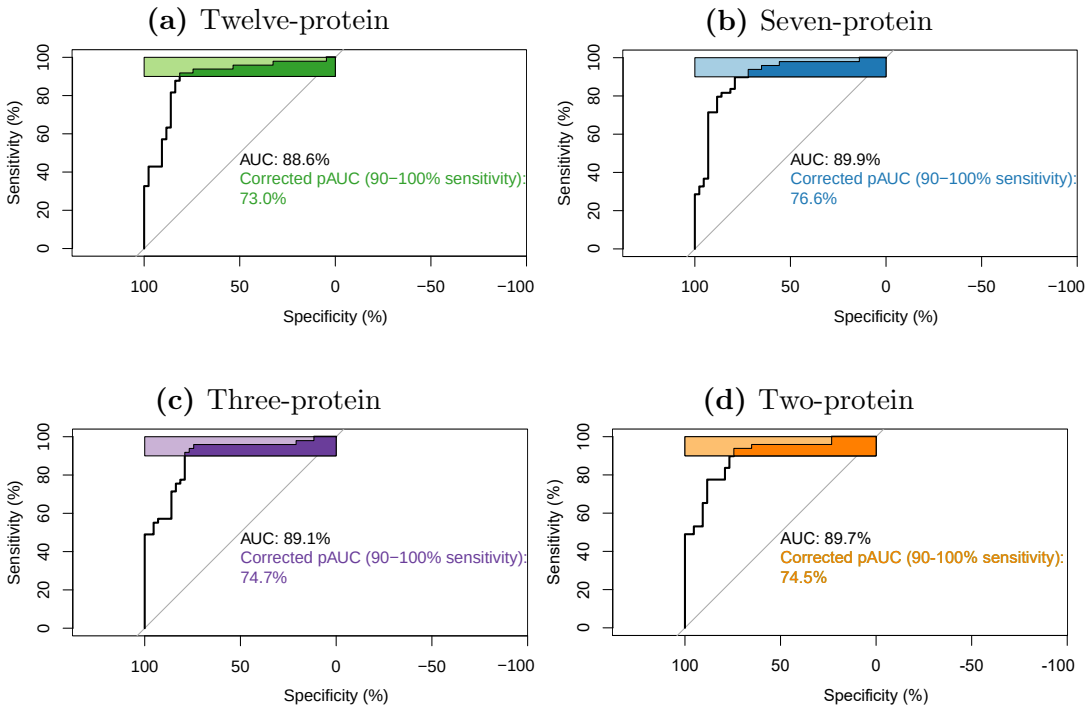


Figure 3.13: Receiver operating characteristic (ROC) curves for four different protein models differentiating between the acute bacterial and acute viral groups, using the training dataset. The area under the curve (AUC) presented for each with a partial AUC at sensitivity range from 90% to 100% also shown.

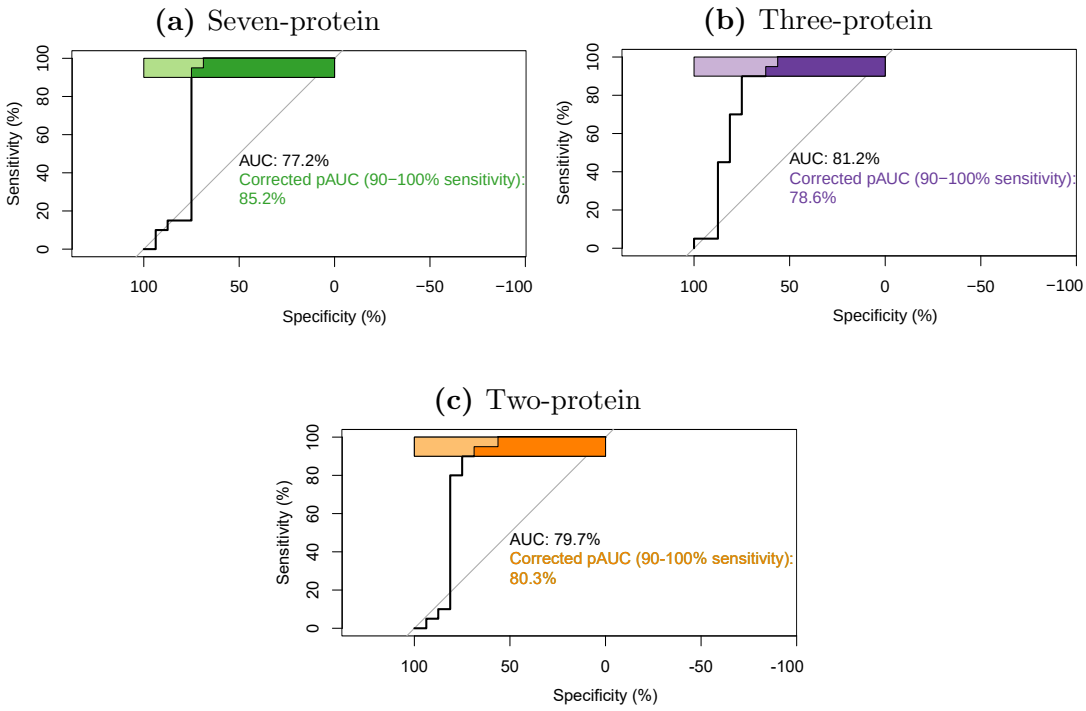


Figure 3.14: Receiver operating characteristic (ROC) curves for four different protein models differentiating between the acute bacterial and acute viral groups, using the test data set. The area under the curve (AUC) presented for each with a partial AUC at sensitivity range from 90% to 100% also shown.

Model performance - Test data

As all four models had similar results in the training data, the three models with fewer numbers of proteins were brought forward for analysis in the test dataset.

The performance statistics for the seven-protein, three-protein and two-protein models were similar. All three models had a sensitivity of 75% with the three-protein model having a slightly higher specificity, 95% compared with 90% in the seven-protein and two-protein models.

Table 3.6: Protein signature performance statistics in the test data with three different protein models compared.

Performance metric	**Seven-protein signature**		**Three-protein signature**		**Two-protein signature**	
Proteins included	A2GL, SAA2, S10A9, AACT, PLTP, IDH3A, VWF		A2GL, SAA2, LBP		A2GL, SAA2	
	Reference - bacterial	**Reference - viral**	**Reference - bacterial**	**Reference - viral**	**Reference - bacterial**	**Reference - viral**
Predicted bacterial	12	2	12	1	12	2
Predicted viral	4	18	4	19	4	18
Accuracy (95% CI)	0.83 (0.67, 0.94)		0.86 (0.71, 0.95)		0.83 (0.67, 0.94)	
Sensitivity	0.75 (0.48, 0.93)		0.75 (0.48, 0.93)		0.75 (0.48, 0.93)	
Specificity	0.9 (0.68, 0.99)		0.95 (0.75, 1)		0.9 (0.68, 0.99)	
Positive predictive value	0.86 (0.61, 0.96)		0.92 (0.64, 0.99)		0.86 (0.61, 0.96)	
Negative predictive value	0.82 (0.66, 0.91)		0.83 (0.67, 0.92)		0.82 (0.66, 0.91)	

3.3.6 Differential protein abundance - pathogen groups

LRTI groups were also classified by the likely pathogen found in different cases. RSV was the most common viral pathogen, and made up 56% (40/71) of the definite viral cases. Influenza was the next most common viral pathogen, detected in 24% (17/71) of the definite viral cases. There were fewer cases of parainfluenza, detected in 8% (6/71) of definite viral cases and human metapneumovirus, detected in 6% (4/71) of the definite viral cases.

S. pneumoniae was the most common bacterial pathogen found on bacterial or pleural culture, 55% (11/20), followed by *S. aureus*, 15% (3/20) and *S.* Typhi, 10% (2/20). As discussed in Section 2.2.3, participants with a nasopharyngeal sample culture-positive for *S. pneumoniae* Serotypes 1 or 5 were included in a probable pneumococcal group. This is because these were found to be the pneumococcal serotypes most commonly associated with invasive pneumococcal disease at the research site.[56]

As RSV, influenza viruses and *S. pneumoniae* were the most common pathogens these were the focus of the sub-group analyses.

RSV group

When acute samples from participants where RSV was detected were compared with convalescent samples from participants with viral LRTIs, 72 proteins were differentially abundant with an adjusted p-value >0.001, and an absolute log2FC>1, see Figure 3.15.

Next, acute samples from the RSV group were compared with acute samples

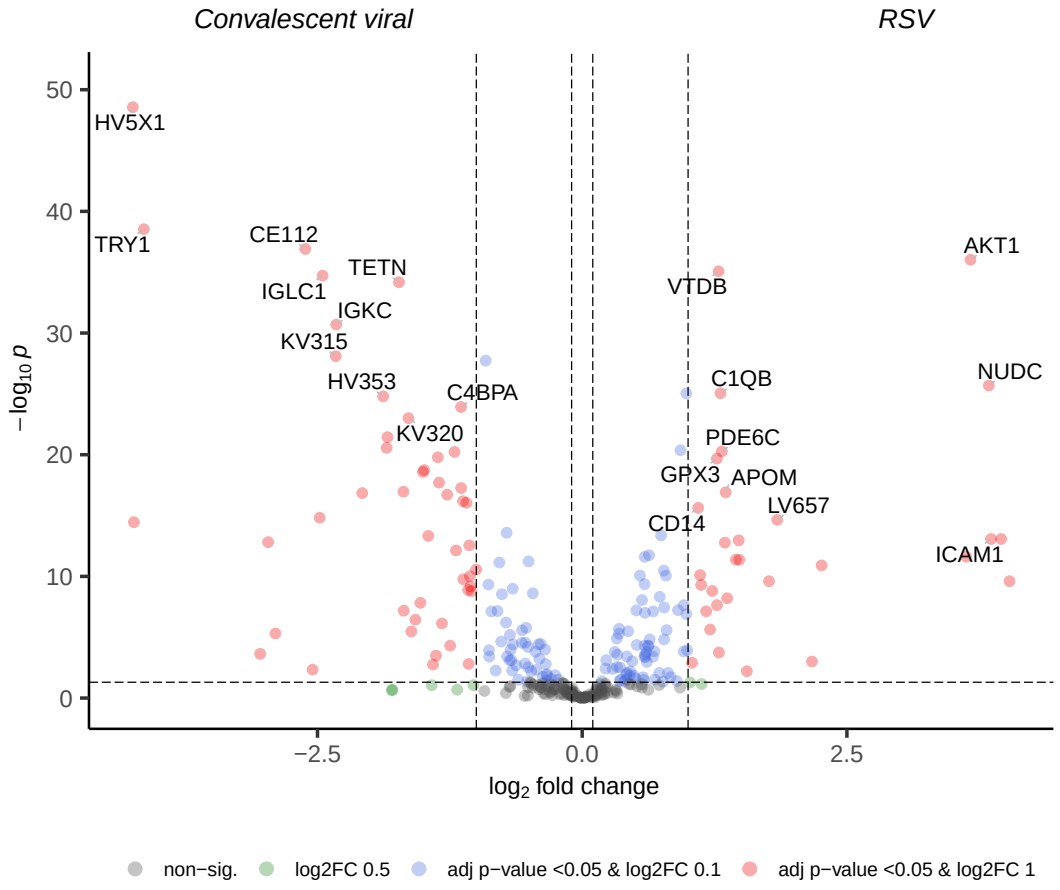


Figure 3.15: Volcano plot showing differentially abundant proteins between participants in the acute RSV group and convalescent samples from participants who had a viral LRTI. All participants in the cohort included. The labelled proteins are the top ten up- and down-regulated proteins, ranked by Benjamini-Hochberg adjusted p-values.

from participants with other viral LRTIs to look for proteins which may be specific to RSV infections, see Figure 3.16. Only two proteins were significantly more abundant in the RSV group, ZTB12 and QSOX1.

ORA pathway analysis was performed using the proteins differentially abundant between the acute RSV group and convalescent samples from participants with viral LRTIs. Enrichment analysis was carried out using the GO-BP ontology, 22 pathways were significantly over-enriched. Figure 3.18 shows the top over-represented pathways, ranked by p-value. Figure 3.19 is a network plot showing the top over-represented pathways and how the proteins contributing to each pathway are connected. Complement activation and humeral response pathways are among the top over-enriched pathways. The box plots in Figure 3.20 show the proteins contributing to these pathways.

Influenza group

When acute samples from participants where influenza was detected were compared with convalescent samples from participants with viral LRTIs, 71 proteins were differentially abundant, using an adjusted p-value >0.001, and an absolute log2FC>1, see Figure 3.21.

Next, acute samples from the influenza group were compared with acute samples from participants with other, non-influenza, viral LRTIs to look for proteins which may be specific to influenza infections, see Figure 3.22. Only one protein was significantly more abundant in the other acute viral group compared with the influenza group, immunoglobulin heavy variable 4-34 (HV434).

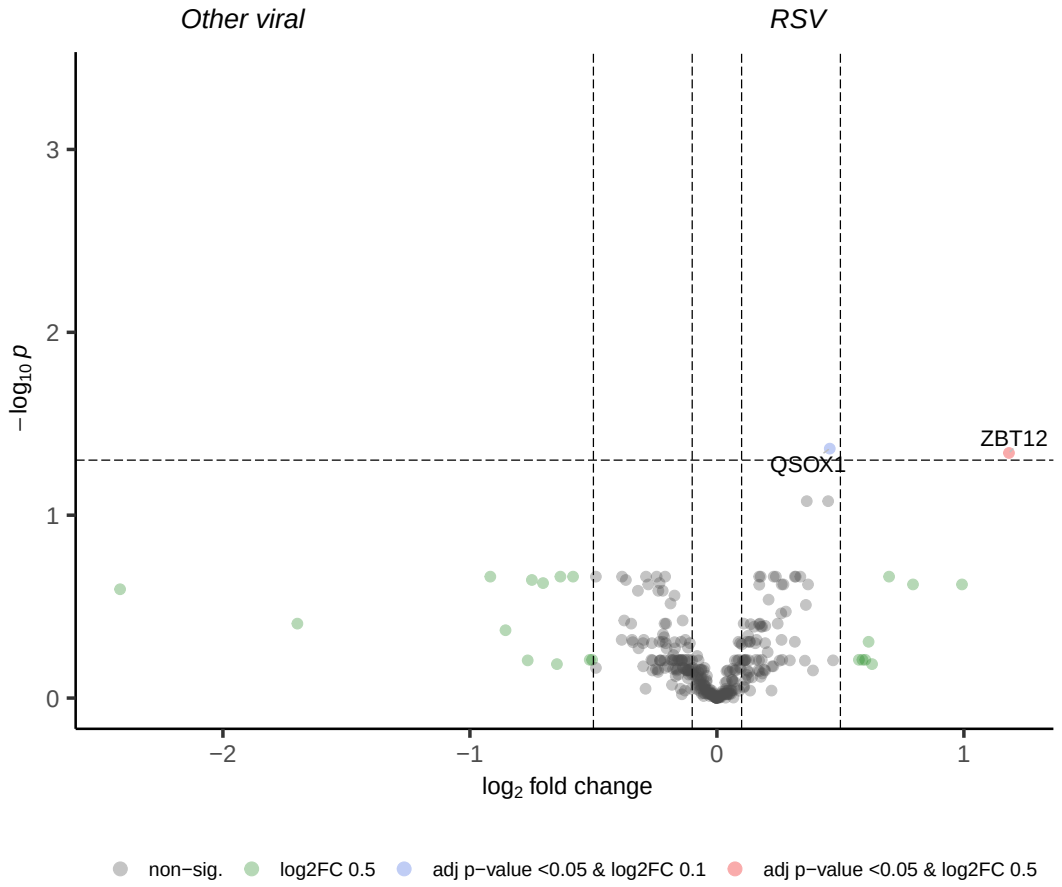


Figure 3.16: Volcano plot showing differentially abundant proteins between participants in the acute RSV group and acute samples from participants with other viral, non-RSV, LRTI. All participants in the cohort included. The labelled proteins are the top ten up- and down-regulated proteins, ranked by Benjamini-Hochberg adjusted p-values.

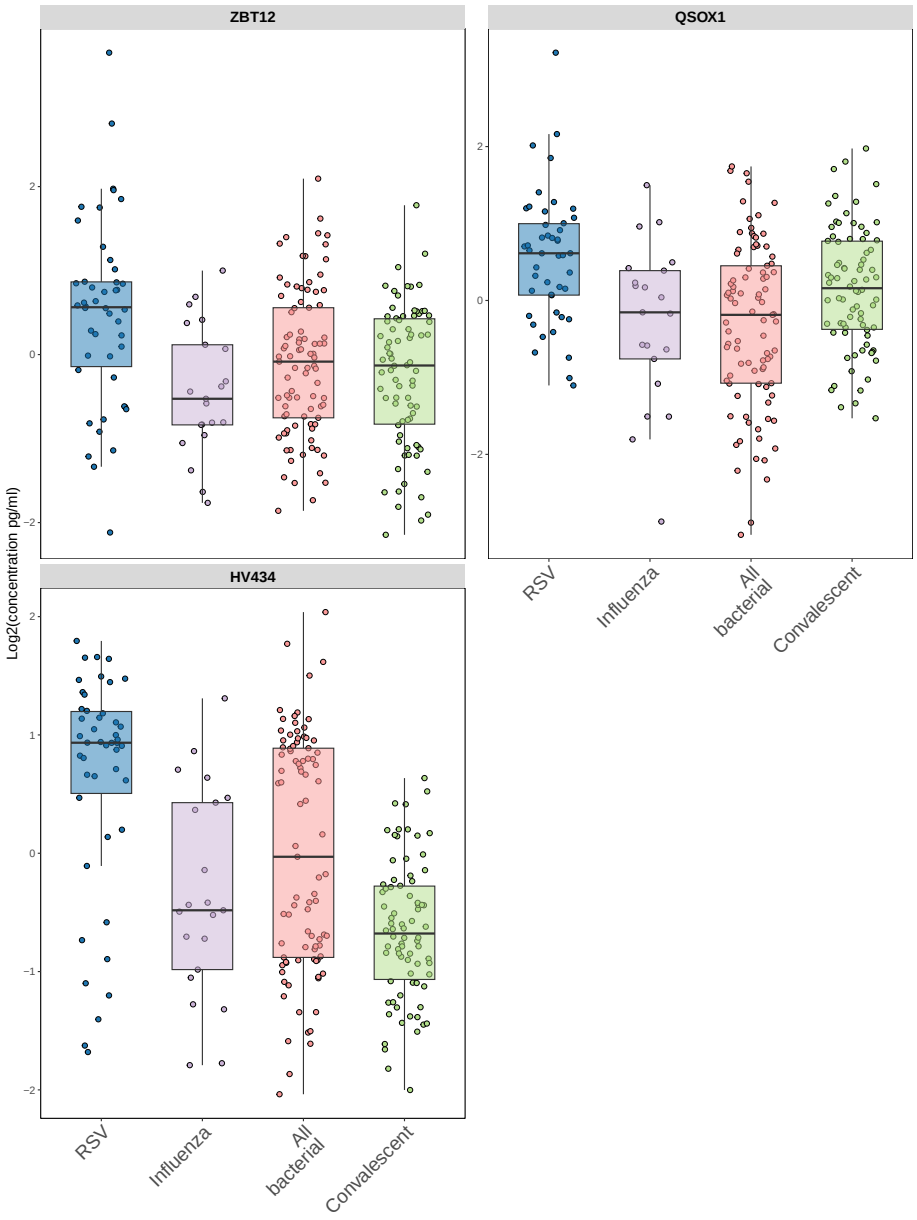


Figure 3.17: Box plots of the two proteins which have an adjusted p-value <0.05 when acute RSV samples are compared with acute samples from the other, non-RSV, viral group, ZBT12 and QCOX1. Also shown is HV434, the only protein that was differentially abundant when acute influenza samples were compared with acute samples from the other, non-influenza, viral group. All of the bacterial sample results and all of the convalescent sample results are grouped for comparison.

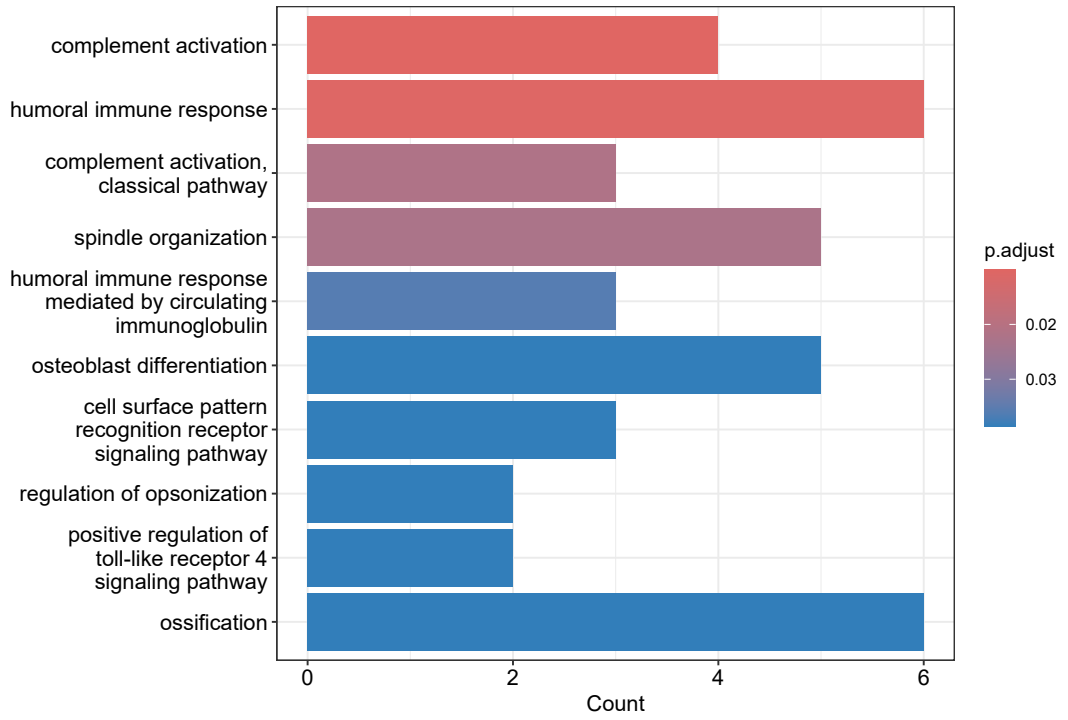


Figure 3.18: Results of over-representation analysis of the differentially abundant proteins between participants in the acute RSV group and convalescent samples from participants with viral LRTIs. The pathways were ranked by BH adjusted p-value. The length of the bars represents the number of proteins contributing to the over-enrichment in the pathway.

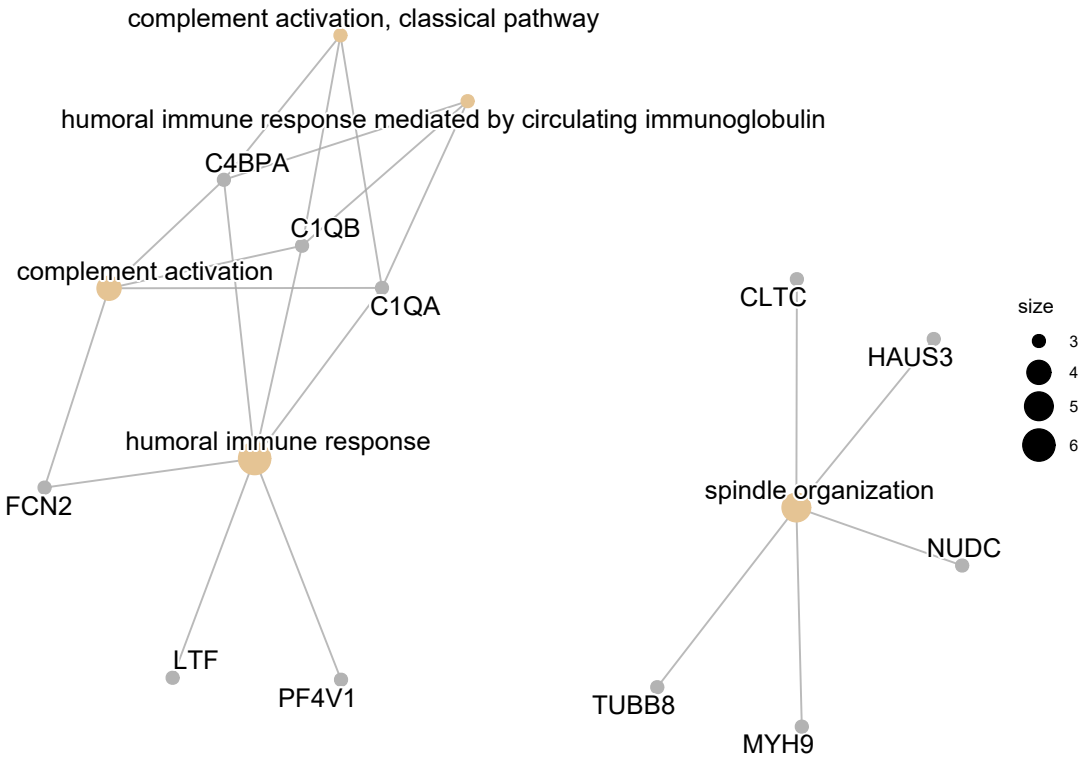


Figure 3.19: Over-representation analysis results of the differentially abundant proteins between participants in the acute RSV group and convalescent samples from participants with viral LRTIs. All participants in the cohort were included. The pathways were ranked by BH adjusted p-value. The size of dots represents the number of proteins contributing to the over-enrichment in the pathway.

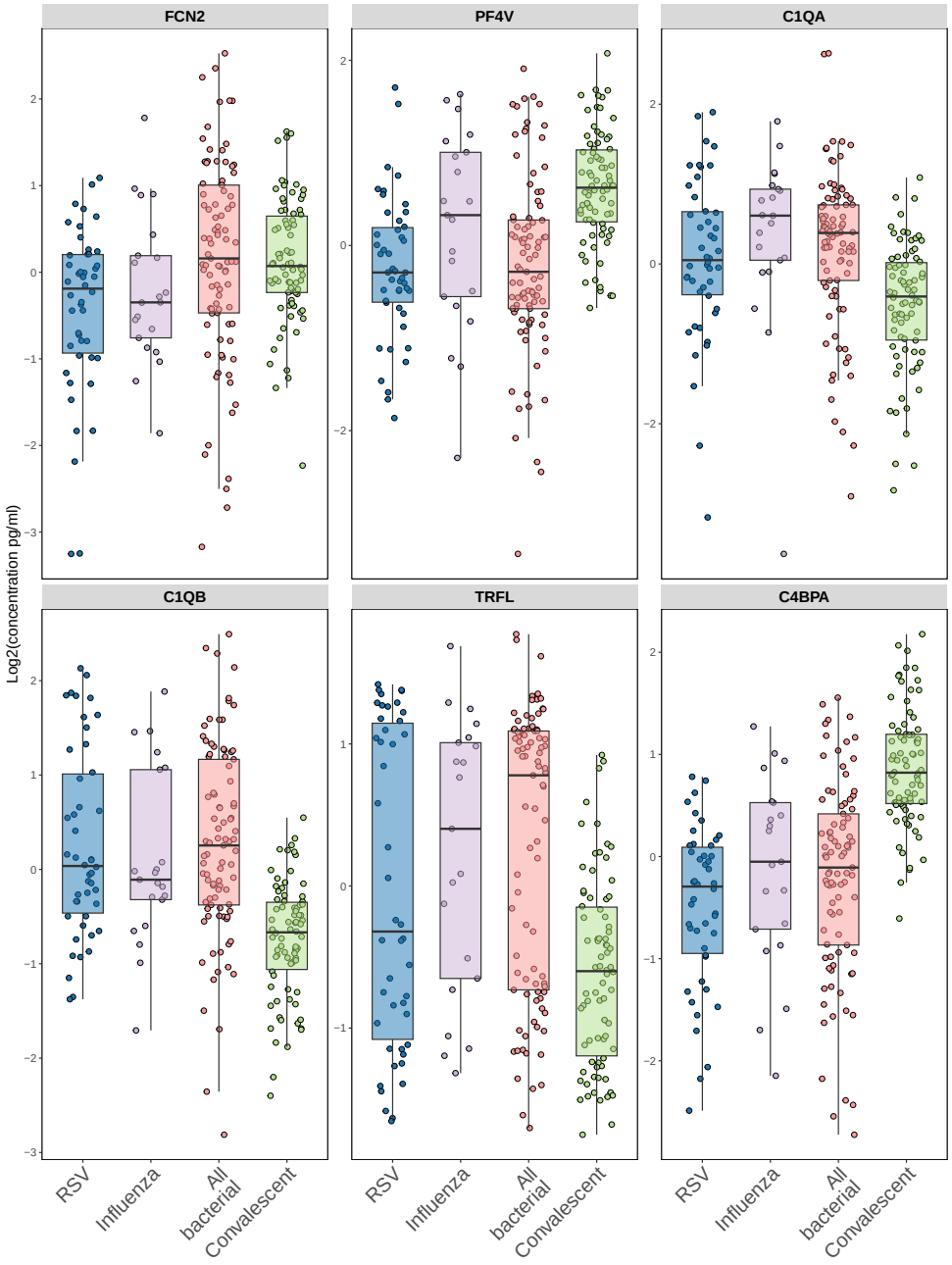


Figure 3.20: Proteins which contribute to the humoral immune response and complement activation pathways from the over-representation analysis results when participants in the acute RSV group are compared with convalescent samples from participants with viral LRTIs. All of the bacterial sample results and all of the convalescent sample results are grouped for comparison.

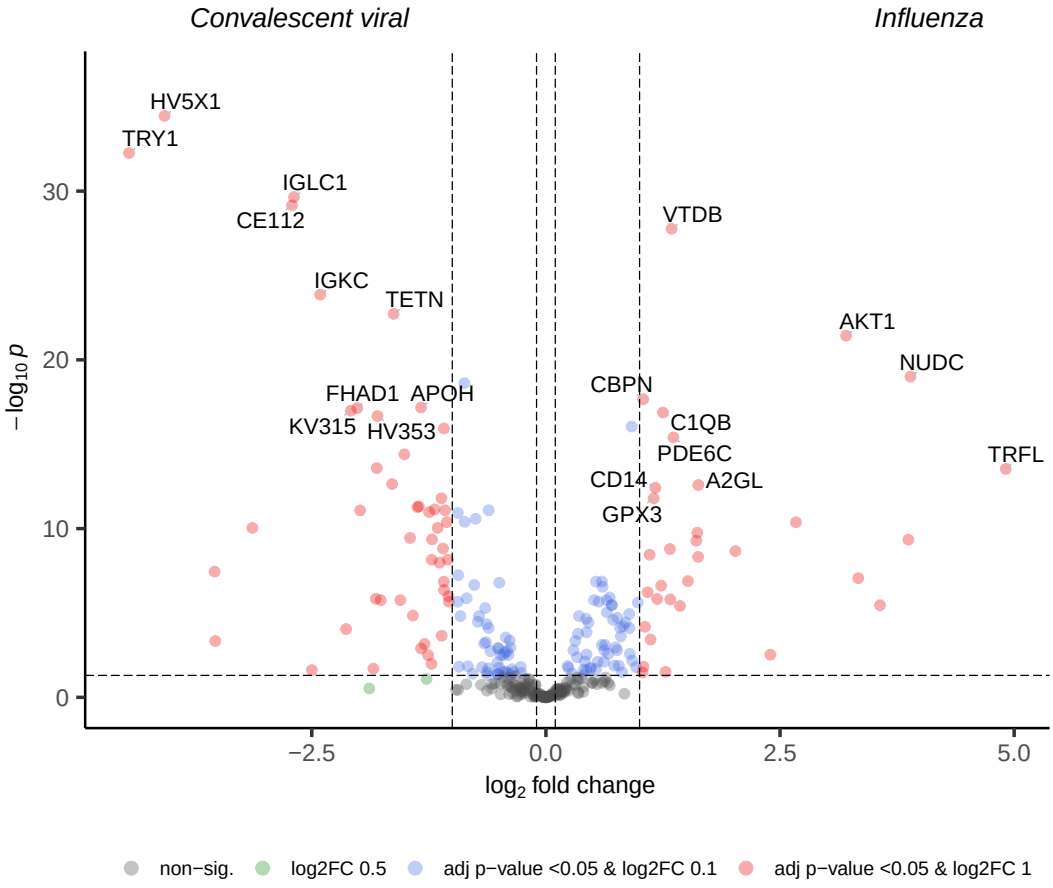


Figure 3.21: Volcano plot showing differentially abundant proteins between participants in the acute influenza group and convalescent samples from participants with a viral LRTI. The labelled proteins are the top ten up- and down-regulated proteins, ranked by Benjamini-Hochberg adjusted p-values.

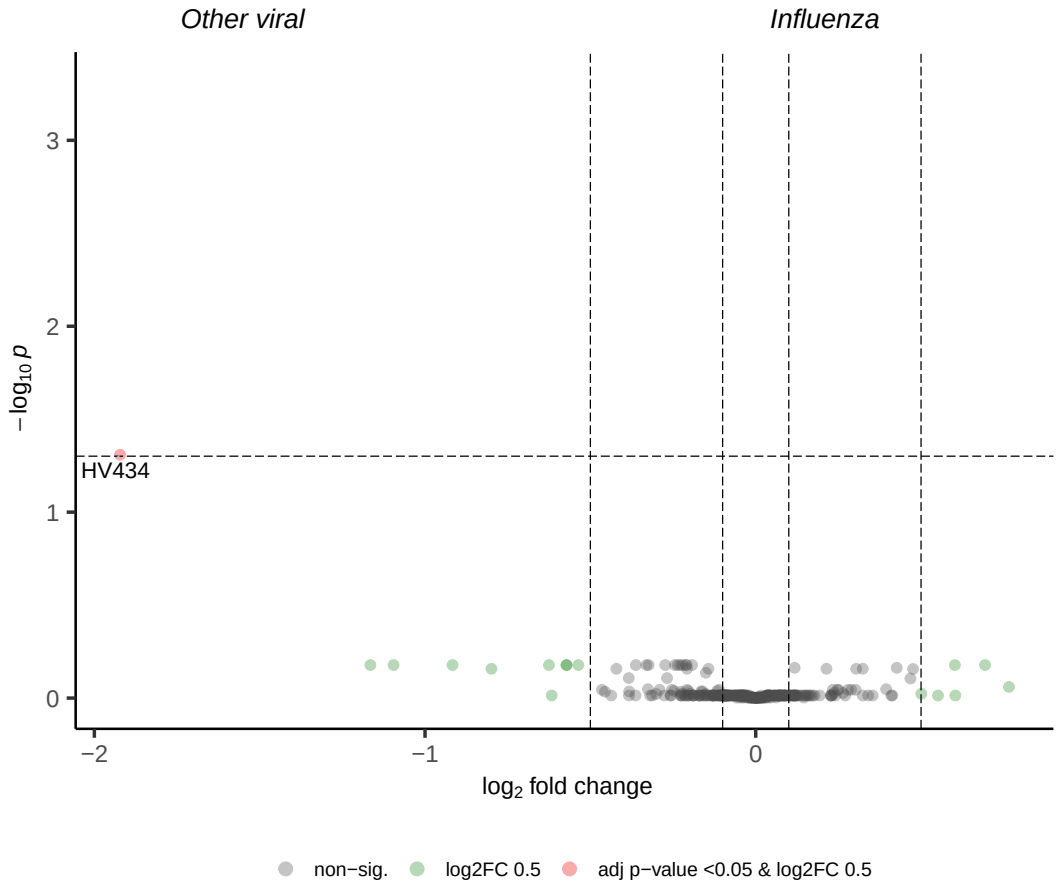


Figure 3.22: Volcano plot showing the one differentially abundant protein between participants in the acute influenza group and acute samples from participants with other viral, non-influenza, LRTI. The labelled proteins are the top ten up- and down-regulated proteins, ranked by Benjamini-Hochberg adjusted p-values.

ORA was performed using the proteins differentially abundant between the acute influenza group and the convalescent samples of participants who had viral LRTIs. Enrichment analysis was carried out using the GO-BP ontology, 57 pathways were significantly over-enriched.

Figure 3.23 shows the top over-represented pathways, ranked by adjusted p-value. Figure 3.24 is a network plot showing the top over-represented pathways and how the proteins contributing to each pathway are connected. Complement activation and humeral response pathways are among the top over-enriched pathways. The box plots in Figure 3.20 show the proteins contributing to these pathways. Separate clusters relate to spindle organisation and high-density lipoprotein particle clearance, see Appendix C.

Pneumococcal group

When acute samples from participants where influenza was detected were compared with convalescent samples from participants with bacterial LRTIs, 84 proteins were differentially abundant with an adjusted p-value >0.001, and an absolute log2FC>1, see Figure 3.25.

Differential abundance analysis of blood or pleura fluid culture-confirmed pneumococcal LRTIs compared with other, non-pneumococcal, bacterial LRTIs found no differentially abundant proteins using an adjusted p-value threshold of 0.05 and a log2-fold-change of 0.5. The results were the same when the probable pneumococcal group were included in the analysis and compared with other acute bacterial samples, see Appendix C.

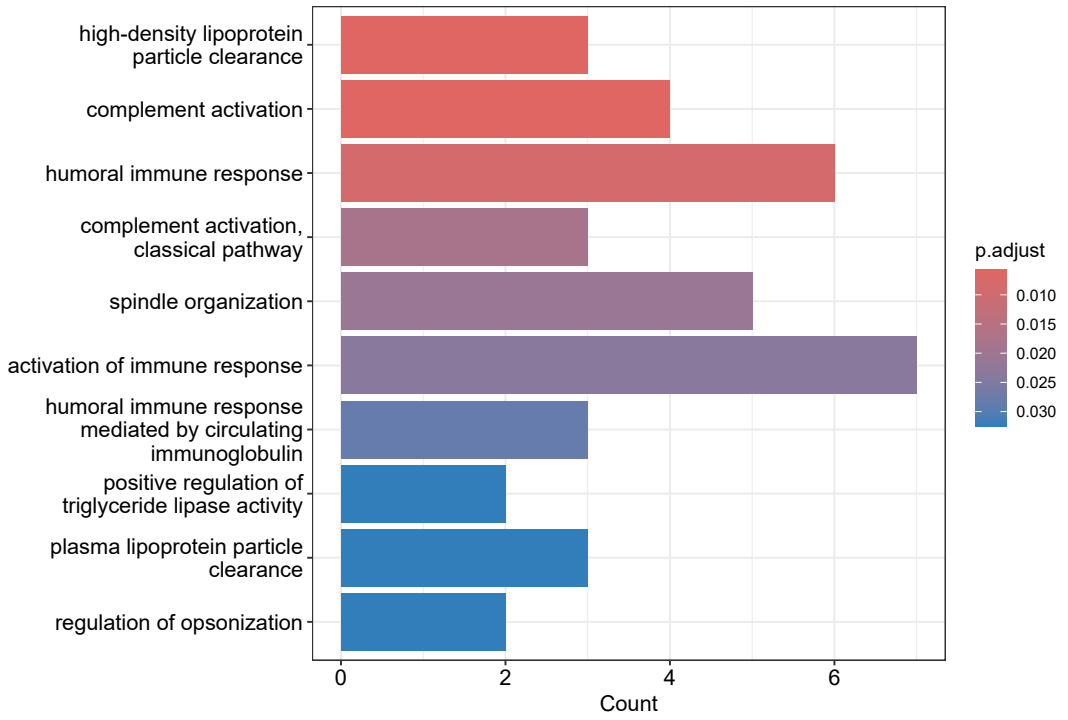


Figure 3.23: Over-representation analysis of the differentially abundant proteins between participants in the acute influenza group and convalescent samples from participants with viral LRTIs. The pathways were ranked by BH adjusted p-value. The length of the bars represents the number of proteins contributing to the over-enrichment in the pathway.

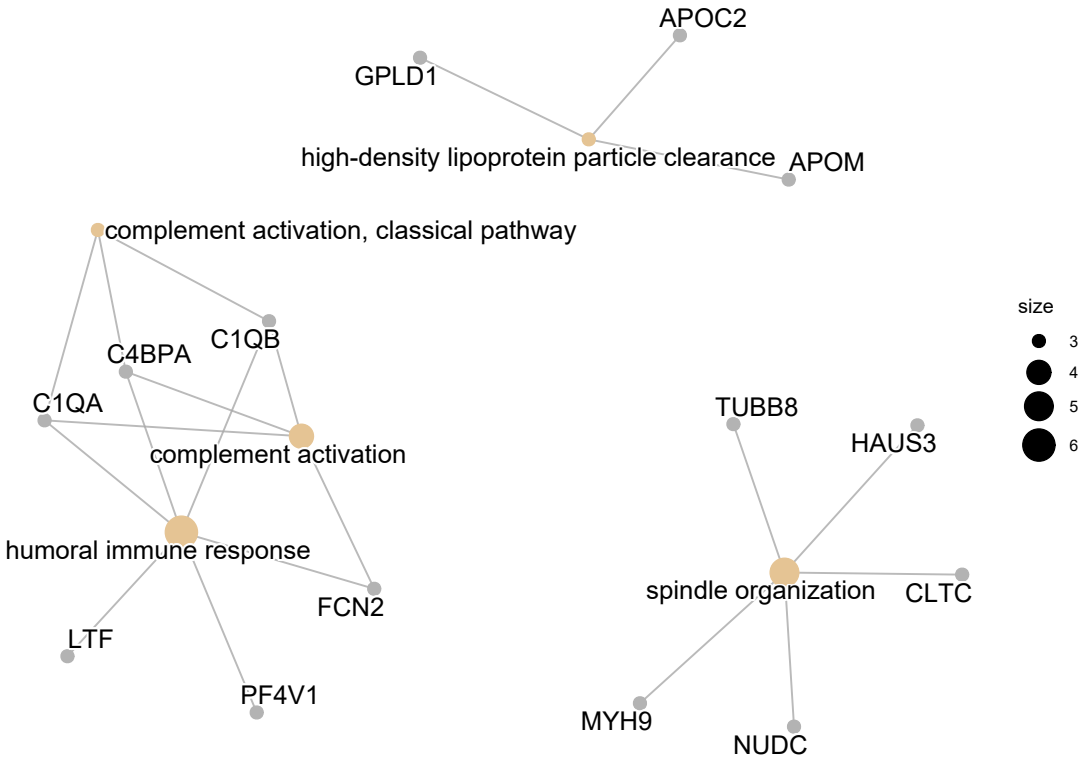


Figure 3.24: Over-representation analysis of the differentially abundant proteins between participants in the acute influenza group and convalescent samples from participants with viral LRTIs. The pathways were ranked by BH adjusted p-value. The size of the dots represents the number of proteins contributing to the over-enrichment in the pathway.

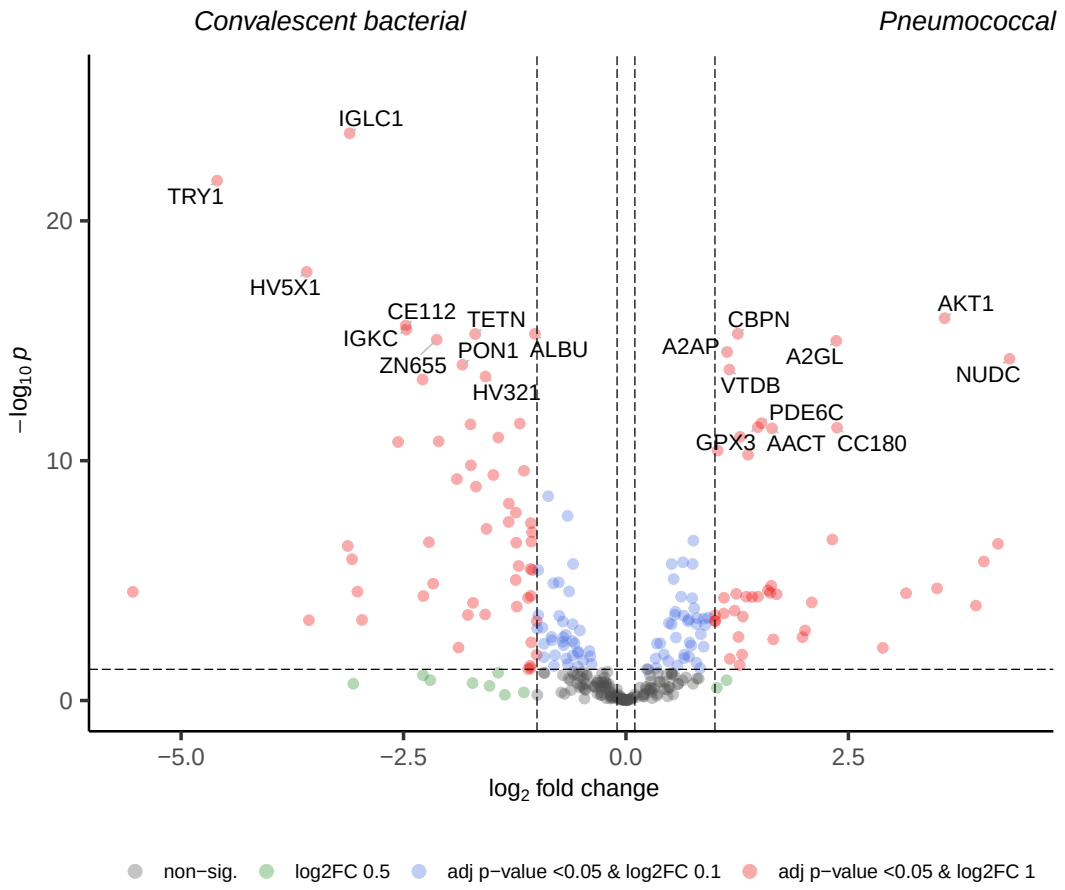


Figure 3.25: Volcano plot showing that there were no differentially abundant proteins between participants in the acute pneumococcal group and acute samples from participants with convalescent samples from participants with bacterial LRTI. Only samples that were blood or pleural fluid positive for *S. pneumoniae* were included in the pneumococcal group. The labelled proteins are the top ten up- and down-regulated proteins, ranked by Benjamini-Hochberg adjusted p-values.

ORA was performed using the proteins differentially abundant between the acute pneumococcal group and the convalescent samples of participants who had presented with bacterial LRTIs. Enrichment analysis was carried out using the GO-BP ontology, 57 pathways were significantly over-enriched.

Figure 3.26 shows the top over-represented pathways, ranked by adjusted p-value. Figure 3.24 is a network plot showing the top over-represented pathways and how the proteins contributing to each pathway are connected. Acute-phase response and negative regulation of peptidase activity are among the top two over-enriched pathways. The box plots in Figures 3.28 and 3.29 show the proteins contributing to these pathways. Plasminogen activation is also a top-ranked pathway shown in the network, see Appendix C for proteins related to this pathway.

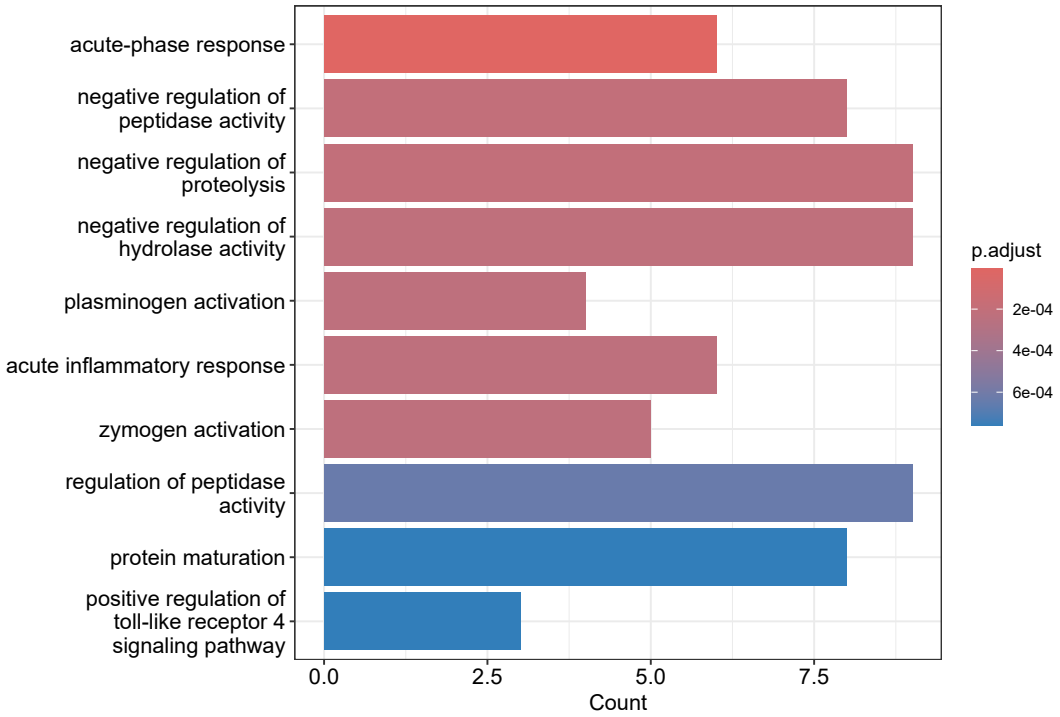


Figure 3.26: Over-representation analysis of the differentially abundant proteins between participants in the acute pneumococcal group the convalescent samples of participants who presented with bacterial LRTIs. Only samples that were blood or pleural fluid positive for *S. pneumoniae* were included in the pneumococcal group. The pathways were ranked by BH adjusted p-value. The length of the bars represents the number of proteins contributing to the over-enrichment in the pathway.

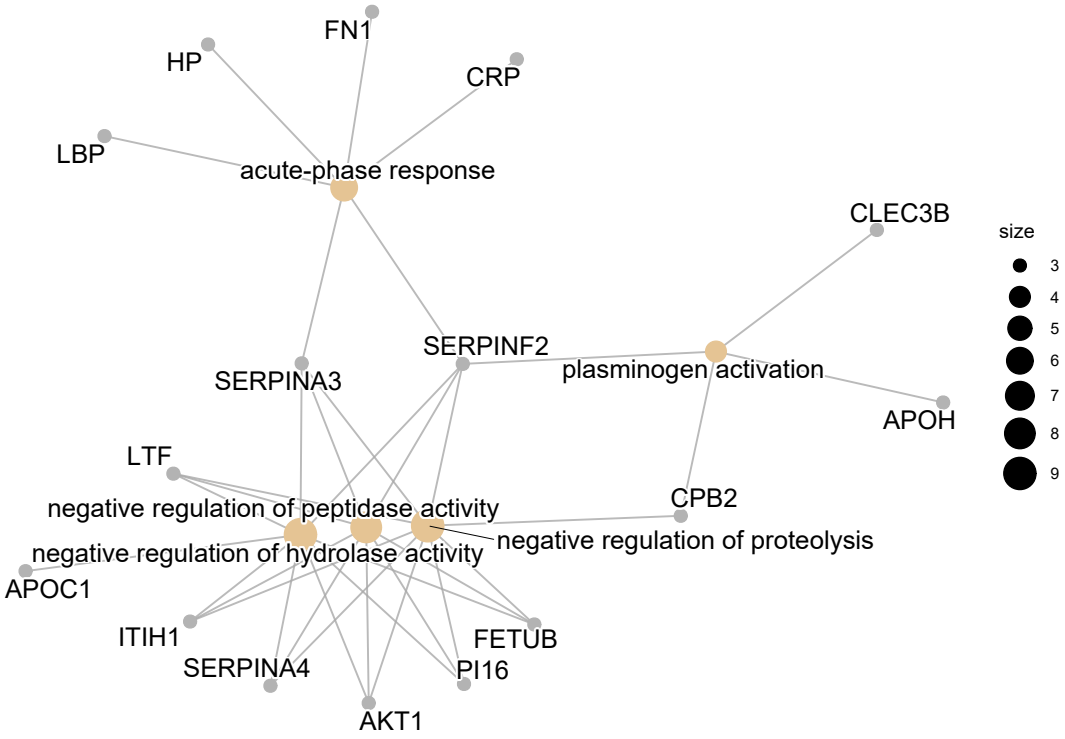


Figure 3.27: Over-representation analysis of the differentially abundant proteins between participants in the acute pneumococcal group the convalescent samples of participants who presented with bacterial LRTIs. Only samples that were blood or pleural fluid positive for *S. pneumoniae* were included in the pneumococcal group. The pathways were ranked by BH adjusted p-value. The size of the dots represents the number of proteins contributing to the over-enrichment in the pathway.

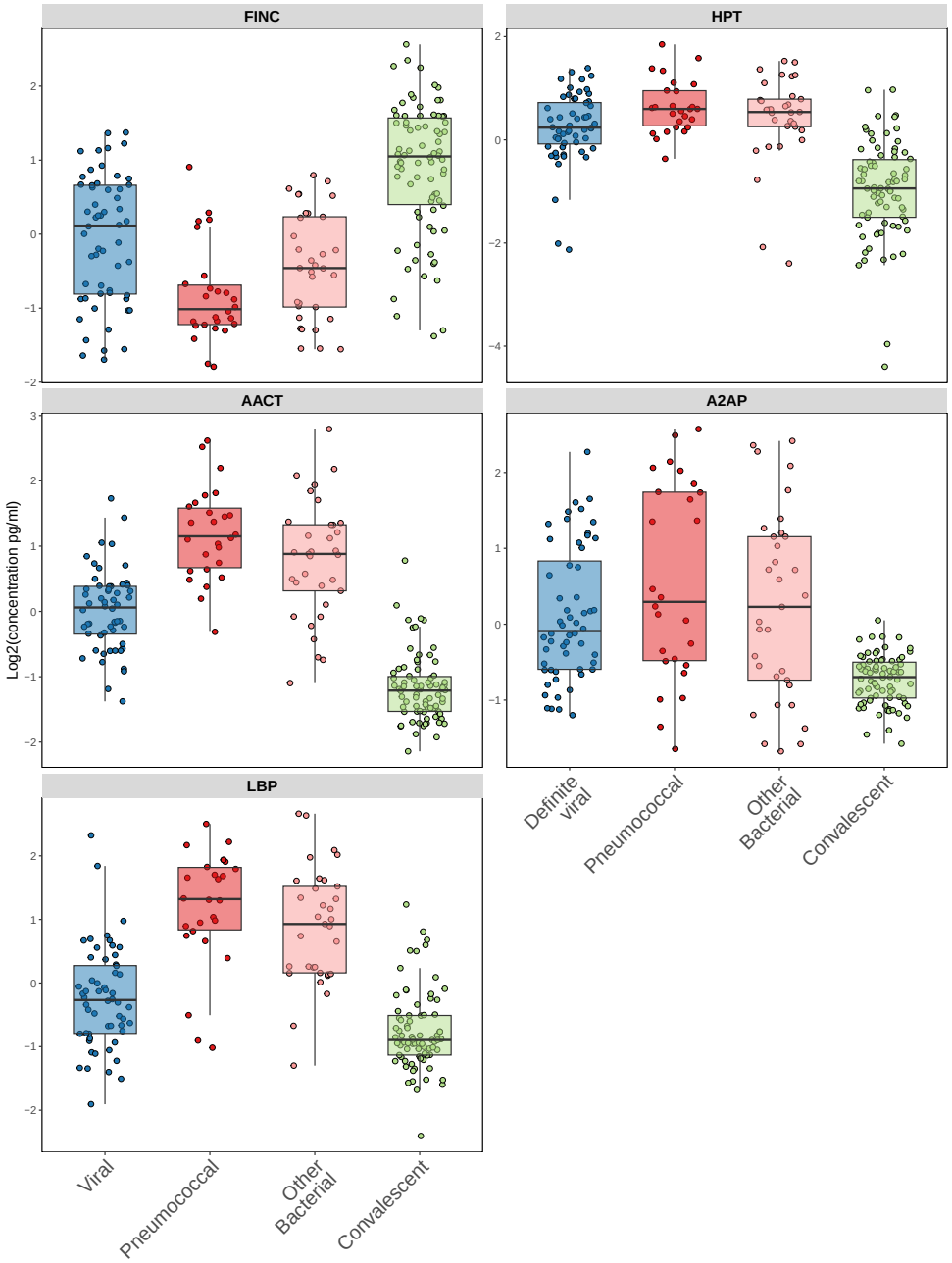


Figure 3.28: Proteins which contribute to the acute-phase response pathway from the over-representation analysis results when participants in the acute pneumococcal group are compared with convalescent samples from participants with bacterial LRTIs. All of the viral sample results and all of the convalescent sample results are grouped for comparison.

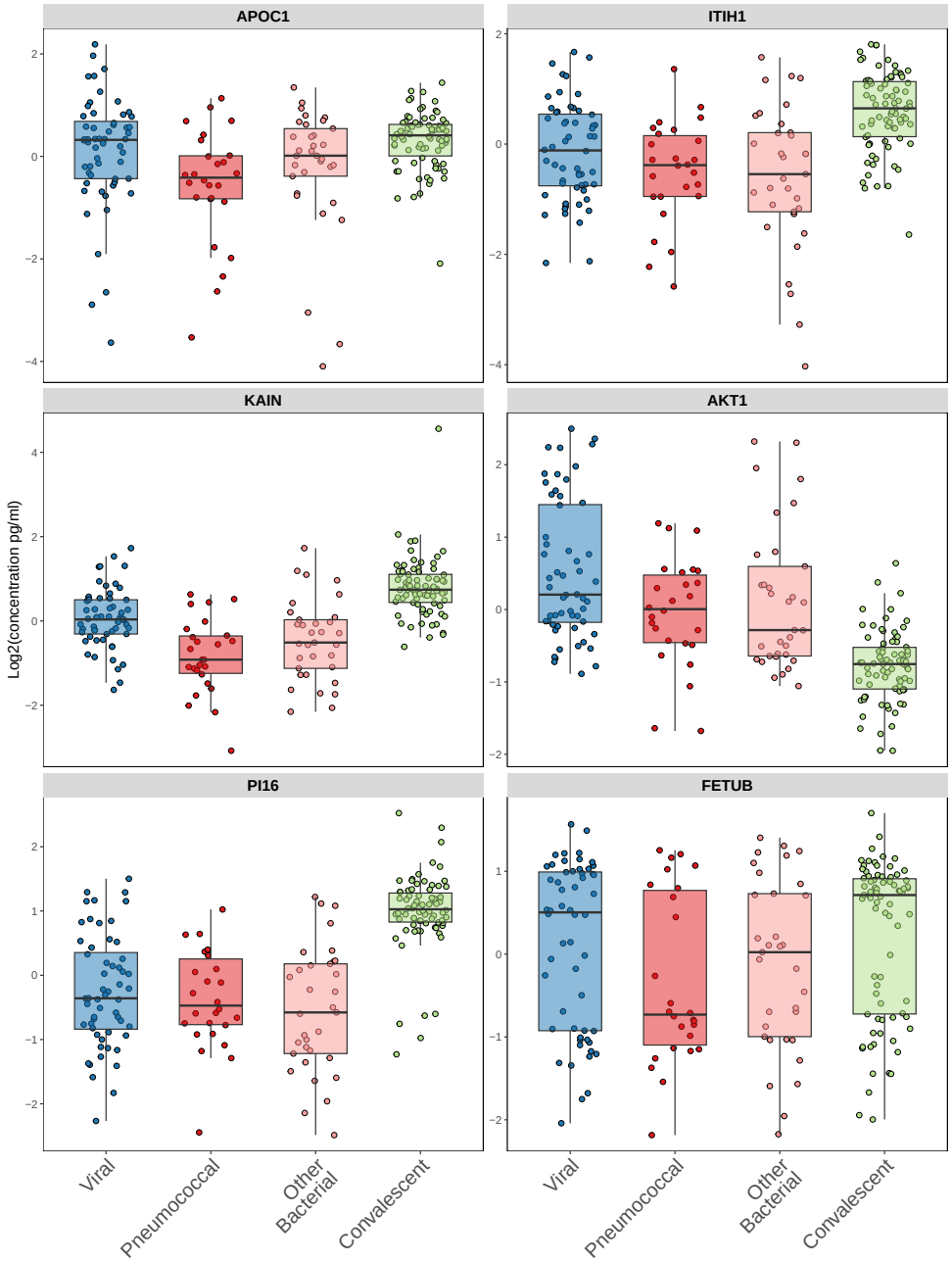


Figure 3.29: Proteins which contribute to the negative regulation pathways from the over-representation analysis results when participants in the acute pneumococcal group are compared with convalescent samples from participants with bacterial LRTIs. All of the viral sample results and all of the convalescent sample results are grouped for comparison.

3.4 Discussion

Differential abundance analysis between bacterial and viral LRTIs highlighted 13 differentially expressed proteins. A two-protein signature identified bacterial infections with moderate sensitivity.

Pathway analysis of the RSV and influenza groups showed that proteins in similar immunological pathways were over-presented in both groups, compared with convalescent samples. Results of pathway analysis comparing samples from the acute pneumococcal group with convalescent samples highlighted proteins related to the acute immune response and negative regulation of protein breakdown.

3.4.1 Protein signatures

The first question to answer is why do we need new protein tests when we have proteins like CRP and procalcitonin that can differentiate between bacterial and viral infections. As discussed in Section 1.6, neither of these tests is perfect and there is room for improvement. A meta-analysis of the utility of CRP and procalcitonin to identify bacterial infections reported a sensitivity of 85% and a specificity of 80% for CRP, and a sensitivity of 76% with a specificity of 69% for procalcitonin. The two-protein signature identified in this chapter had comparable results to these commonly used tests, with a sensitivity of 75% and a specificity of 90%. It is difficult to compare the two-protein signature directly with CRP in our data, as CRP was used as part of the classification process.

A combination of proteins may offer increased accuracy in identifying bacterial infections. When considering how many proteins to include in our model, the

accuracy increased substantially when a second protein was added to the model. The potential utility of combining proteins has been reported previously; TRAIL, CRP and IP-10 were reported to have improved accuracy for identifying bacterial infections when compared with using any of these three proteins alone. Eden *et al.* reported that in a study of febrile adults and children presenting to the emergency department over four years, the three proteins combined had a sensitivity of 96% (95% CI 78% to 100%) with a specificity of 93% (95% CI 87% to 97%).[82] The Eden *et al.* study based their results on a relatively low number of bacterial cases, however, the potential for improved accuracy when combining proteins is encouraging.

3.4.2 Differentially abundant proteins

The two-protein signature to differentiate between bacterial and viral LRTIs included A2GL and SAA2.

A2GL is involved in normal physiological activity in the nervous system but has been linked to disease processes.[212] In 2013, Honda *et al.* reported that A2GL was an upstream regulator of transforming growth factor-beta.[213] It is through this pathway that A2GL is thought to influence disease processes.[214] Two 2023 studies reported promising results when using A2GL as a disease biomarker. Shimoyama *et al.* reported A2GL as a useful biomarker for disease activity in inflammatory bowel disease, and Arredondo Montero *et al.* suggested urine levels of A2GL could be a useful marker in urine during paediatric appendicitis.[215, 216] To my knowledge, A2GL has not been suggested as a marker to identify bacterial LRTIs prior to this thesis. However, recent reports proposing A2GL

could be a biomarker for inflammatory conditions suggest that further exploration of the role of A2GL in paediatric infection is warranted.

The second protein in the two-protein signature, SAA2 has a more recognised role in the immune response. Both serum amyloid A-1 (SAA1) and SAA2 were increased in bacterial LRTIs in our cohort. Serum amyloid A has several different isotypes, including SAA1 and SAA2, which are proteins derived from almost identical genes. SAA1 and SAA2 are acute-phase reactants produced by the liver.[217] In a study of 106 paediatric bacterial pneumonia cases and 87 non-bacterial pneumonia cases serum amyloid A was used in combination with CRP and procalcitonin to improve the identification of bacterial versus non-bacterial pneumonia in children. It is not clear from the study how the cases were classified and if a particular isotype of serum amyloid A was measured. The sensitivity for the combination of serum amyloid A, CRP and procalcitonin to identify bacterial infections was 97% with 98% specificity. For serum amyloid A alone, sensitivity was 78% and specificity was 93% for identifying bacterial pneumonia.[218] Serum amyloid A could be a useful biomarker to help in identifying bacterial infections in children, and should be considered when looking for combinations of proteins to use in future diagnostic tests.

3.4.3 Pathway analysis and pathogen groups

When ORA was used to compare acute RSV cases with convalescent samples, complement activation and humoral immune response pathways were the highest-ranked. Similar pathways were also highest-ranked when acute influenza samples were compared with convalescent samples. Pathway analysis in the RNA-seq

cohort in Section 2.4 also highlighted common biological pathways across RSV and influenza comparisons. In the RNA-seq results, genes related to TRAF6 had increased expression in influenza and RSV groups compared with bacterial and convalescent samples, see Figure 2.31. As discussed in Section 2.4, TRAF6 is linked to type I interferon production.[197] Type I interferons are implicated in the early antiviral humoral immune response.[219] This could explain why humoral immune pathways were significant pathways in the RSV and influenza pathway analyses in this chapter.

When acute RSV samples were compared with acute samples from other, non-RSV, viral LRTIs one protein had increased differential abundance in the RSV group, zinc finger and BTB domain-containing protein 12 (ZBT12); ZBT12 had increased abundance in RSV LRTIs compared with other viral LRTIs. ZBT12 is a Broad-complex, Tramtrack, and Bric-à-brac (BTB)-domain containing protein. BTB domains have diverse cellular functions with specific inhibition of protein interactions.[220] ZBT12 has a role in haemopoiesis and is a predicted transcription factor. Methylation of ZBT12 has been associated with increased inflammatory potential of blood cells.[221] The ZBTB12 gene, which translates to ZBT12, has been associated with patients who are elite controllers of HIV.[222] To my knowledge, no other study has linked ZBT12 to RSV infection.

When acute samples from participants where influenza was detected were compared with other, acute non-influenza LRTI samples, HV434 was decreased in influenza. No other proteins were significantly different between the groups. The immunoglobulin HV434 gene encodes autoreactive antibodies. HV434 antibodies are usually undetectable in healthy sera.[223] HV434 antibodies have been ob-

served to rise in acute infections including Epstein-Barr virus and Mycoplasma pneumoniae.[224, 225] In the present study, levels of HV434 were similar in the convalescent and acute influenza groups, with the highest levels in the acute RSV group; several bacterial cases had levels of HV434 that were similar to the RSV group, which suggests that high levels of HV434 are not specific to RSV and instead HV434 may be a more general marker of inflammation.

As neither ZBT12 nor HV434 have been associated with RSV in the previous publications, these findings would need to be validated in follow-up studies to evaluate their significance.

3.4.4 Limitations

As discussed in Section 2.4.4, there were limitations which made it more complicated to interpret the results, including that gold standard diagnostics may not correctly classify cases and the difference in ages between the groups. The age differences have been adjusted for in the analyses, however, it would have been easier to compare groups if the ages had been more similar.

CRP was used as part of the LRTI classification system. Definite bacterial cases were classified independently of their CRP result, but the probable bacterial group was classified using a CRP >60 mg/L. CRP interacts with other proteins during the acute inflammatory response.[226] Proteins linked to CRP were more likely to also be elevated in the cases classified by high CRP levels.

This was a single-site study in a Nepali hospital. To validate the signatures presented in this chapter, the results would need to be replicated in other cohorts.

Future work to address this is outlined in the next section.

3.4.5 Future directions

The protein signatures identified in this chapter need to be evaluated in other populations. Another prospective cohort study, enrolling children with any sign of infection, is underway at Patan Hospital; this new cohort study can act as a validation cohort for the results presented in this chapter. The new study will allow these protein signatures to be tested in a population with a broader range of infections. If the findings can be replicated this would validate the findings in another Nepali cohort, however, the protein signatures would need to then be validated in a non-Nepali setting to ensure utility in other settings.

The addition of a small protein signature to the available protein-based diagnostic tests has the potential to improve diagnostic accuracy. A future protein biomarker study which evaluated the sensitivity of CRP, procalcitonin and the proteins included in the signatures presented in this chapter would show the potential added value of combining these new protein signatures with established biomarkers.

In Chapter 6, I integrate data from this MS proteomics experiment with data from the RNA-seq experiment in Chapter 2 and the cytokine panel in Chapter 4 to see if a multi-platform signature performs better than the protein signatures presented in this chapter.

3.4.6 Conclusions

The protein signatures identified in this chapter could be used as a diagnostic test to differentiate between bacterial and viral infections. However, using these protein signatures in combination with protein biomarkers already in clinical use may be the best way forward.

Chapter 4

Cytokine changes in Nepali children with lower respiratory tract infections

4.1 Introduction

4.1.1 Chapter in context

This chapter describes the analysis of a plasma cytokine panel in Nepali children with LRTIs at Patan Hospital, Nepal. RNA-seq and MS proteomics results from the same cohort study were presented in Chapters 2 and 3.

As previously described in Chapter 1, LRTIs are an important cause of infection in children worldwide.[20] Current diagnostic tests lack the sensitivity to confidently rule out bacterial infection in a timely manner.[17, 60, 61] Improved diagnostic tests could assist clinicians in managing patients appropriately and identifying children at risk of serious bacterial infections. More accurate diagnostic tests should lead to more judicial use of antibiotics, which can help to tackle the growing problem of antimicrobial resistance.[17, 3, 20] Many different types of molecules have been suggested as potential diagnostic tests. This chapter focuses on the potential utility of cytokines in infectious disease diagnostics.

Cytokines are important proteins in the immune response and have the potential to be useful biomarkers of infection. The cytokine results in this chapter are from the prospective cohort study described in Section 2.2.1. A panel of cytokines were measured in samples taken from children soon after their admission to hospital with signs of pneumonia. Cytokine levels were also measured when these same children had recovered from illness, to act as control levels in health.

The results presented in this chapter show the relationship between different cytokine levels and paediatric LRTI aetiology. In the following sections, I discuss

the role of cytokines and their potential utility as diagnostic tests.

4.1.2 Function of cytokines

Cytokines are a heterogeneous group of small proteins that are important for communication within the immune system. A complex network of interacting cells, cytokines and receptors is required to produce a coordinated immune response to infection.[115, 116]

Cytokines act in different ways and are divided into different groups depending on their physical properties, cells from which they are secreted or primary functions. Cytokines can be divided based on whether they have predominantly pro- or anti-inflammatory properties, as discussed below.[116] Interleukins, colony-stimulating factors, tumour necrosis factors, interferons and chemokines are important cytokine groups. These cytokine groups were discussed in Section 1.7.5.

Several studies have reported the use of different cytokines as potential biomarkers in various disease states. The evidence supporting the use of the cytokines measured in this chapter is discussed in Section 4.1.4.

4.1.3 Platforms for detecting cytokines

The reliable measurement of cytokines can be used to diagnose and understand disease processes. Different platforms are available for detecting cytokines, each with their own advantages and disadvantages.

As discussed in Chapter 3, advances in MS proteomics, including the use of DIA approaches mean that a greater variety of proteins can be detected using MS approaches. However, limitations still exist in detecting lower abundance proteins, such as cytokines. The sensitivity of MS proteomics is usually insufficient to confidently detect cytokines at the range of concentration that these small proteins are present in samples.[114] However, this may improve with newer proteomic technologies.

ELISA is one of the most commonly used approaches for detecting analytes in biological samples. There are different methods used in ELISA protocols but the basic principles are usually the same. ELISA uses the specificity of detection antibodies to measure analytes of interest. ELISA assays usually rely on the addition of a substrate which can be converted to a coloured product, and the colour change is proportional to the amount of analyte in the sample.[129, 227]

The MSD platform was used in the experiment discussed in this chapter. MSD is similar to ELISA except MSD measures electrochemiluminescence rather than colour change. MSD has several potential advantages over conventional ELISA including decreased inter-assay variability and increased multiplexing capabilities with smaller sample volumes.[130]

MSD assays use electrochemiluminescence methods to detect and measure multiple targets simultaneously. Electrochemiluminescence involves the production of light from an electrochemical reaction. An electric current generates a chemical reaction which produces energy and emits light.[228] Electrochemiluminescence techniques can detect protein targets with a high sensitivity and specificity.[229, 230] The MSD platform allows measurements of up to ten pro-

teins in a 96-well multi-spot plate.[231] The principle of the MSD assay is a sandwich antibody format, see Figure 4.1. At the bottom of each well on an MSD plate, there are up to ten graphite electrodes. Ten linkers are combined with ten biotinylated capture antibodies and added to each well. Each linker has a biotin-binding site that allows it to bond with the biotinylated capture antibody. Each linker also has a domain which binds to the matching electrode at the base of the well. When a clinical sample is added to the well, each capture antibody will recognise a specific protein. Following this, detection antibodies conjugated to light-emitting labels are added to the well. This completes the sandwich immunoassay. A voltage is then applied to the plate causing the labels to emit light. The amount of light is measured to calculate the abundance of each protein present in the sample.[231]

4.1.4 Cytokines included in Meso Scale Discovery panel

It is difficult to interpret the action of any one cytokine in isolation. As part of the immune response, the production of one cytokine can often alter the release of other cytokines. A combination of cytokine results would provide more useful information about the immune response, than looking at the levels of single cytokines in isolation.

To guide the work of this chapter, and the thesis in general, a systematic review of host biomarkers in viral LRTIs was undertaken. The focus on viral LRTIs was due to the importance of viral pathogens in children, and the number of published biomarker studies made a systematic review of all biomarkers impractical.

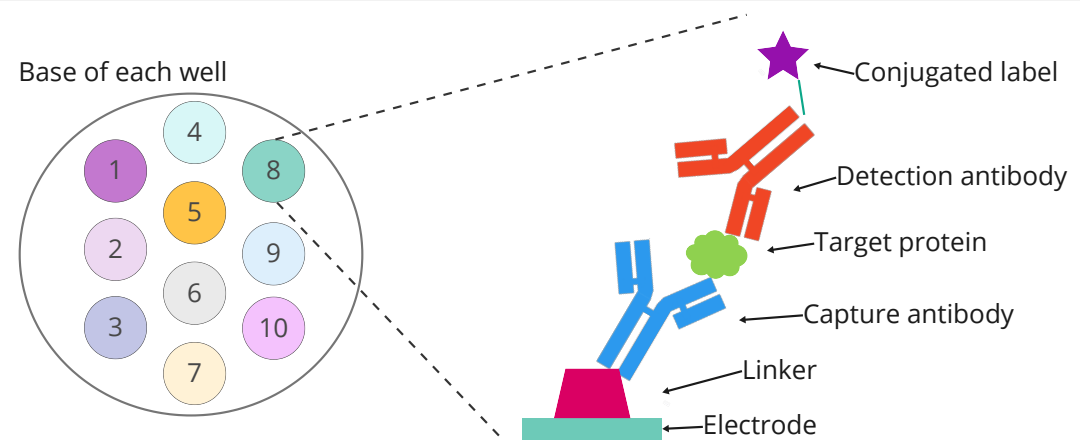


Figure 4.1: Meso Scale Discovery sandwich antibody method. The figure shows the ten electrodes at the bottom of each well in the Meso Scale Discovery U-PLEX 96-well plate. Electrode 8 is magnified and a graphical representation of one sandwich antibody configuration is shown. A linker-capture antibody complex is added to the plate, the linker then attaches to the electrode. When the target protein is added it attaches to the capture antibody. Finally, the detection antibody, with conjugated light-emitting label, attaches to the target protein. Figure created using Miro.com.

The methods and further results related to this review are summarised in Appendix A.

Following the systematic review, ten cytokines, the maximum number per MSD plate, were chosen for this experiment. MSD has a large list of possible cytokine assays. Cytokines were chosen for inclusion if they were of biological interest in LRTIs or had been reported as potential biomarkers in other disease states. Findings of particular interest were cytokine levels which could differentiate between different aetiology of disease and cytokine levels in LRTIs.

In the following paragraphs, I summarise some of the functions and relevant publications for the cytokines selected for use in this MSD experiment.

IL-6 is secreted by B and T lymphocytes, fibroblasts and macrophages. IL-6 has both pro- and anti-inflammatory roles. The primary functions of IL-6 include B cell differentiation and stimulation of acute phase proteins.[117] IL-6 has been widely studied as a biomarker in infectious disease. For example, IL-6 levels have been shown to be different in MIS-C compared with SARS-CoV-2 LRTIs.[232] In a study of adults with community-acquired pneumonia, IL-6 levels correlated well with disease severity scores.[233] In a cohort of infants with RSV bronchiolitis, higher levels of IL-6 have been associated with increased severity.[234]

IL-8, also known as C-X-C motif chemokine ligand 8 (CXCL8) is a chemotactic cytokine produced by many cell types and has a significant role in the inflammatory response. Various stimuli including viruses, lipopolysaccharide and other pro-inflammatory cytokines induce IL-8 production.[235] Studies have reported IL-8 to be elevated in similar populations to our cohort.IL-8 was elevated in a cohort of children with viral bronchiolitis who required admission to hospital.[236]

Gern *et al.* looked at IL-8 levels in paediatric LRTIs. IL-8 levels were associated with clinical disease symptoms in influenza A infections but not in RSV or parainfluenza infections.[237] IL-8 has a longer half-life to IL-6 and may be a more consistent biomarker for diagnostics.[238]

IL-10, also known as human cytokine synthesis inhibitory factor is predominantly an anti-inflammatory cytokine which inhibits the production of other cytokines. However, IL-10 can also stimulate natural killer cells and cytotoxic T cells. IL-10 is produced by several cell types including macrophages and epithelial cells within the lung.[239] Lower levels of IL-10 have been associated with increased mortality in patients with acute respiratory distress syndrome.[239] In children with RSV bronchiolitis IL-10 levels were increased in convalescent samples compared with samples taken during acute infection.[240]

IL-18 is expressed by macrophages and dendritic cells. IL-18 is a member of the IL-1 family and is generally pro-inflammatory. IL-18 requires activation by caspase. Once activated IL-18 uses the same signalling pathway as IL-1 to activate NF-kappaB and induce release of inflammatory mediators including chemokines. IL-18 is involved in activation of natural killer (NK) cells, Th1 cells, and macrophages.[241, 242] Changes in IL-18 levels have been reported in several inflammatory disease processes. In children with sepsis, IL-18 levels were higher than well children or children without signs of sepsis.[243] In animal studies, IL-18 has been shown to be important in protection against influenza A by increasing INF-gamma production and activating NK cells.[244, 245]

IL-17, also known as IL-17A, is a pro-inflammatory cytokine with a role in host defence and drives inflammation in infection that inflammatory conditions.

IL-17 is prominent in innate immunity through activation of neutrophils.[246] Changes in IL-17 have been reported in response to extracellular and intracellular bacteria including *Mycobacterium tuberculosis*.[247] Higher levels of IL-17 have been associated with reduced clinical symptoms in RSV bronchiolitis.[248]

IL-15 is a cytokine produced by many different cell types in many different tissues. In the immune response, IL-15 is produced by monocytes, macrophages and dendritic cells. When secreted IL-15 plays a role in neutrophil and NK cell activation and survival. Despite IL-15 being produced by multiple different cell types it can be difficult to detect as IL-15 is transported while bound to IL-15 receptor alpha.[249] IL-15 has been reported to have a role in the immune response to various viruses and bacteria including HIV and *M. tuberculosis*.[250, 251] In a study of children admitted with viral LRTIs, IL-15 was elevated cases versus controls; serum IL-15 levels were highest in those children who required intensive care.[252]

G-CSF, also known as colony-stimulating factor 3 is produced by different cells including macrophages, monocytes and bone marrow stromal cells. Several different stimuli can increase G-CSF production including lipopolysaccharide, TNF-alpha and IL-17.[253] G-CSF is one of the most important cytokines in controlling the differentiation and development of neutrophils.[254] G-CSF is difficult to detect in healthy individuals but levels can be markedly raised in acute infection.[120] In samples from patients with *S. aureus* and *K. pneumoniae* sepsis, G-CSF was shown to be higher in the patients with sepsis versus controls. In the same study, G-CSF and IL-6 were shown to have different levels among patients with *S. aureus* or *K. pneumoniae* infections.[255]

TNF-alpha is a major pro-inflammatory cytokine. TNF-alpha initiates a response when Gram-negative bacteria are detected, and this response leads to the production of lipopolysaccharide. TNF-alpha has been shown to have a wide variety of functions including inducing fever, activating the acute-phase response in the liver, activating the coagulation cascade, depressing cardiac function, and inducing hypoglycaemia. Due to these activities, TNF-alpha is thought to be important in septic shock.[122] In a cohort of adults with chronic LRTIs, TNF-alpha and IL-6 levels were reported to be higher in patients versus controls; the study also reported that TNF-alpha and IL-6 levels could potentially differentiate between different disease pathogens with TNF-alpha and IL-6 levels higher in *Pseudomonas aeruginosa* infections than in those with *H. influenzae*.[256]

IFN-gamma is the only type II interferon. IFN-gamma is primarily produced by immune cells including lymphocytes and NK cells. IFN-gamma production can be activated by microbial infection, tissue damage or by other cytokines including IL-18. While IFN-gamma alters other immune cells, macrophages are the main target for IFN-gamma. IFN-gamma increases phagocytosis by macrophages.[123] In a study of 50 hospitalised children with viral bronchiolitis, IFN-gamma has been reported to be lower in acute infections compared with controls; however, there was no difference in IFN-gamma levels between levels during acute infection and convalescent samples taken three to four weeks later from the same participants.[240] Aberle *et al.* reported that LRTIs caused by different viral pathogens produced different IFN-gamma levels, with parainfluenza infections having higher IFN-gamma levels compared with RSV.[257]

IP-10, also known as C-X-C motif chemokine ligand 10 (CXCL10) is a chemokine

which is secreted by IFN-gamma-stimulated cells. Neutrophils, eosinophils, monocytes and lymphocytes secrete IP-10 in response to inflammation. Secretion of IP-10 attracts various immune cells, including Th1 cells, macrophages, NK and dendritic cells, to sites of high IP-10 concentration.[258] Machado *et al.* reported that in vivo RSV infection resulted in high IP-10 levels but *S. pneumoniae* did not elicit an increase in IP-10; co-infection with RSV and *S. pneumoniae* did show an increase in IP-10 levels.[259]

In this chapter, I report the levels of ten cytokines, and how these results could be combined as a diagnostic test, in children with acute LRTIs.

4.1.5 Hypotheses

Two hypotheses are tested in this chapter:

1. Plasma cytokine levels differ between bacterial and viral LRTIs.
2. Plasma cytokine levels differ depending on the pathogen causing the LRTI.

4.1.6 Aims and objectives

Linked to hypothesis one - Plasma cytokine levels differ between bacterial and viral LRTIs This study aims to investigate if cytokine levels differ across LRTI classifications. Levels will be compared across the following groups of LRTIs:

1. Acute viral group compared with convalescent samples from the same individuals.

2. Acute bacterial group compared with convalescent samples from the same individuals.

3. Acute viral LRTIs compared directly with acute bacterial LRTIs.

A combination of cytokines will also be evaluated to differentiate between acute viral and acute bacterial LRTIs.

Linked to hypothesis two - Plasma cytokine levels differ depending on the pathogen causing the LRTI In this study I investigate if plasma cytokine levels differ depending on the pathogen causing infection. To explore this, LRTIs groups linked to RSV, influenza and *S. pneumoniae* were compared with other LRTIs.

4.2 Methods

4.2.1 Clinical study

The clinical methods are outlined in detail in Section 2.2.1, and summarised briefly below.

Study design

A prospective cohort study was undertaken at Patan Hospital, Lalitpur, Nepal. In this hospital setting, children between 2 months and 14 years of age were eligible for enrolment if they were assigned an admission diagnosis of pneumonia,

as decided by the clinical team. Children were enrolled between March 2015 and December 2017. Blood samples for RNA and protein analyses, as well as nasopharyngeal samples for molecular diagnostics, were taken on presentation to hospital. Demographics and medical record data were recorded, including the results of clinical investigations, to assist in correctly classifying the different LRTIs for further analysis. Participants were asked to return six to eight weeks after enrolment to provide a convalescent blood sample when they had recovered.

Study procedures

If the clinical team decided that the admission diagnosis was pneumonia, the research team approached the participant's parent/guardian and sought informed consent. Research samples were taken within the first 48 hours of the admission, and demographics were recorded. See Figure 2.1.

At time of discharge, data on investigations undertaken during the admission were recorded. Six to eight weeks after admission the parent/guardian were asked to return with the participant for repeat research blood samples. These convalescent samples were used as control samples.

Healthy children, taking part in a separate vaccine study were recruited to act as controls. After informed consent was obtained nasopharyngeal and blood samples were taken on one occasion only from these participants.

Research samples

At enrolment a nasopharyngeal sample and a blood sample were taken from each participant. A 3 ml blood sample was taken and 1 ml was transferred to an RNA-stabilising tube (Tempus Blood RNA Tube) and 2 ml transferred to a heparinised centrifuge tube. Research samples were transferred to the laboratory within two hours.

At convalescence the same research blood samples were taken.

4.2.2 Laboratory methods

Nasopharyngeal samples

An aliquot of the transport media from each nasopharyngeal sample was transferred to an agar plate for pneumococcal culture. If *S. pneumoniae* was found, a colony underwent serotype testing, as outlined in Section 2.2.2.

Another aliquot of transport media from each nasopharyngeal sample was frozen at -80°C and later transferred to Micropathology Ltd., University of Warwick Science Park, UK. At this UK laboratory, nucleic acid was extracted and analysed using the NxTAG™ Respiratory Pathogen Panel. This panel detects 16 viral and 3 bacterial targets, see Figure 6.1.

Blood samples

The processing of the RNA-stabilising tubes was discussed in Section 2.2.2. Plasma samples obtained from the heparinised centrifuge tube were used for the

analyses presented in this chapter.

Within four hours of blood sampling, fresh whole blood in the centrifuge tube was separated into plasma, PBMCs and a mixture of red cells and polymorphonuclear cells. The use of the PBMCs, red cells and polymorphonuclear cells is outside of the scope of this thesis and will not be discussed further.

The plasma was frozen at -80°C and initially stored at Patan Hospital. Later the plasma was transferred to the laboratories at the University of Oxford, United Kingdom. At the CCVTM, University of Oxford, 25 microlitres of plasma from each participant were used to measure cytokine levels using the MSD platform. These methods are discussed in detail in Section 4.2.4, below.

4.2.3 Classification of cases

Participants were classified into different LRTI groupings based on the likely cause of their illness. The results of routine investigations and the NxTAG respiratory PCR panel were used for classification, see Figure 2.3.

The most important groups for the analyses in this chapter are the definite viral and bacterial groups, these are described here. For a full description of case classification, see Section 2.2.3.

Participants were classified as definite bacterial if they had a culture-positive pathogenic bacteria isolated from a normally sterile site (blood or pleural fluid). The probable bacterial group included participants with high inflammatory markers (CRP >60mg/L) in the absence of a culture-positive bacteria and the absence of an LRTI-associated virus.

Participants were classified as definite viral if they had an LRTI-associated virus isolated in their nasopharynx, with low inflammatory markers (CRP $<$ 60mg/L and neutrophils $\leq$ 12 x $10^9/L$, and with no pathogenic bacteria isolated from a sterile site. LRTI-associated viruses were RSV, influenza, parainfluenza viruses and human metapneumovirus, as discussed in Section 1.4.1.

Bacterial-viral co-infections were classified as those who met the criteria for the definite bacterial group, and who also had an LRTI-associated virus isolated from their nasopharynx. Secondary groupings were used to look at the individual pathogens, RSV, influenza and *Streptococcus pneumoniae*.

4.2.4 Meso Scale Discovery panel

The ten cytokines included in the MSD panel were IFN-gamma, IL-6, IL-8, IL-10, IL-15, IL-17, IL-18, TNF-alpha and G-CSF, see Figure 4.2.

Where sample volume allowed, samples from the same participants who underwent RNA-seq were selected for MSD cytokine panel, see Section 6.2.9 for more details on sample selection. This was to allow for comparisons between the RNA and protein results, see Chapter 5 for comparisons across platforms.

The acute samples included the following classifications: definite viral, viral syndrome with high inflammatory markers, unknown class, bacterial-viral co-infection, probable bacterial and definite bacterial LRTIs. Full details of how the classification system can be found in Section 2.2.3.

A selection of convalescent samples was included as comparison groups for the analysis. If a selected acute sample had a paired convalescent sample available,

this sample was included for cytokine analysis. Healthy control samples were also included to help to adjust for any experimental differences across the different MSD plates used.

Laboratory protocol

A step-by-step protocol is available in Appendix D. The main steps in using the U-PLEX MSD platform are summarised in the following sections. This laboratory protocol was followed for each of the plates in this experiment.

Create individual U-PLEX linker-coupled antibody solutions The ten biotinylated capture antibodies and the ten different U-PLEX linker solutions were used in this first step. Each biotinylated capture antibody was combined with a unique linker and the antibody identity was recorded next to the linker number on the spot map, see Figure 4.2.

Prepare the multiplex coating solution and coat the plate All ten U-PLEX linker-coupled antibodies were combined to create the multiplex coating solution. Fifty microlitres of the multiplex coating solution was added to each well of the U-PLEX 96-well plate. The plate was incubated, with shaking, for one hour. The plate was washed three times with MSD wash buffer. At this stage, the coated plate was ready for use.

Prepare the calibrator standards Three calibrators were provided with each MSD kit. The diluent supplied was added to each calibrator. The three

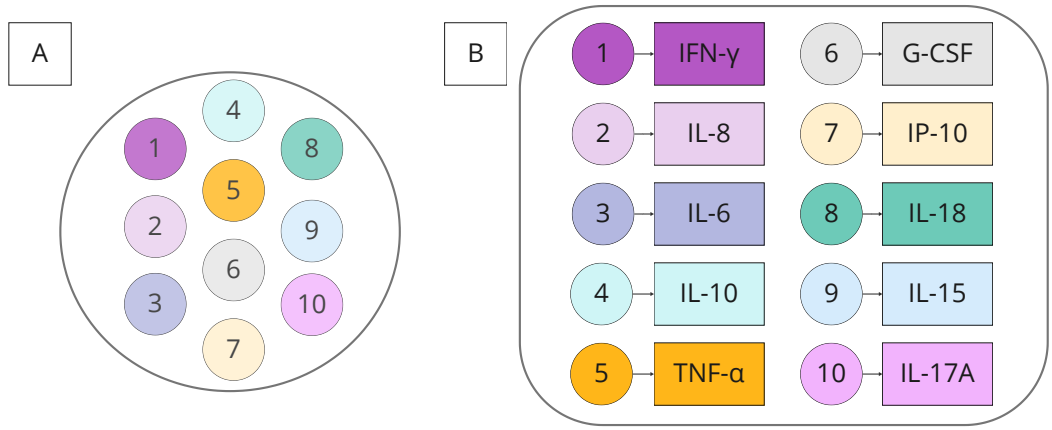


Figure 4.2: Panel A shows the spot map representing the ten electrodes at the base of each well of the U-PLEX 96-well plate. Each plate contained 96 wells, each with the same stop map distribution. Panel B shows the specific cytokine linked to each numbered electrode at the base of each well for this experiment.

calibrators were combined. The reconstituted calibrator solution was allowed to equilibrate at room temperature for 15-30 minutes. The calibrator solution was serially diluted to create the eight calibrator standards used to generate the standard curve. Each calibrator solution was combined with three-parts diluent to create a four-fold dilution between each step on the calibrator curve. The final calibrator standard (lowest point on the curve) contained no calibrator solution.

Add calibrators and samples to plate A diluent volume of 25 µL was added to each well. Then 25 µL of the different calibrator standards were added to the wells selected to create the standard curve. Then 25 µL of the plasma samples from the participants were added to the wells designated for participant samples on the U-PLEX MSD plate. The plate was then sealed and incubated at room temperature, with shaking, for one hour.

Prepare the detection antibody solution Each detection antibody solution is designed to detect a specific cytokine. The ten detection antibody solutions were combined in a single tube with the specified amount of diluent.

Wash plate and add detection antibody solution The plate was washed three times with the MSD wash buffer. A volume of 50 µL of the combined detection antibody solution was then added to each well. The plate was sealed and incubated again at room temperature for one hour while shaking.

Wash and read plate. The plate was again washed three times with the MSD wash buffer. A volume of 150 µL of MSD GOLD Read Buffer B was then added to each well. The plate was placed in the MSD reader instrument, Meso QuickPlex SQ 120MM. The MSD instrument read the electrochemiluminescent signals generated in each well to estimate the amount of each cytokine present. The actual photon count from each protein was the reported signal. These formed the raw output results. The MSD Methodical Mind program which accompanies the MSD reader was used to interpret the results. Methodical Mind is a software system that allows users to capture, process and analyse assay data from MSD plates.[260] When the Methodical Mind program was finished running, the raw results from the MSD plate were available in a .txt file.

Samples included per plate

In each 96-well plate, the 8 standards were included in duplicate across 16 wells. Each plate also included 40 participant samples, in duplicate. The experiment

was designed so that each plate had a combination of acute, convalescent and healthy control samples; within the acute samples there was a mix of viral and bacterial LRTIs. The same samples from healthy controls were tested on multiple plates, this was to allow for inter-plate validation and to adjust for any batch effect across the different plates.

Quality control of results

MSD Discovery Workbench 4.0 analysis software is a Windows application which allows the QC and analysis of MSD plate data.[260] The raw MSD plate results were loaded into Discovery Workbench. A template was built to use in assessing the quality of the results. The template included the plate layout used in the experiment and the proteins that were assigned to each spot at the base of the wells, Figure 4.2, Panel B. The template also included the following QC rules. Results were included in the final analysis if:

1. The standard values produced an acceptable standard curve.

    (a) Values for standards were only included if the measured concentration matched the expected concentration, percentage recovery between 80% and 120%. This expected concentration was calculated from the known concentration of the initial calibrator solutions.

2. The two duplicate results agreed with each other, within an acceptable error range, coefficient of variation (CV) less than 50%.

The CV is the standard deviation divided by the mean and multiplied by 100.[260]

In addition to the above rules, if samples were outside of the calculated reference range the following rules were applied:

1. For results below the reference range – These results were converted to half the lower limit of detection (LLOD) for each analyte.

2. For results above the reference range – Following discussion with the scientific team at MSD, these results were deemed to be unreliable and were excluded from analyses.

4.2.5 Analysis of cytokine panel results

MSD Discovery Workbench generated a results report for the experiment. This report includes the raw electrochemiluminescent signals, the adjusted electrochemiluminescent signals and the calculated concentrations in picograms per millilitre (pg/ml). The concentration was calculated by comparing the electrochemiluminescent signal to the standard curve.

After results which failed QC were removed, the results table was uploaded to RStudio for analysis. These analyses were carried out using R version 4.3.1.The calculated concentration in pg/ml was used as the main measure for the analyses. The mean value of the two duplicates was used.

Samples with more than 50% of analyte results excluded were deemed low-quality samples. Low-quality samples were repeated if there was sufficient sample volume to do so. If low-quality samples did not have sufficient sample volume to be repeated these sample results were excluded from further analyses.

Normalisation and quality control plots. After assessing the cytokine results for skewness, and confirming that the data were skewed, normalisation was carried out using a log2 transformation. PCA was carried out looking for outlier samples and factors which might bias the results including differences between experimental plates, age and sex. Samples with incomplete results were excluded from PCA as complete results are required to calculate the principle components. Further outlier examination was performed by creating box plots of mean results per sample.

Batch correction. The experimental plate was found to be a confounding factor. A batch correction for the experimental plate was performed using the ComBat function in R. ComBat is an algorithm that adjusts for batch effects using an advanced modification of mean centering.[207] PC biplots were repeated post correcting for experimental plate to ensure the batch correction step was successful, and that there were no other confounding factors seen post-batch correction.

Fitting the linear model. Any confounding factors, found during PCA, were included in a linear model. An empirical Bayes approach was used to apply this model and correct for confounding factors. This approach is common in high dimensional data but has also been applied to lower dimension data such as cytokine panels.[261, 262, 263]

Comparisons of interest

To test the hypotheses listed in Section 4.1.5, comparisons were made between the different LRTI classifications of interest, see Table 4.1. Box plots were used to visualise these data. Mann-Whitney U tests were used for statistical testing between groups of interest across the ten different cytokines. Results were adjusted for multiple testing using the BH Procedure.[264]

For each of the box plots comparing LRTI groups in this chapter, the 50th (median) and 25th and 75th centiles are represented by the box plot, the line above and below each box extends to 1.5 times the length of the inter-quartile range (IQR). Individual results are shown separately as dots. The horizontal variation of dots within each category is determined randomly to improve visualisation. Mann-Whitney U tests were used to look for statistical significance between the groups, results were adjusted for multiple testing using the Benjamini-Hochberg Procedure.

4.2.6 Cytokine panel to differentiate between bacterial and viral lower respiratory tract infections

To determine if the combined results from the cytokine panel described in this chapter can be used to differentiate accurately between bacterial and viral LRTIs, the following steps were performed.

Bacterial and viral samples were selected, as per the results of the semi-supervised approach taken in Section 2.3.3. A 70:30 split was used to divide the data into training and test datasets.

Table 4.1: Comparison across different LRTI groups used to test hypotheses in the MSD cytokine panel experiment. * Bacterial and viral groups are the groups defined using the semi-supervised approach in Section 2.3.3

Hypothesis	LRTI Group of Interest*	Contrast
Plasma cytokine levels differ between bacterial and viral LRTIs	Acute viral	Convalescent samples from the same participants (convalescent viral)
	Acute bacterial	Convalescent samples from the same participants (convalescent bacterial)
	Acute bacterial	Acute viral
	Acute co-infections	Acute bacterial/acute viral groups
Plasma cytokine levels differ depending on the LRTI pathogen	Acute pneumococcal	Convalescent samples from the same participants (convalescent pneumococcal)
	Acute pneumococcal	Acute non-pneumococcal bacterial
	Acute RSV	Convalescent samples from the same participants (convalescent RSV)
	Acute RSV	Acute non-RSV viral
	Acute influenza	Convalescent samples from the same participants (convalescent influenza)
	Acute influenza	Acute non-influenza viral

Training dataset Cytokines were ranked in order of importance using a PLS approach. A variable importance measure was calculated based on the weighted sums of the absolute regression coefficients.[210]

Using the caret function in R, a correlation matrix was created to look for highly correlated cytokines, Pearson's correlation coefficient >0.75.

Recursive feature elimination was performed using 10-fold cross-validation, to remove potentially redundant features.[211] Two different cytokine signatures for differentiating bacterial from viral LRTIs were identified and taken forward for testing.

Performance statistics were calculated for the different protein signatures. Sensitivity, specificity, ROC-AUC, positive and negative likelihood ratios were calculated for the training set initially. As a test which has a high sensitivity

to detect bacterial infections would be more useful clinically, I restricted the sensitivity range to 90% to 100% and re-calculated the AUC. This pAUC showed how well the model was able to predict a correct result when the required true positive rate was high.[168]

The performance of the selected cytokine signatures was then evaluated in the test dataset.

4.3 Results

4.3.1 Results included for analysis

Forty samples, in duplicate were included per plate, nine plates were included in the experiment. Figure 4.3 shows the different quality control steps and the number of cytokine results and samples after each step.

The standards for IL-8 failed to produce an acceptable standard curve on two plates. All the IL-8 results from these two plates were excluded from further analyses, Plates 5 and 6. The issue with the failure of the IL-8 standards was related to the linker antibody provided with one of the kits. Due to the limited sample volume and the number of kits available for the experiment, these failed IL-8 results were not repeated on a new kit. One convalescent sample was an outlier on PCA and this sample was excluded. Twenty-eight low-quality samples were removed.

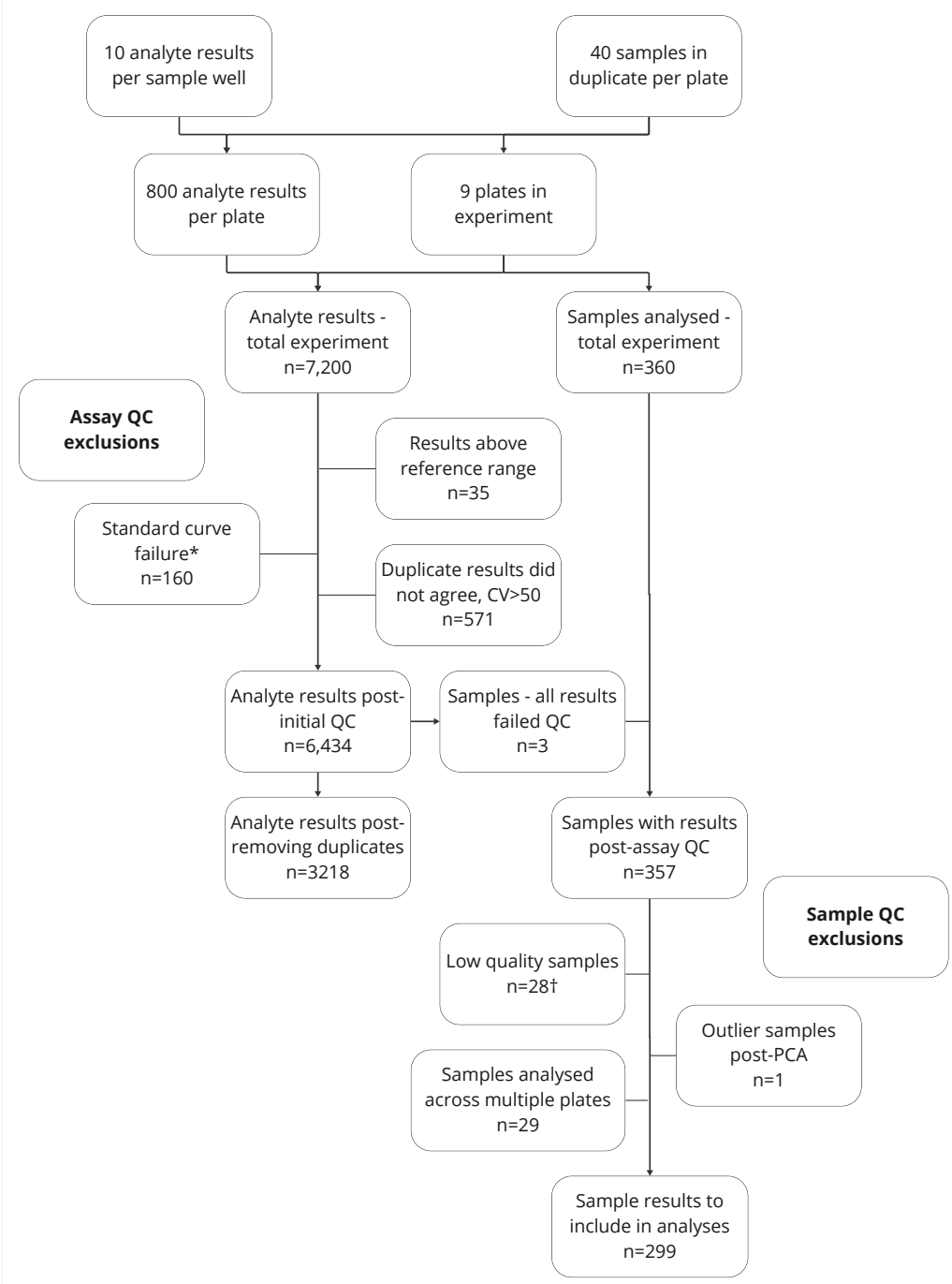


Figure 4.3: Number of samples and number of cytokine results included after each QC step in the MSD cytokine panel experiment. *The standard curve for IL-8 failed on two plates. †Low quality samples where over 50% of results failed QC were excluded. QC, quality control; CV, coefficient of variation ([standard deviation/mean]*100)

Table 4.2: Demographic and key clinical/laboratory data for the participants with cytokine results included after quality control steps, grouped by LRTI classification. CRP, C-reactive protein; WCC, white cell count.

	Definite viral	Viral syndrome - high inflammatory markers	Unknown	Co-infection (bacterial-viral)	Probable bacterial	Definite bacterial	Convalescent	Healthy controls
Number of cases	70	20	36	5	53	16	67	32
Age ,years (median)	0.8	1.5	1.5	0.7	4.2	5.6	1.7	0.8
Female (%)	37%	45%	36%	40%	42%	25%	43%	72%
Duration of hospital stay (median)	6	7	4	13.5	7	6	NA	NA
CRP (mg/l) (median)	11.8	102.1	16.2	117.7	157.5	154.1	NA	NA
WCC x10^9/L (median)	10.9	16.4	17.6	8.9	21.3	18.9	NA	NA
Neutrophils x 10^9/L (median)	5.4	12.3	8.8	3.2	16.6	15.6	NA	NA

4.3.2 Description of cohort

Following quality control and removal of samples repeated across multiple plates, there were samples from 299 participants with results included for further analyses. These samples were made up of 200 acute, 67 convalescent and 32 healthy control samples.

The acute samples were obtained from children early in their admission with signs of LRTI. The samples were divided into LRTI classifications, as described in Section 5.2.3. Table 4.2 shows basic demographics and clinical data for these participants. Of note, the median age in the bacterial groups is higher than the viral groups; this has been adjusted for in the analyses.

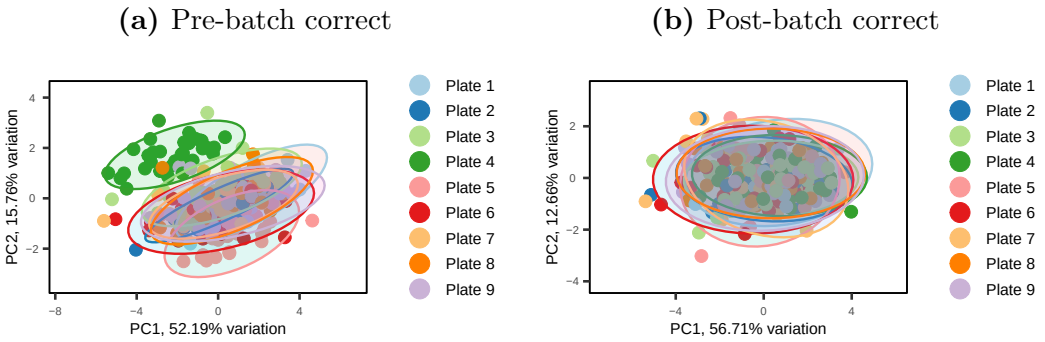


Figure 4.4: PCA of samples complete results when IL-8 results are excluded from all plates. Principal component biplot pre-batch correction, panel A and post-batch correction, panel B coloured by experimental plate. PC1 and PC2 plotted.

4.3.3 Principal component analysis

PC biplots were examined looking for variables which should to adjusted for in the analyses.

When samples with missing analyte results are excluded, 200 samples with complete results were included in the PCA. A PC biplot coloured by experimental plate showed that the results from Plate 4 separated from the other results, Figure 4.4a. There was no known technical or biological reason for Plate 4 to produce differing results to the other plates. Batch correction using the ComBat function was carried out and the PC biplot was repeated, Figure 4.4b. The results from all of the plates now overlap, as expected.

PCA did not highlight any other issues with the data, see Appendix D for other PC plots.

4.3.4 Cytokine levels across different acute lower respiratory tract infections

To address the first hypothesis that cytokine levels differ depending on LRTI aetiology, results from different acute LRTI groups were compared. To give an overview of the cytokine results before looking at specific group comparisons, Figure 4.5 shows the cytokine levels across all of the different LRTI classifications. The co-infection group includes only five acute samples and has heterogeneous results across some of the cytokines. It would be difficult to obtain any other meaningful results from this group. No further analyses will be carried out on the co-infection group. The co-infection cases will be included with the bacterial group when testing models to differentiate between bacterial and viral infections in Section 4.3.7 but co-infections have been kept separate for the results presented in the next section, Section 4.3.5.

4.3.5 Cytokine levels in bacterial and viral groups

First, results were compared between the acute and convalescent samples in the bacterial and viral groups separately. This allowed for identification of cytokines that were elevated in acute bacterial and viral infections. Then I directly compare the acute bacterial and acute viral results, these results were adjusted for the age differences across the bacterial and viral groups.

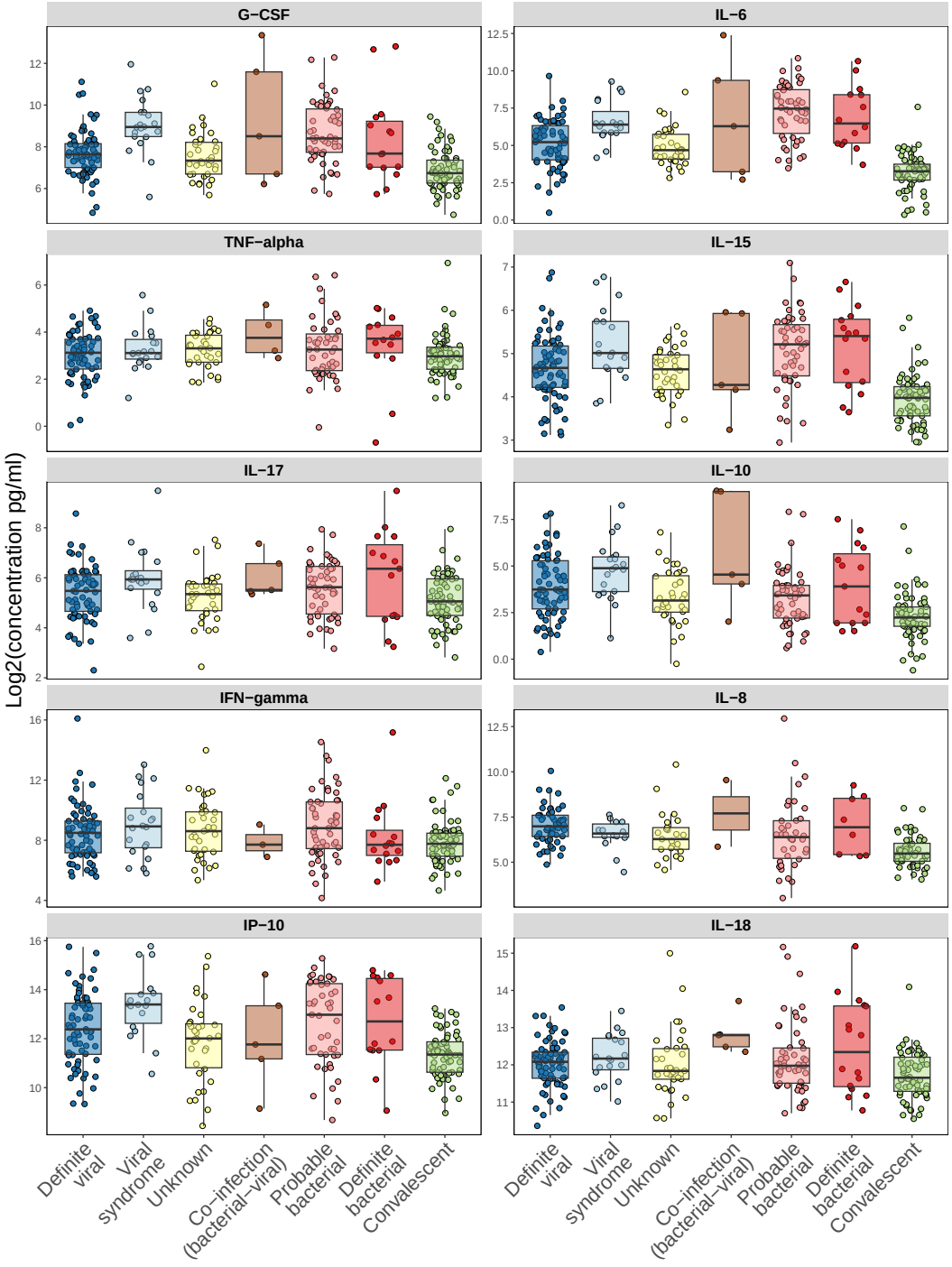


Figure 4.5: Results from the acute samples for each of the ten cytokines tested, across each of the study groups. All convalescent samples are grouped together and shown in the final column. The healthy control group, used to correct for experimental batch effects across the plates, were not included in these plots.

Acute viral LRTIs compared with convalescent samples

To examine the cytokine profile in viral LRTIs, the acute viral group was compared with samples from the same participants in convalescence.

Seventy acute viral and 36 convalescent samples were included in this comparison. Figure 4.6 shows the cytokine levels in the acute viral and convalescent groups. After adjusting for multiple testing seven out of ten analytes had significantly higher results in the acute viral samples. The differences between TNF-alpha, IL-17 and IFN-gamma results across the groups failed to reach statistical significance.

Acute bacterial LRTIs compared with convalescent samples

Cytokine levels were compared between the acute bacterial group and convalescent samples from these same individuals, see Figure 4.7. The bacterial group included all definite and probable bacterial cases. Sixty-nine acute bacterial and 15 convalescent samples were included in this comparison. After adjusting for multiple testing eight out of ten analytes had significantly higher results in the acute bacterial samples. The differences between TNF-alpha and IL-18 results across the groups failed to reach statistical significance.

Acute viral compared with acute bacterial lower respiratory tract infections

Cytokine results from acute definite viral LRTIs were directly compared with results from acute bacterial LRTIs, probable and definite bacterial groups.

Sixty-nine acute bacterial and 15 convalescent samples were included in this

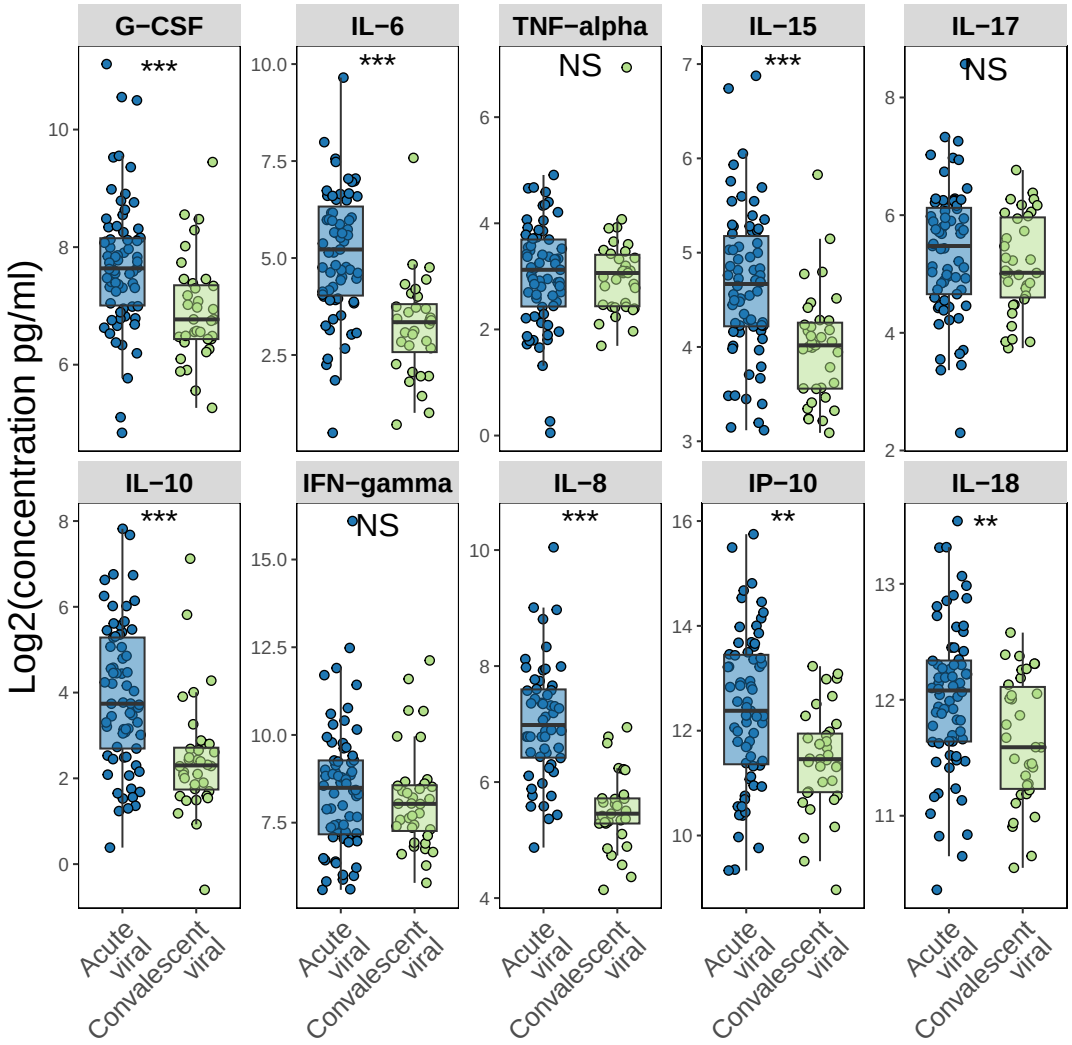


Figure 4.6: Results from the acute viral LRTI group, n=70, for each of the ten cytokines tested, compared with convalescent samples from the same individuals, n=36. *adjusted p-value <0.05; **adjusted p-value <0.01; ***adjusted p-value <0.001; NS, non-significant, adjusted p-values >0.05.

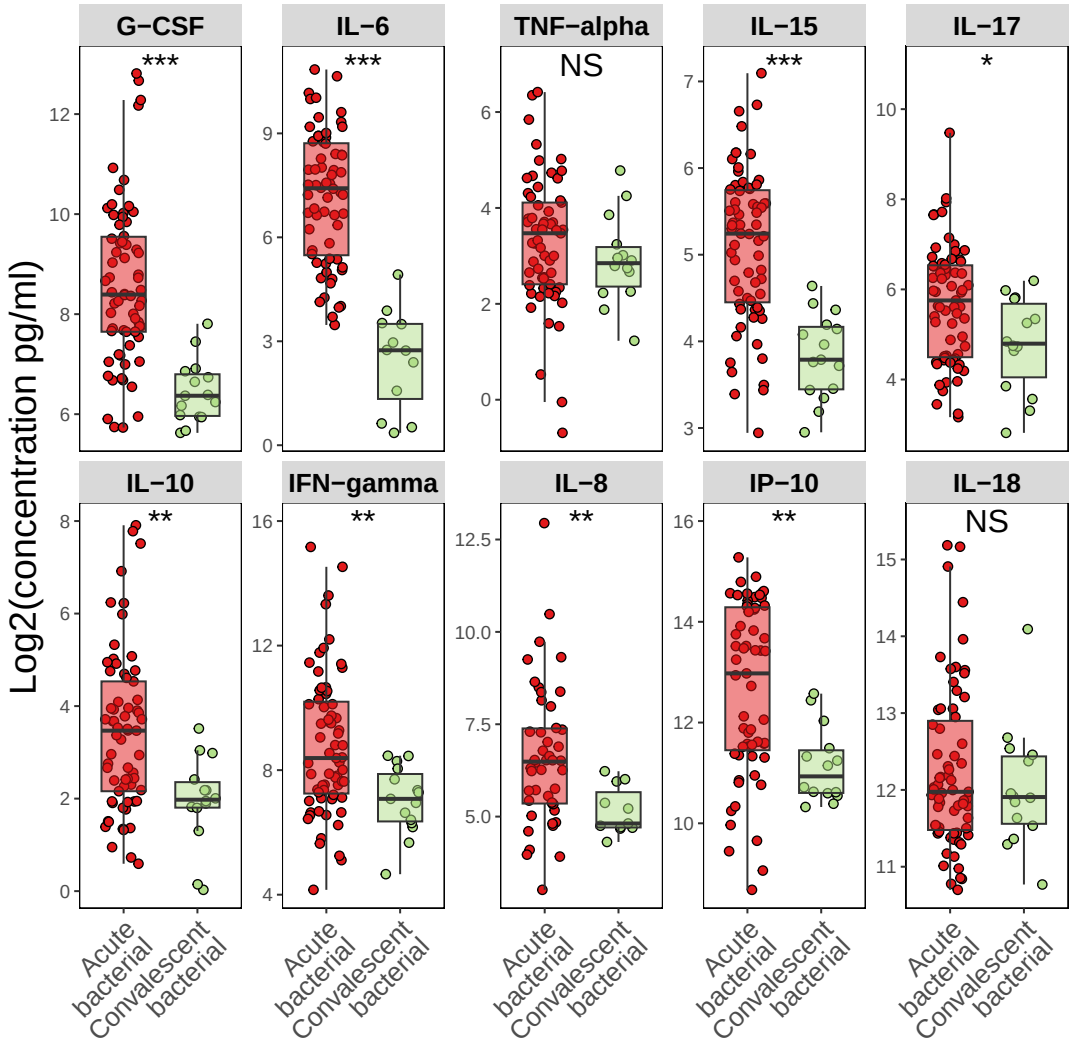


Figure 4.7: Results from the bacterial viral LRTI group, n=69, for each of the ten cytokines tested, compared with convalescent samples from the same individuals, n=15. *adjusted p-value <0.05; **adjusted p-value <0.01; ***adjusted p-value <0.001; NS, non-significant, adjusted p-values >0.05.

comparison. Figure 4.8 shows box plots for results from each of the ten cytokines, divided into acute bacterial and acute viral groups. After adjusting for multiple testing, three out of ten analytes had significantly higher results in the acute bacterial samples. G-CSF, IL-6 and IL-15 were significantly higher in the acute bacterial group compared with the acute viral group.

When acute definite viral LRTIs were compared with acute definite bacterial LRTIs only, the pattern of results was similar to the above comparisons, however, none of the cytokine level differences between groups reached statistical significance. See Appendix D.

4.3.6 Pathogen groups

To investigate if the cytokines results varied by different pathogens, samples were grouped by their pathogen classification. RSV, influenza and *S. pneumoniae* are the most common pathogens found in the cohort and these three pathogen groups are examined in the following sections.

Acute RSV LRTIs compared with other acute LRTIs. Acute RSV LRTIs were compared with acute samples from LRTIs of other aetiology. The LRTIs of other aetiology included bacterial and other viral LRTIs. Participants from the unknown group were excluded as these could have included RSV-associated LRTIs. Forty-nine acute RSV and 109 other aetiology samples are included in this comparison, see Figure 4.9. After adjusting for multiple testing three out of ten cytokines had significantly higher results in the acute bacterial samples. G-CSF, IL-6 and IL-15 were significantly higher in the acute other aetiology

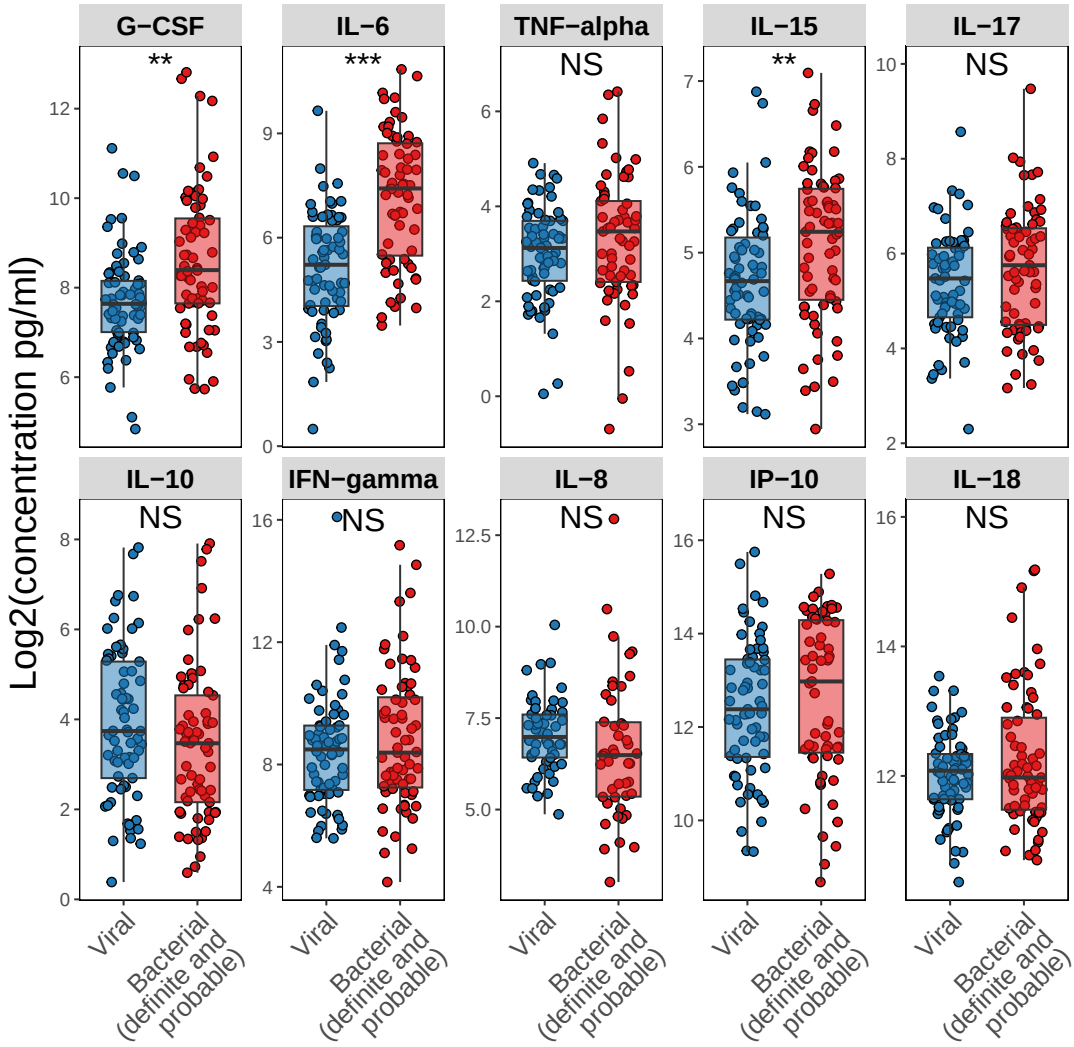


Figure 4.8: Results from the viral LRTI group, n=70, for each of the ten cytokines tested, compared with the bacterial LRTI group, n=69. *adjusted p-value <0.05; **adjusted p-value <0.01; ***adjusted p-value <0.001; NS, non-significant, adjusted p-values >0.05.

samples compared with the acute RSV samples. These were the same three cytokines that were significantly elevated when the acute bacterial group was compared with the viral group. The results of the comparison between RSV and other aetiology samples, likely reflects the high percentage of bacterial samples in the other aetiology group (69/109, 63%). For this reason, I excluded the bacterial samples and repeated the comparison between RSV and other acute viral LRTIs to look for cytokines which are significantly different in RSV LRTIs compared with other viral pathogens. The other viral group included influenza, parainfluenza and human metapneumovirus-associated LRTIs.

Forty-nine acute RSV and 35 other viral aetiology samples were included in this comparison. Figure 4.10 shows box plots for each of the ten cytokines with the samples grouped into acute RSV and other viral aetiology samples. After adjusting for multiple testing, no analytes had significantly different results between the groups.

Acute influenza LRTIs compared with other acute LRTIs. When comparing the acute influenza LRTI group with other viral and bacterial groups combined, there were no significant differences across the ten cytokines. Just as with the RSV results, there were no significantly different cytokine results when acute influenza LRTIs were compared with samples from acute other viral LRTIs. See Appendix D for these results.

Acute pneumococcal LRTIs compared with other acute LRTIs. The acute pneumococcal group included samples which were classified as definite pneumococcal (culture-positive from a normally sterile site) and probable pneu-

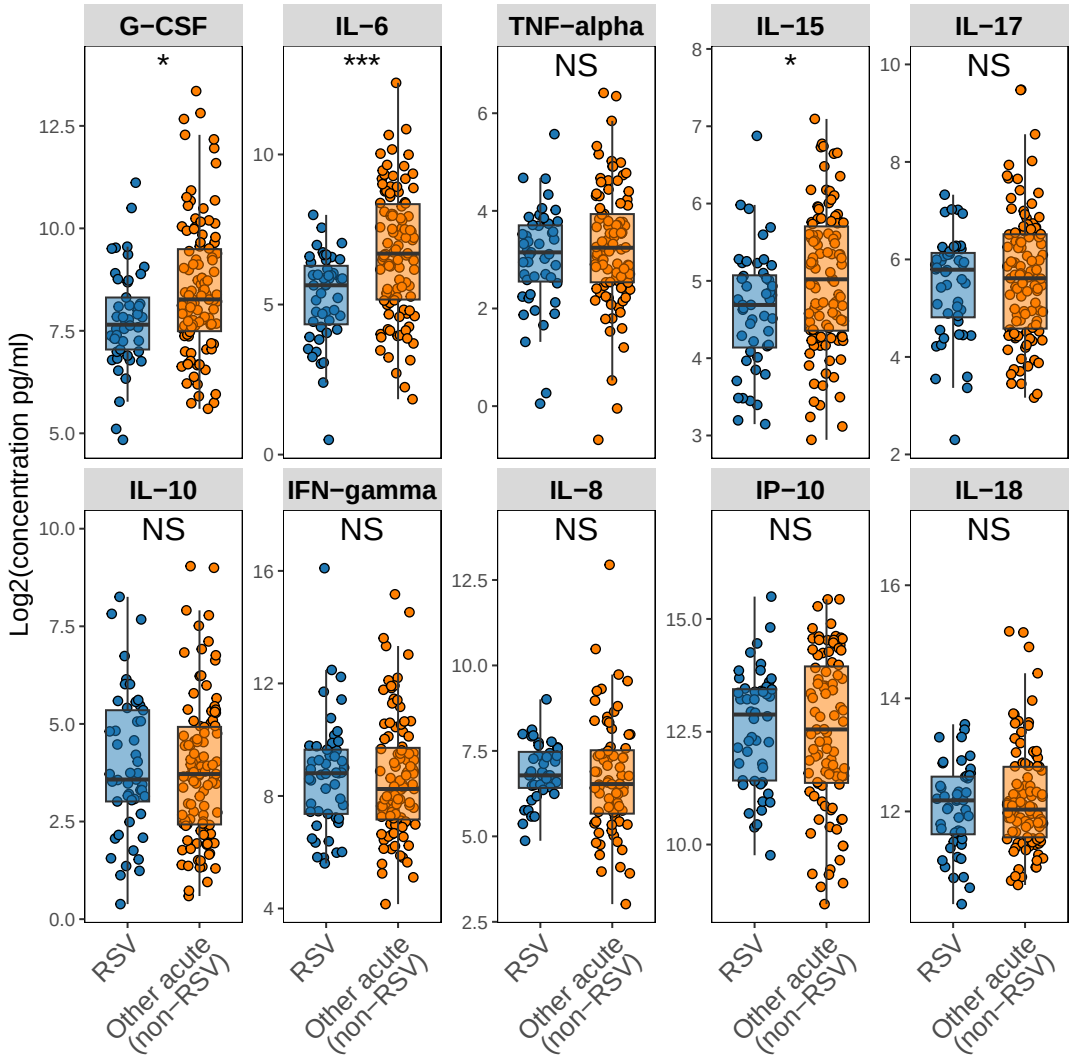


Figure 4.9: Results from the acute RSV LRTI group, n=49, for each of the ten cytokines tested, compared with LRTI samples from participants with other acute classifications, n=109. *adjusted p-value <0.05; **adjusted p-value <0.01; ***adjusted p-value <0.001; NS, non-significant, adjusted p-values >0.05.

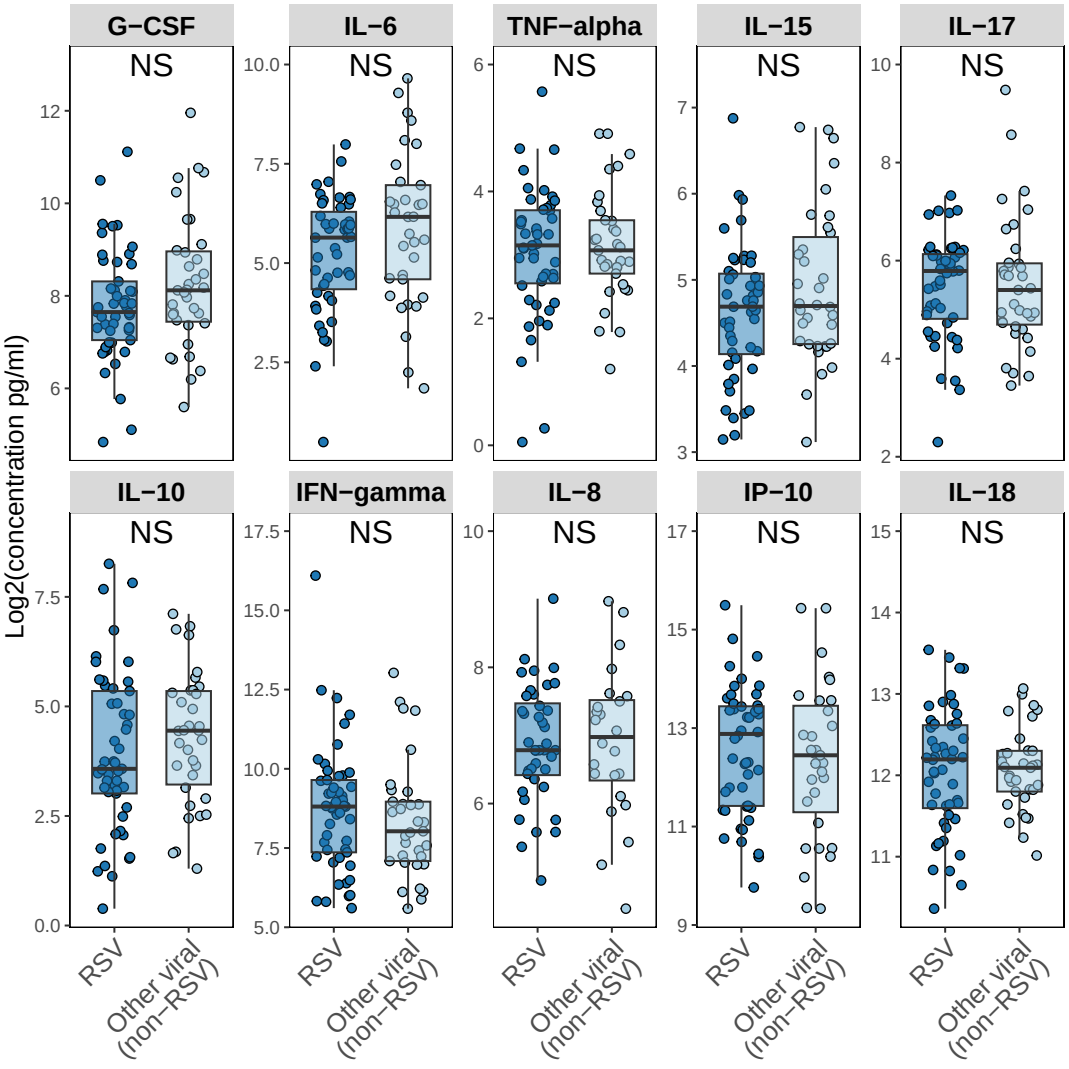


Figure 4.10: Results from the acute RSV LRTI group, n=49, for each of the ten cytokines tested, compared with LRTI samples from participants with other acute viral classifications, n=35. *adjusted p-value <0.05; **adjusted p-value <0.01; ***adjusted p-value <0.001; NS, non-significant, adjusted p-values >0.05.

mococcal (probable bacterial cases with pneumococcal serotype 1 or 5 isolated from their nasopharynx. The justification for using these serotypes is explained in Section 2.2.3. There were no significant differences across the ten cytokines when results in the acute pneumococcal group were compared with results from all other acute bacterial and acute viral samples.

The acute pneumococcal group were then compared with results from other acute bacterial samples only. Again, there were no significant differences found across any of the ten cytokine results. See Appendix D for more details.

4.3.7 Cytokine levels to differentiate between bacterial and viral lower respiratory tract infections

To examine if this cytokine panel could be used to differentiate between bacterial and viral LRTIs I used a similar approach to that used in the RNA and protein model creation in Chapters 2 and 3. The groups of bacterial and viral samples that were selected using the semi-supervised approach in Section 2.3.3 were again used to train the model. These groups were split into training and test datasets, in a 70:30 ratio. In the training dataset, there were 49 viral and 38 bacterial cases. In the training dataset, there were 20 viral and 15 bacterial cases. Included in the bacterial group were three co-infections in the training dataset and two co-infections in the test dataset.

Table 4.3: Correlation matrix showing Pearson's correlation coefficient (r) between the cytokine levels considered for the bacterial versus viral protein signature. No cytokine correlations reached the r >0.75 threshold.

	G-CSF	IL-6	TNF-alpha	IL-15	IL-17	IL-10	IFN-gamma	IL-8	IP-10	IL-18
G-CSF	1	0.62	0.35	0.56	0.44	0.35	0.37	0.05	0.37	0.23
IL-6	0.62	1	0.35	0.48	0.30	0.25	0.18	0.15	0.04	0.23
TNF-alpha	0.35	0.35	1	0.59	0.54	0.38	0.38	0.43	0.19	0.46
IL-15	0.56	0.48	0.59	1	0.69	0.39	0.52	0.16	0.38	0.49
IL-17	0.44	0.30	0.54	0.69	1	0.42	0.55	0.19	0.49	0.37
IL-10	0.35	0.25	0.38	0.39	0.42	1	0.47	0.42	0.19	0.39
IFN-gamma	0.37	0.18	0.38	0.52	0.55	0.47	1	0.19	0.51	0.30
IL-8	0.05	0.15	0.43	0.16	0.19	0.42	0.19	1	0.11	0.30
IP-10	0.37	0.04	0.19	0.38	0.49	0.19	0.51	0.11	1	0.12
IL-18	0.23	0.23	0.46	0.49	0.37	0.39	0.30	0.30	0.12	1

Feature selection

When selecting cytokines for inclusion in the model, a correlation matrix was created looking for highly correlated cytokines, see Table 4.3. No cytokine comparisons were above the correlation threshold, Pearson's correlation >0.75.

The cytokines were then ranked in order of importance by the sums of the absolute regression coefficients, using a PLS approach. IL-6 had the highest importance followed by G-CSF and IL-15. The weights calculated are a function of the reduction of the sum of the squares across the PLS components, and are calculated separately for each outcome.[211]

Recursive feature elimination was used to determine the optimal number of features to include in the model. Possible confounding factors were included such as experimental batch, age and sex.

An eight-cytokine model had the highest accuracy, with age included as a known

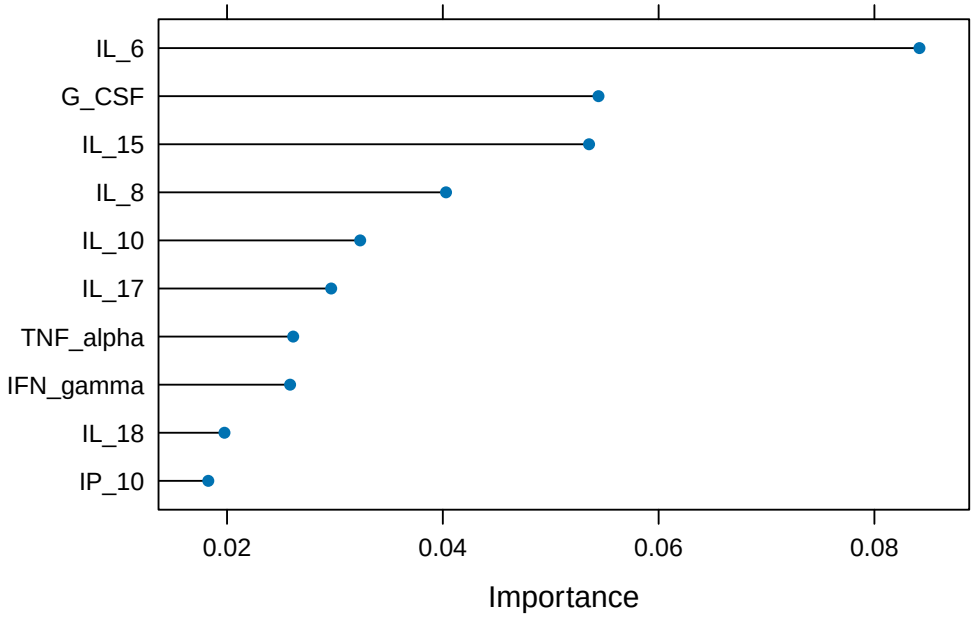


Figure 4.11: The ten cytokines included in the MSD panel, and considered for inclusion in a model to differentiate between bacterial and viral LRTIs, using a partial least squares (PLS) method. The cytokines are ranked by importance which is calculated as the weighted sums of the absolute regression coefficients. The weights calculated are a function of the reduction of the sum of the squares across the PLS components, and are calculated separately for each outcome.

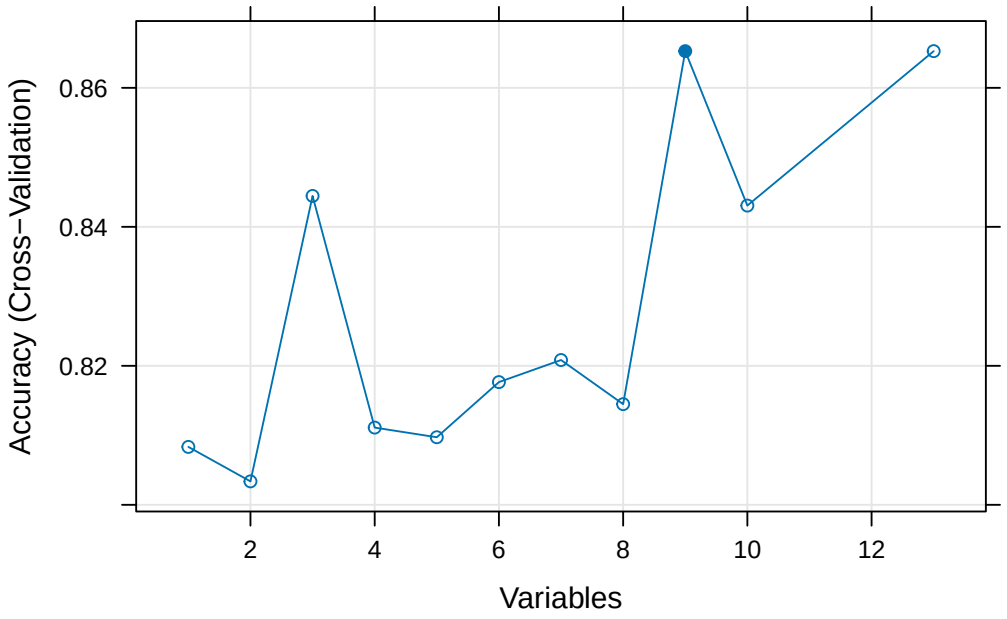


Figure 4.12: Using recursive feature elimination to determine the accuracy of different models and the ideal number of variables to include, considering all ten cytokines. A model including nine features, age and eight cytokines, had the highest accuracy level. The ten-cytokines, age, sex and batch were considered for inclusion in the model.

confounder. However, there was only a marginal difference in accuracy between a two-cytokine model and the eight-cytokine model. Two models were brought forward for testing. The eight-cytokine model excluded IL-10 and IFN-gamma. The two-cytokine model included IL-6 and G-CSF. See Figure 4.12 for the accuracy of models with different numbers of variables included.

The accuracy of cytokine models without including age was also tested. These models performed marginally worse and these results can be found in Appendix D.

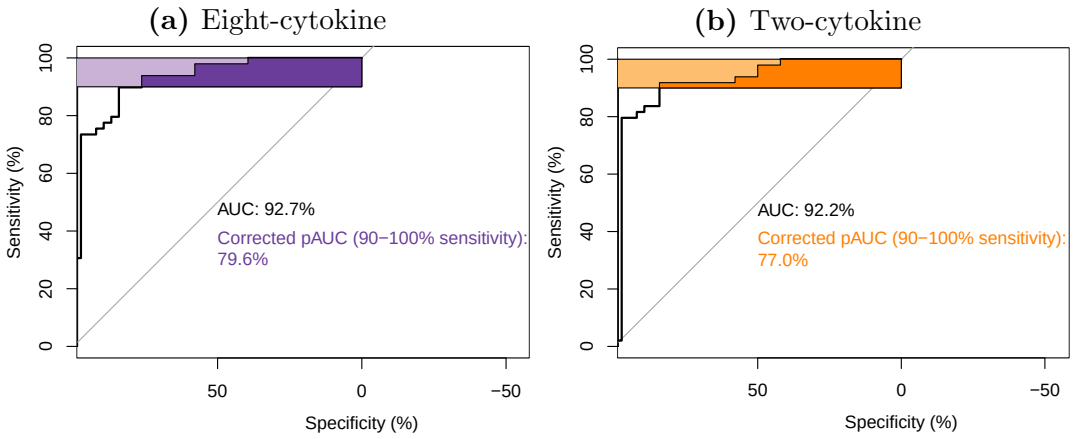


Figure 4.13: Receiver operating characteristic curves for two different cytokine models to differentiate between the acute bacterial and acute viral groups, using the training dataset. Both models include age as a covariate. The area under the curve (AUC) presented for each with a partial AUC at sensitivity range from 90% to 100% also shown.

Model performance - Training dataset

In the training dataset, the eight-cytokine and two-cytokine models performed similarly when differentiating between bacterial and viral LRTIs. The specificity was the same for both models, 86%. The two-cytokine model had a slightly higher sensitivity, 84%, and specificity, 92%, when compared with the eight-cytokine model, 82% sensitivity and 90% specificity, see Table 4.4.

The AUC for the ROC curves was similar for both models, 92.7% for the eight-cytokine model and 92.2% for the two-cytokine model. When the AUC was restricted to a high-sensitivity range, 90%-100% the pAUC was 79.6% for the eight-cytokine model and 77% for the two-cytokine model, see Figure 4.13.

Table 4.4: Cytokine model performance statistics in the training data. Comparison of results when an eight-cytokine and a two-cytokine model are used to differentiate between bacterial and viral LRTIs. Age was included in both models to adjust for this confounding factor.

Performance metric	**Eight-cytokine signature**		**Two-cytokine signature**	
Proteins included	IL-6, G-CSF, IL-15, IL-8, IL-17, TNF-alpha, IL-18, IP-10		IL-6, G-CSF	
	Reference - bacterial	**Reference - viral**	**Reference - bacterial**	**Reference - viral**
Predicted bacterial	31	5	32	4
Predicted viral	7	44	6	45
Accuracy (95% CI)	0.86 (0.77, 0.92)		0.89 (0.80, 0.94)	
Sensitivity (95% CI)	0.82 (0.66, 0.92)		0.84 (0.69, 0.94)	
Specificity (95% CI)	0.90 (0.78, 0.97)		0.92 (0.8, 0.98)	
Positive predictive value (95% CI)	0.86 (0.72, 0.94)		0.89 (0.76, 0.95)	
Negative predictive value (95% CI)	0.86 (0.76, 0.93)		0.88 (0.8, 0.94)	

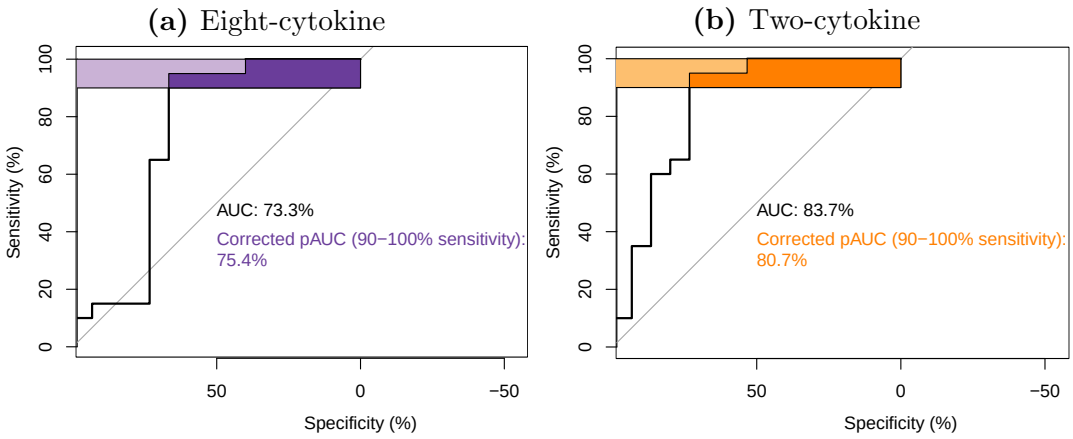


Figure 4.14: Receiver operating characteristic curves for two different cytokine models to differentiate between the acute bacterial and acute viral groups, using the test data set. The area under the curve (AUC) presented for each with a partial AUC at a sensitivity range from 90% to 100% also shown.

Model performance - Test dataset

The performance of both models was measured in the test dataset. The numbers of cases correctly predicted as bacterial and viral were similar for both models. The eight-cytokine model performed slightly better with a sensitivity of 67% and a specificity of 95% compared with a sensitivity of 60% and a specificity of 95% for the two-cytokine model, see Table 4.5.

The AUC for the ROC curves were AUC of 73.3% for the eight-cytokine model, and 83.7% for the two-cytokine model. Restricting the AUC to the sensitivity range of 90% to 100% resulted in a pAUC of 75.4% for the eight-cytokine model, and a pAUC of 80.7% for the three-cytokine model, see Figure 4.14.

Table 4.5: Cytokine model performance statistics in the test data. Comparison of results when an eight-cytokine and a two-cytokine model are used to differentiate between bacterial and viral LRTIs. Age was included in both models to adjust for this confounding factor.

Performance metric	**Eight-cytokine signature**		**Two-cytokine signature**	
Proteins included	IL-6, G-CSF, IL-15, IL-8, IL-17, TNF-alpha, IL-18, IP-10		IL-6, G-CSF	
	Reference - bacterial	**Reference - viral**	**Reference - bacterial**	**Reference - viral**
Predicted bacterial	10	1	9	1
Predicted viral	5	19	6	19
Accuracy (95% CI)	0.83 (0.66, 0.93)		0.8 (0.63, 0.92)	
Sensitivity (95% CI)	0.67 (0.38, 0.88)		0.6 (0.32, 0.84)	
Specificity (95% CI)	0.95 (0.75, 0.99)		0.95 (0.75, 0.99)	
Positive predictive value (95% CI)	0.91 (0.59, 0.99)		0.9 (0.56, 0.98)	
Negative predictive value (95% CI)	0.80 (0.65, 0.89)		0.76 (0.63, 0.86)	

4.4 Discussion

In this chapter, levels of ten different cytokines were measured in a cohort of children with LRTIs, near time of admission to hospital and again during convalescence. When acute bacterial LRTI results were compared with acute viral LRTI results, the levels of G-CSF, IL-6 and IL-15 were significantly higher in the acute bacterial group.

Two classification models which combined results from the cytokines measured in this study identified bacterial LRTI cases with moderate sensitivity.

4.4.1 Cytokine levels in viral and bacterial lower respiratory tract infections

Mean levels of G-CSF, IL-6 and IL-15 were higher in acute bacterial LRTIs compared with viral LRTIs.

IL-6 is an important cytokine in the response to infection and has been suggested as a biomarker for bacterial infection in critically ill adults. A systematic review suggested that IL-6 had moderate diagnostic value at identifying sepsis in adults.[265] In a case-control study of children with sickle cell disease, IL-6 was suggested as a marker for serious bacterial infections in these patients.[266] However, raised IL-6 levels have also been implicated in viral infections. A study of 50 children with diarrhoeal disease, reported that IL-6 levels were higher in participants with a viral infection compared with a group with bacterial diarrhoea.[267] In a cohort of children with RSV bronchiolitis, IL-6 levels were

elevated in participants with more severe disease.[234] The data in this chapter support IL-6 as a marker for bacterial infection. As bacterial infections are associated with increased severity it may be that the raised IL-6 levels in the previous bronchiolitis study were indicative of bacterial co-infections rather than as an independent marker of severity.

To my knowledge, there is no previous publication that reports higher levels of IL-6 when directly comparing bacterial with viral paediatric LRTIs. Karhu *et al.* report no difference between bacterial and viral groups with severe CAP, however, the sample size in their study was relatively small.[268] The results presented in this chapter suggest that IL-6 could be a useful marker for bacterial infection, in combination with other tests as discussed below. However, it can be difficult to distinguish markers for severity and markers for bacterial infection, as bacterial infections often have more severe disease compared with viral infections.

G-CSF was elevated in the bacterial LRTI group compared with the viral group. This makes sense biologically. Neutrophil levels are known to be higher in bacterial infections. G-CSF is important in the differentiation and development of neutrophils.[254] G-CSF has been suggested as a potential biomarker for bacterial infection.Mouse studies have shown raised levels of G-CSF in *S. aureus*, *K. pneumonia* and *Listeria monocytogenes* infections.[269, 255] A small study measuring G-CSF in adults with different infections reported that G-CSF was increased in acute disease, compared with convalescent samples from the same patients.[270] With neutrophil measurement being easy and widely available, the use of G-CSF clinically would have to show a benefit over routine neutrophil measurement.[271]

The third cytokine elevated in acute bacterial LRTIs compared with the viral LRTIs was IL-15. Out of the three cytokines raised in the bacterial group compared with viral LRTIs, this was the least predictable finding. IL-15 has previously been reported to be a significant cytokine in paediatric RSV bronchiolitis. In a cohort of children with RSV bronchiolitis, Leahy *et al.* reported that serum IL-15 levels highest in participants with severe disease.[252] In our study IL-15 was raised in acute RSV infections compared with convalescent samples. However, there was no difference when RSV was compared with other viral LRTIs. Similarly to the IL-6 findings, our results suggest that it may be difficult to say that IL-15 is an independent factor for disease severity, and undiagnosed bacterial co-infections may be linked to increased IL-15 levels in severe bronchiolitis. IL-15 has been linked with bacterial infections in other studies. IL-15 stimulated NK cells have been reported to be important in defence against *Salmonella* Typhimurium infection in mice.[272] *Salmonella* Typhi is a common pathogen in Nepali children.[34] Two cases in the definite bacterial LRTI group were due to *S.* Typhi, and undiagnosed *S.* Typhi is likely to contribute to a proportion of the undiagnosed bacterial pathogens in the probable bacterial group; this may be a factor as to why IL-15 is raised in this cohort.

IL-15 has been suggested as a biomarker for active tuberculosis, Alzheimer's disease and interstitial lung disease.[273, 274, 275] To my knowledge, IL-15 has not been reported as a possible marker for acute bacterial LRTIs prior to this study.

Combinations of cytokines were included in models to differentiate between bacterial and viral LRTIs. A model including eight cytokines differentiated

between bacterial and viral LRTIs with moderate sensitivity. A two-cytokine model which used IL-6 and G-CSF performed almost as well as the eight-cytokine model. Neither of these cytokine combinations would be useful clinically if used in isolation. Both models wrongly classified a third of bacterial cases in the test dataset. However, these cytokines may be useful if combined with other potential biomarkers. This is discussed further in Chapter 5 when these cytokine results are combined with potential protein and RNA biomarkers.

4.4.2 Cytokines in pathogen groups

Previous studies have reported changes in cytokine levels between different viral infections, or between different bacterial infections. For example, in a cohort of infants with LRTIs, IL-8 levels were higher in the RSV group compared with the influenza group.[237] IL-6, IL-12 and G-CSF were significantly different between mice infected with *S. aureus* and mice infected with *K. pneumonia*.[255] In the results presented in this chapter, when cytokine levels for RSV or influenza were compared with other viral LRTIs there were no significantly different cytokine results. Similarly, when cytokine levels in the pneumococcal LRTI group were compared with levels in other acute bacterial LRTIs there were no significantly different cytokine results. In this cohort study, differences in cytokine levels were not demonstrated between specific pathogen groups. The number of samples in each pathogen group, particularly the RSV group with 49 samples, would suggest that the sample size should not have been an issue in detecting differences between cytokine levels if there were any true differences present. These results suggest that the cytokines included in this panel may not be useful in differentiating

between LRTI aetiology at a pathogen level.

4.4.3 Results in context and future work

The use of a combination of cytokines as a diagnostic test, with or without other proteins, has been proposed for various conditions. Wang *et al.* tested 16 cytokines as biomarkers for active tuberculosis and suggest that a combination of IFN-gamma, IP-10 and C-X-C motif chemokine ligand 9 (CXCL9) gave the best diagnostic accuracy.[276] A study of adults admitted with COVID-19 suggested that a combination of IL-32, IL-6, IFN-gamma, and CRP predicted severity of disease with reasonable accuracy.[126] Looking at multiple cytokines on one platform can aid biomarker discovery by making better use of limited clinical samples; this can especially be an issue in paediatric populations.

Jackson *et al.* reported a six-protein signature for differentiating bacterial and viral paediatric infections. They used two cytokines in this signature, IFN-gamma which was elevated in bacterial infections and IL-18 which was elevated in viral infections.[101] In our cohort, neither of these cytokine results were significantly different when the acute bacterial group was directly compared with the direct viral group. However, IFN-gamma was raised in acute bacterial LRTIs compared with convalescence samples, and no difference was found in IFN-gamma levels when acute viral LRTIs were compared with convalescent samples. Similarly, IL-18 levels were significantly elevated when viral acute and convalescent groups were compared, but not when acute bacterial and convalescent groups were compared. This would support the findings in the Jackson *et al.* paper and suggest that raised IL-18 is potentially more indicative of viral rather than

bacterial LRTIs.

In Chapter 5, I combine these cytokine results with RNA results from Chapter 2 and protein results from Chapter 3 to see if there is potential for including cytokines in a multi-platform signature to identify bacterial LRTIs.

The follow-up fever cohort, discussed in Chapter 6, can be used to validate these cytokine findings in a different cohort of children. The choice of cytokines to include in a future panel will be different, I would like to add new cytokines to a future panel. However, I would like to test the cytokines which were different between bacterial and viral LRTIs again to confirm these findings in another cohort.

4.4.4 Limitations

As discussed in Section 2.4.4, there are limitations in this clinical study which affect the utility of the results. These limitations include the difficulty in confidently assigning LRTIs to a particular group, the relatively low number of definite bacterial cases and issues with bias introduced by age differences between the different classification groups.

Regarding the MSD platform used in this experiment, the laboratory processes are quite time-consuming and a mistake at any stage could make the results difficult to interpret. There are no checks built into the process, so you do not know if there is a problem with the plate until after the results are obtained. An issue with the linker antibody caused IL-8 results to be unusable on two plates. As with any laboratory platform, the quality of the components of the kit is

important in obtaining usable results.

Performing tests across multiple plates means there is a possibility of introducing a batch effect, as occurred in this experiment. This was adjusted for in the analysis, but this emphasises the importance of a good QC system with robust checks so that we can be confident that we are getting reliable, reproducible results.

4.4.5 Conclusions

Measurable increases in different cytokines can be found in acute LRTIs, depending on the aetiology of disease. The cytokines identified as important in bacterial LRTIs could be incorporated into future diagnostic tests. A combination of cytokine results, with other biomarkers, may offer a promising way forward for confident diagnosis in paediatric LRTIs.

Chapter 5

Integrating results from RNA and protein platforms to classify the aetiology of paediatric lower respiratory tract infections

5.1 Introduction

5.1.1 Chapter in context

The results presented in this chapter come from the same cohort study of Nepali children with LRTIs that was also the source of the results for Chapters 2, 3 and 4. This cohort study recruited children admitted to hospital with signs of pneumonia on admission.

In Chapter 1, I discussed the importance of LRTIs globally, and the issues with current diagnostic tests.[20, 17, 60, 61] I highlighted the importance of developing new diagnostics and outlined different platforms used to develop potential new infectious disease diagnostics.[107, 115, 100, 103]

In this chapter, I integrate data from platforms presented in previous chapters in a multi-platform approach to differentiate between bacterial and viral LRTIs. In Chapters 2, 3 and 4, I presented the results from three different platforms, RNA-seq, MS proteomics and an MSD cytokine panel. Samples from children with bacterial and viral LRTIs were compared in each of these three platforms separately and models were developed to differentiate between bacterial and viral LRTIs.

In Chapter 1, I also discussed the potential benefits and challenges of integrating results across different platforms, which are discussed in more detail in Section 5.1.2 below.[131, 133] In this chapter, I integrate gene count data, protein abundance, and cytokine levels. I investigate if a multi-platform approach produces models which can differentiate between bacterial and viral LRTIs with

improved performance, when compared with individual platforms.

In the following sections, I discuss some of the available evidence surrounding multi-platform approaches in diagnostics. I focus on topics which are relevant to the methods used in this chapter, and I discuss the process of integrating data from different -omic platforms, known as multiomics.

5.1.2 Integrating data across different platforms

In this chapter, I integrate RNA data with two different protein platforms. As messenger RNA (mRNA) is translated into proteins, interpreting these platforms together has the potential to provide more useful information than a single platform separately.

If there was a significant overlap in the information obtained by measuring RNA and protein levels there would be little need to measure both. However, studies measuring mRNA and the proteins these mRNA encode suggest that the majority of transcript and protein results do not correlate well with each other. In a study of lung adenocarcinoma samples, only 21.4% of mRNA and protein results had statistically significant correlations.[277]. A mouse study which measured 22,000 transcripts and 5,000 proteins found that the levels of transcripts and proteins had a positive correlation in about half of the comparisons.[278]

In Section 1.8, I outlined studies that have used multiomic approaches to improve our understanding of infectious diseases. For example, the COMBAT consortium performed multiomic analyses across several platforms, including transcriptomics and proteomics, on samples from a large cohort of patients with COVID-19

disease of differing severity. Their results identified markers of disease severity across different platforms, among other insights into the pathophysiology of COVID-19.[133] It can be difficult to interpret the huge volumes of data in these studies, and many different methods have been proposed to assist with data integration and interpretation.

Multiomic data integration

Several different omic integration tools are available, using different approaches. The tools can be grouped by the principle of the methodology used. The categories of methods used include Bayesian, similarity, network, correlation and various other multivariate analyses. Some of the available tools use a combination of these approaches. The most appropriate tools to use depend on the data available and the aim of the project.[279] In this chapter, the mixOmics package was used as this multivariate analysis package provides a set of supervised and unsupervised methods, to allow the integration of multiple datasets with a focus on variable selection.

The main mixOmics methods used in this chapter are PLS-DA and Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO).[280] PLS-DA uses dimension reduction techniques, while also incorporating a known classification, e.g. incorporating bacterial and viral labels. PLS is a multivariate projection-based method which fits a linear regression model. It can be used to explain the relationship between two datasets containing continuous variables. PLS is useful for measuring datasets with many highly correlated variables, as can be the case in large omic datasets. PLS seeks to reduce the dimensionality of

datasets to a set of components.[281] PLS-DA is an adaptation of PLS where instead of two continuous datasets, the inputs are one continuous dataset and a set of outcome variables. The aim is to reduce the dimensionality of the data while maximising the covariance within the defined classification outcome.[282, 283]

DIABLO allows for the integration of multiple datasets while explaining the relationship of each dataset to a categorical outcome variable. DIABLO can be thought of as an extension of PLS-DA. In DIABLO the sum of the covariance between each pair of datasets is maximised. All pairwise covariances are weighted in the design matrix. When integrating the datasets, the focus can be on maximising the integration of the dataset, or on maximising the ability of the resulting model to discriminate between the outcome of interest. When predicting results for new samples in the model, a prediction value comes from each dataset and a majority vote can be used if there is disagreement.[284] These methods have been used in previously published work involving biomarker discovery and differentiation of disease sub-types. DIABLO methods have been used to integrate proteomic and 16S ribosomal ribonucleic acid (rRNA) sequencing data from stool samples to develop a signature for type 1 diabetes.[285] In a study of breast cancer patients, DIABLO was used to integrate gene, micro RNA and protein data from 150 tissue samples and develop signatures to distinguish between breast cancer sub-types.[286]

5.1.3 Hypotheses

Two hypotheses are tested in this chapter:

1. Results from RNA and protein platforms can be integrated to create a model which can differentiate between bacterial and viral LRTIs.

2. When designing models to differentiate between bacterial and viral LRTIs, integrating data from different platforms results in improved model performance.

5.1.4 Aims and objectives

Linked to Hypothesis 1 - Results from RNA and protein platforms can be integrated to create a model which can differentiate between bacterial and viral LRTIs RNA-seq gene counts, MS protein abundance and MSD cytokine level results were integrated and used to build a multi-platform model to differentiate between bacterial and viral LRTIs.

Linked to Hypothesis 2 - When designing models to differentiate between bacterial and viral LRTIs, integrating data from different platforms results in improved model performance This hypothesis was addressed by performing the following steps:

1. Compare the performance of the multi-platform signature created in this chapter with the signatures created using data from the individual platforms using RNA-seq results, Chapter 2, MS proteomic results, Chapter 3, and MSD cytokine panel results, Chapter 4.

2. Examine the features selected in the multi-platform model, and compare them with features from the individual-platform signatures built in previous

chapters.

5.2 Methods

5.2.1 Clinical study

The clinical methods are outlined in detail in Section 2.2.1, and summarised briefly below.

Study design

A prospective cohort study was undertaken at Patan Hospital, Lalitpur, Nepal. In this urban hospital setting, children between 2 months and 14 years of age were eligible for enrolment if they were assigned an admission diagnosis of pneumonia, as decided by the clinical team. Children were enrolled between March 2015 and December 2017. Blood samples for RNA and protein analyses, as well as nasopharyngeal samples for molecular diagnostics, were taken on presentation to hospital. Demographics and medical record data were recorded, including the results of clinical investigations, to assist in correctly classifying the different LRTIs for further analysis. Participants were asked to return six to eight weeks after enrolment to provide a convalescent blood sample when they had recovered.

5.2.2 Laboratory methods

Research samples

At enrolment a nasopharyngeal sample and a blood sample were taken from each participant. A 3 ml blood sample was taken and 1 ml was transferred to an RNA-stabilising tube (Tempus Blood RNA Tube) and 2 ml transferred to a heparinised centrifuge tube. Research samples were transferred to the laboratory within two hours.

Nasopharyngeal samples were sent to Micropathology Ltd., University of Warwick Science Park, UK where nucleic acid was extracted and analysed using the NxTAG™ Respiratory Pathogen Panel, see Section 2.2.2

The processing of the RNA-stabilising tubes is outlined in Section 2.2.2. Processing of plasma samples obtained from the heparinised centrifuge tube is explaine in Section 3.2.2.

The RNA-stabilising tubes underwent RNA-seq at WTCHG, University of Oxford. Plasma samples were sent for MS proteomics at the Discovery Proteomics Facility of the Target Discovery Institute, University of Oxford and MSD cytokine testing at the CCVTM, University of Oxford, as outlined in Sections 3.2.4 and 4.2.4.

5.2.3 Classification of cases

Participants were classified into different LRTI groupings based on the likely cause of their illness. Only acute samples that were classified as bacterial or

viral, as per the semi-supervised approach in Section 2.3.3 were included in the analyses in this chapter.

5.2.4 Sample selection for multiomic analysis

Acute LRTI samples with a classification of bacterial or viral that were tested across all three platforms were eligible for inclusion in the multiomic analyses. Samples with incomplete results for the cytokine panel were excluded so that each of the three datasets contained complete results. There were no missing values in the RNA-seq results and following imputation during the pre-processing in Section 3.2.5, the MS proteomic dataset also contained complete results.

IL-8 was excluded from this multiomic analysis. Due to the issue with the standard curve failing for IL-8 on two of the MSD plates, there were no IL-8 results available for these plates, see Section 4.3.1. Since a full set of cytokine results were required for inclusion in the multiomic analysis, including IL-8 would have meant excluding all samples from the two plates without IL-8 results.

CRP was also removed from the MS proteomics results, as CRP was used as part of the classification system, see Section 2.2.3.

5.2.5 mixOmics methods

Pre-processing and selection of data for mixOmics

mixOmics tools require that all pre-processing steps are carried out prior to using the mixOmics packages. These pre-processing steps were carried out for each of

the platforms during the individual platform analyses in Chapters 2, RNA-seq, Chapter 3, MS proteomics and Chapter 4, MSD cytokine panel.

The mixOmics methods do not allow to adjust for covariates in their model creation and the researchers who created the tools recommend any confounding factors be adjusted for prior to using the mixOmics tools.[280] For this reason, the data from the three platforms was batch-corrected and adjusted for age prior to inclusion in mixOmics analysis. See Section 2.2.4 for batch correction steps undertaken in the RNA-seq results, Section 3.2.5 for batch correction in the MS proteomics dataset and Section 4.2.5 for batch correction in the MSD cytokine data. For mixOmics methods, it is recommended by the researchers who created the tools to limit the number of features selected per platform to less than 10,000. This is to help limit the processing time required for the model-tuning steps.[280] This was not an issue for the MS proteomics dataset, with 309 proteins included post-filtering, or the nine cytokines included, after removal of IL-8 results. However, following filtering there were 22,132 transcripts included for the RNA-seq analysis. In Section 2.3.5, DEA comparing bacterial and viral LRTIs produced a list of 500 significant genes. Rather than performing a more stringent filtering step at the pre-processing stage, I decided to use the 500 genes that had already been shown to be statistically differentially expressed between bacterial and viral LRTIs for the mixOmics analysis.

5.2.6 mixOmics steps in R

The steps followed here are taken from the cited mixOmics methods publications, and the vignettes and online courses available at mixOmics.org.[280, 284]

Preparing the data

Tables containing the gene count results of the 500 genes selected from the RNA-seq experiment, the 309 protein abundance results from the MS proteomics experiment, excluding CRP, and the nine cytokine results, excluding IL-8, were loaded into RStudio.

The rest of the analyses were carried out in R version 4.3.1. The main packages used were the mixOmics, PCAtools and ggplot2. For a detailed list of the steps undertaken in R see Appendix E.

The results in each of the three platforms were scaled and centred to allow for comparisons across platforms. The same scaling method was used as per Section 2.2.6. Results were centred by subtracting the mean for each gene/protein result. Results were then scaled by dividing by the standard deviation of each gene/protein result, producing Z-scores for each result. The 88 included samples were randomly split into test and train datasets using a 70:30 split for model training and performance testing, keeping the ratio of bacterial:viral groups equal within training and test datasets.

Dimension reduction

PCA was carried out for each dataset separately. PC biplots were created to visualise the data before classification was taken into consideration. PC biplots allow a check for outlier samples and to see how well the bacterial and viral groups cluster when the variance is maximised.

PLS-DA biplots were created for individual datasets to visualise the data when

covariance in the defined groups was maximised. These plots showed how well the groups clustered when the outcome variable was incorporated.

N-integration with DIABLO

Integrating two or more datasets where results are available for the same samples is called N-integration in mixOmics. The DIABLO framework was used to integrate the datasets.

First, a design matrix was built using a data-driven approach. PLS was used to calculate pair-wise correlations between the three datasets. The correlation values were used to calculate the weights for the design matrix.

Using the design matrix, a PLS-DA model is fitted to each dataset to assess the global performance of the model and choose the number of components to include. Ten-fold cross-validation was repeated ten times for the maximum components considered, which was five. Error rates were calculated for models including one to five components. The discriminant analysis uses prediction distances to assign values to new variables added to the model. Prediction distances can be calculated in different ways including maximum distance, centroid distance and Mahalanobis distance. The supervised models work by creating a dummy indicator matrix to indicate the class membership of each sample. Maximum distance was used in calculating the model in this chapter. Maximum distance is where each new sample added to the model is assigned to a predicted class based on the dummy variable value that is highest is the outcome class.[283, 282]

Variable selection

Tuning of the model to help decide on the number of variables to include was performed. Models were fitted using ten-fold cross-validation with different numbers of variables considered each time, up to a maximum chosen number of variables for each dataset, twenty variables for the RNA-seq and MS proteomics datasets and all nine cytokines were considered for inclusion. The number of features selected from each platform were included in the final model.

The final DIABLO model was then created, including the weighted design matrix and calculated optimal number of variables. This model was examined and the variables selected were extracted. The correlation between each platform in the final model was calculated. Plots were created to examine the correlation between different model features.

Model performance

Model performance was initially assessed in the training data, the DIABLO model with selected features was assessed using ten-fold cross-validation repeated ten times. The balanced error rate for features across the three platforms was measured. The maximum distance measure for the weighted predicted class was used to classify cases. A confusion matrix was used to calculate the performance statistics including sensitivity, specificity, positive and negative predictive values and AUC-ROC. As we are interested in a diagnostic test that has a high sensitivity for bacterial LRTIs, a partial AUC was calculated when the sensitivity range was 90% to 100%.

Table 5.1: The number of samples in the bacterial and viral groups across the different RNA and protein platforms. Samples with results across all three platforms were included in the multiomic analysis. The number of bacterial and viral cases was defined using the semi-supervised approach in Section 2.3.3. *All IL-8 results were excluded as two plates did not have any IL-8 results. As full results were needed for inclusion, including IL-8 would have meant excluding all results from these two plates. †Proteomic results that were imputed as part of the pre-processing of these data were included.

	Bacterial	**Viral**	**Total**
RNA-sequencing	65	75	140
Mass spectrometry proteomics	55	69	124
Cytokine panel	53	69	122
Cytokine panel with full results*	34	60	94
Samples with complete results across all three platforms†	31	57	88

The performance of the model was then assessed in the test dataset, using the same performance measures as in the training dataset.

5.3 Results

5.3.1 Description of cohort

Table 5.1, shows the number of samples available across each of the three platforms for inclusion in the multiomic analysis. Eighty-eight acute samples with complete results were included for multiomic analysis.

Table 5.2 shows basic demographics and results of routine clinical investigations

at admission. Age, which is lower in the viral group, has been adjusted for in all three of the individual datasets, as this was a known confounding factor. The bacterial group had a longer hospital stay, and higher inflammatory markers on admission compared with the viral group, as expected.

5.3.2 Create training and test datasets

The results were split into training and test datasets, using a 70:30 split. The training dataset was made up of 22 bacterial and 40 viral cases. The test dataset consisted of 17 viral and 9 bacterial cases.

5.3.3 Principal component analysis

Using the training dataset, PCA was carried out and PC biplots were created for each of the three individual datasets.

The PC biplots in Figures 5.1, 5.2 and 5.3 show that there is a varying degree of separation between the groups, across the different platforms, before the group classification is incorporated in the next section.

5.3.4 Partial least squares - discriminant analysis

The PLS-DA approach was used to maximise covariance in the defined classification groups, PLS-DA biplots are presented with component one and component two plotted. The PLS-DA biplots in Figures 5.4 and 5.5 show that there is very little overlap in the bacterial and viral groups in the RNA-seq and MS proteomics

Table 5.2: Baseline data for cases to be included in the multi-platform analysis, divided into bacterial and viral lower respiratory tract infection classifications. Basic demographics and investigations at the time of hospital admission also presented. Samples with results across all three platforms, RNA-seq, MS proteomics and MSD cytokine panel, were included in this analysis. The number of bacterial and viral cases was defined using the semi-supervised approach in Section 2.3.3. CRP, C-reactive protein; WCC, blood white cell count.

	Viral	**Bacterial**	**Total cases**
Number of cases	57	31	88
Age, years (median)	0.8	4.4	1.8
Female (%)	37	45	40
Duration of hospital stay, days (median)	6	8.5	6
CRP, mg/L (median)	7	141	16
WCC x10^9/L (median)	10.6	19.8	12.1
Neutrophils x 10^9/L (median)	5.3	17	6.4

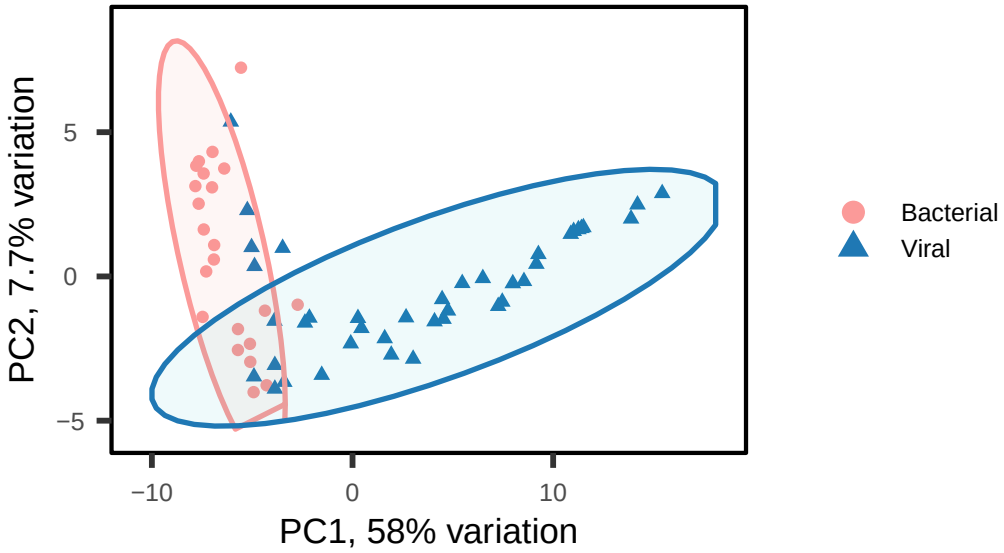


Figure 5.1: Principal component biplot of the RNA-seq training dataset, coloured by bacterial or viral classification, post-normalisation and correcting for the batch effect. Plot only includes samples with results across all three platforms, included for multi-platform analysis. An ellipse is fitted around the groups with assumes a multivariate t-distribution.

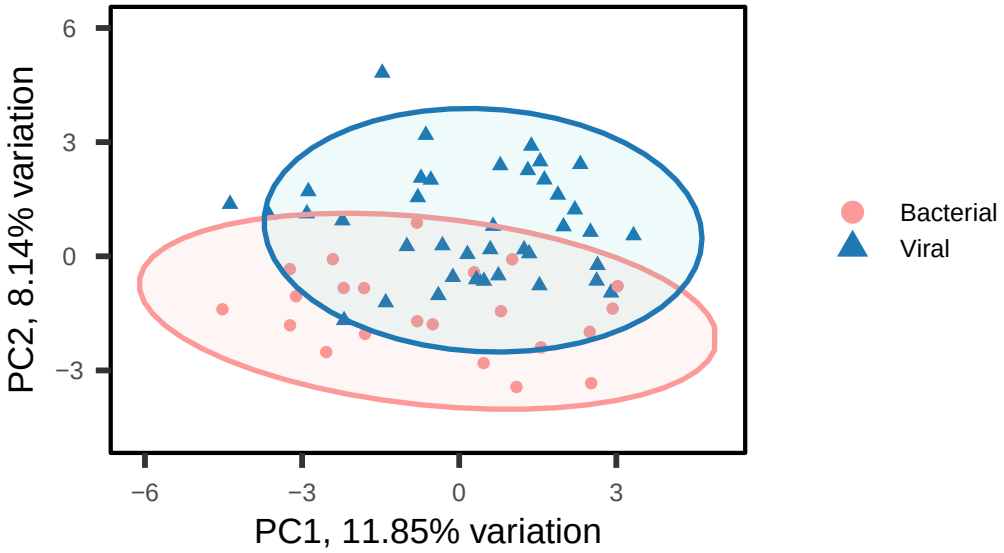


Figure 5.2: Principal component biplot of the proteomic training dataset, coloured by bacterial or viral classification, post-normalisation and correcting for the batch effect. Plot only includes samples with results across all three platforms, included for multi-platform analysis. An ellipse is fitted around the groups with assumes a multivariate t-distribution.

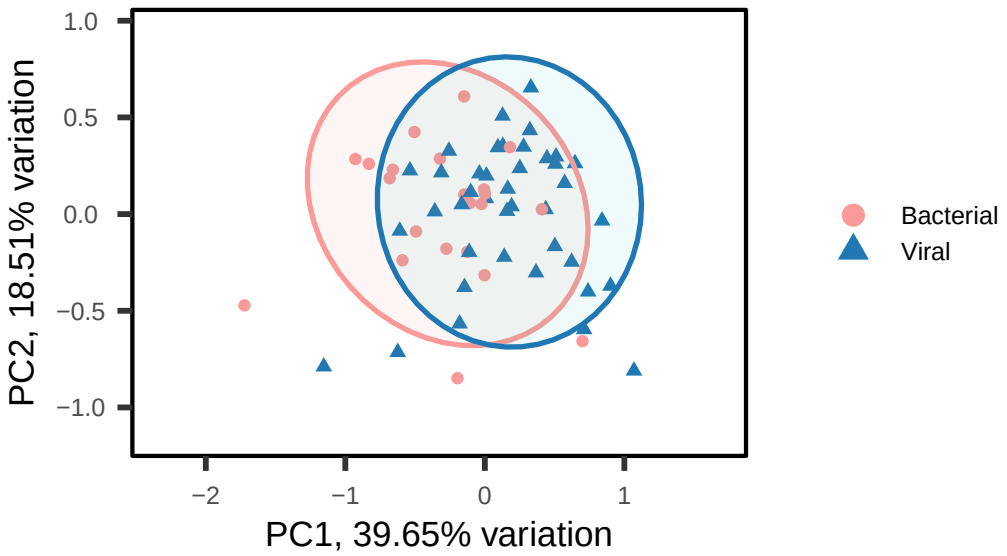


Figure 5.3: Principal component biplot of the cytokine panel training dataset, coloured by bacterial or viral classification, post-normalisation and correcting for the batch effect. Plot only includes samples with results across all three platforms, included for multi-platform analysis. An ellipse is fitted around the groups with assumes a multivariate t-distribution.

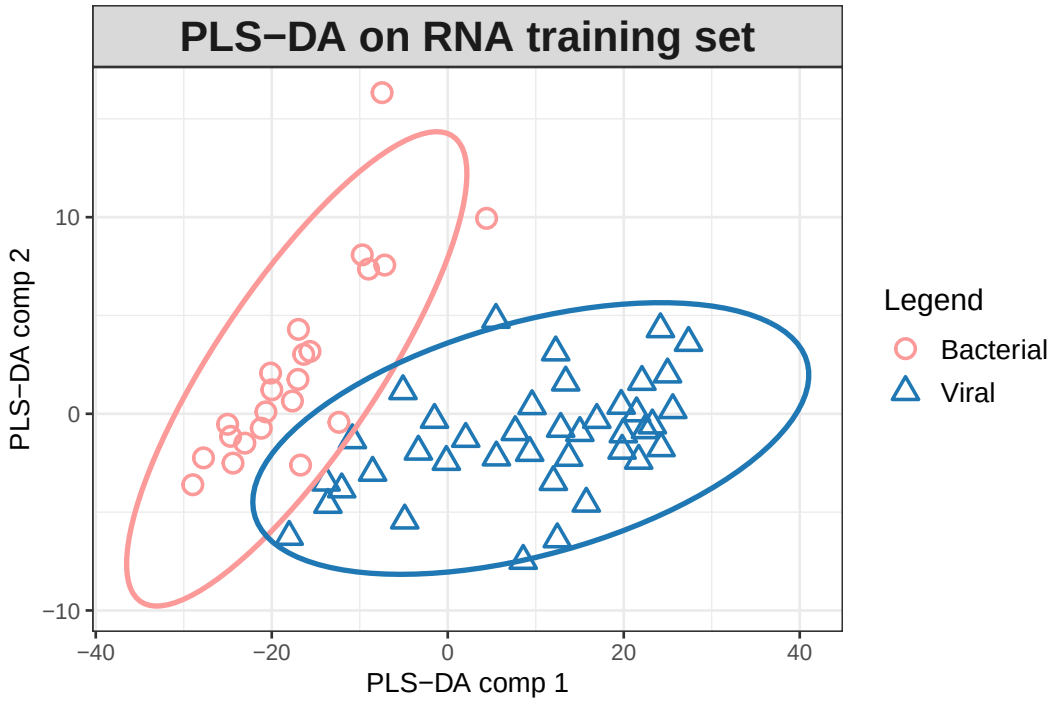


Figure 5.4: Partial least squares – discriminant analysis biplot of the RNA-seq training dataset, coloured by bacterial or viral classification, post-normalisation and correcting for the batch effect. Covariance within classification groups is maximised. Plot only includes samples with results across all three platforms, included for multi-platform analysis.

datasets. The PLS-DA biplot for the cytokine dataset does not show the same separation of the bacterial and viral clusters, see Figure 5.5. For the cytokine dataset, PLS-DA biplots were examined for each of the first five components and none of the component plots resulted in improved clustering, see Appendix E.

5.3.5 DIABLO N-integration

After examining the three datasets individually, the DIABLO framework was used to integrate the datasets.

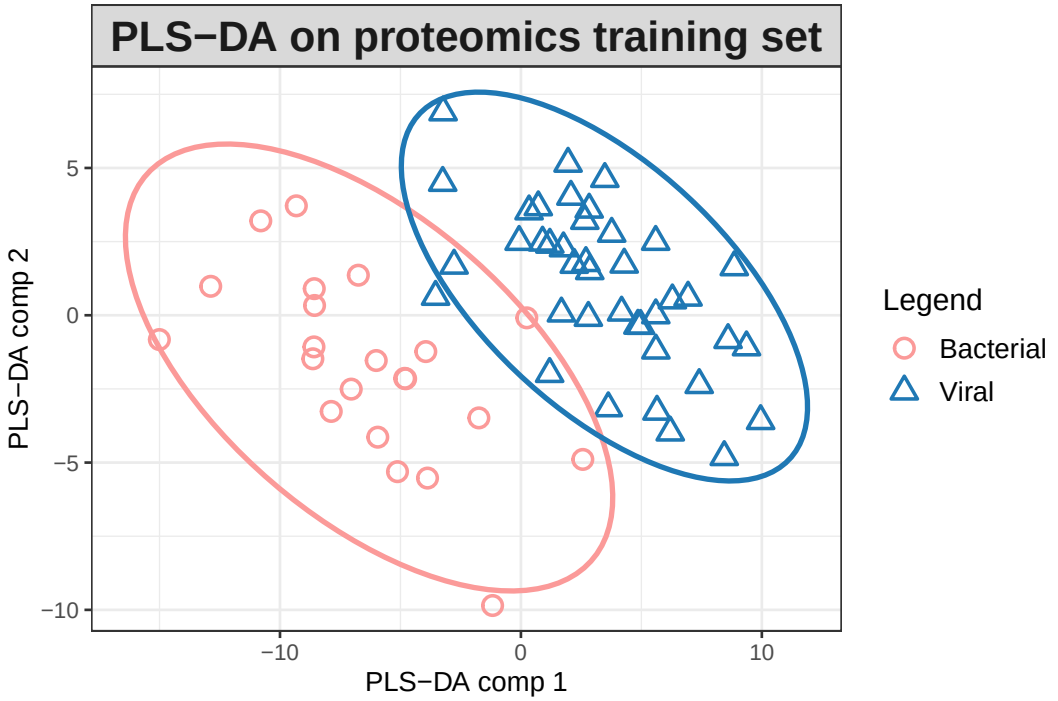


Figure 5.5: Partial least squares – discriminant analysis biplot of the MS proteomics training dataset, coloured by bacterial or viral classification, post-normalisation and correcting for the batch effect. Covariance within classification groups is maximised. Plot only includes samples with results across all three platforms, included for multi-platform analysis.

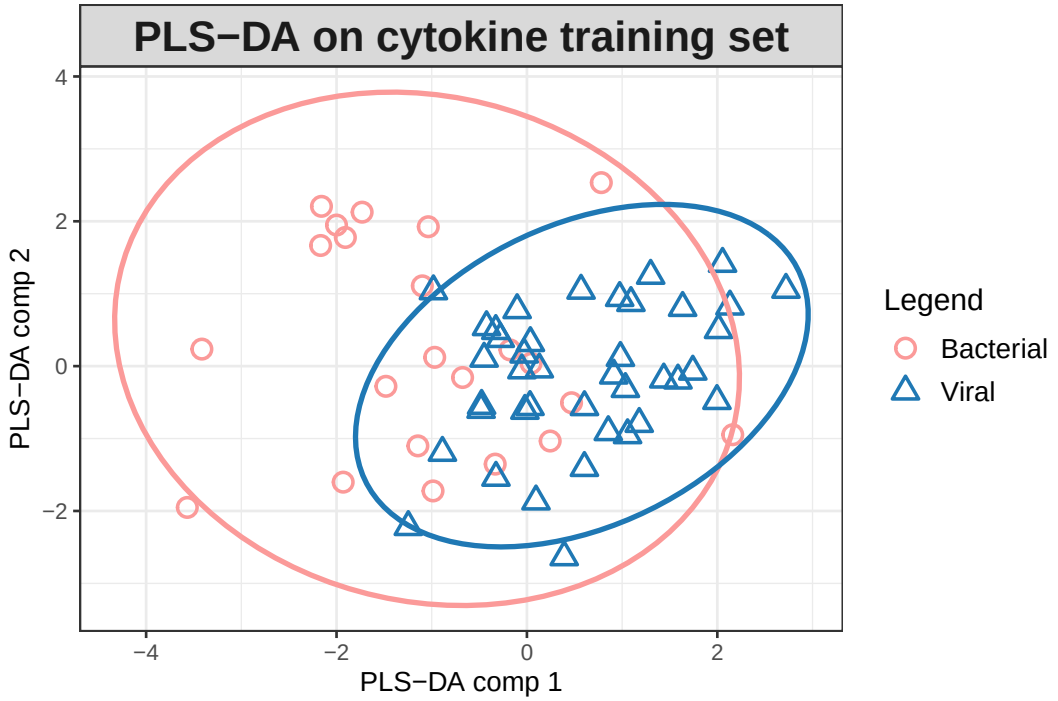


Figure 5.6: Partial least squares – discriminant analysis biplot of the cytokine training dataset, coloured by bacterial or viral classification, post-normalisation and correcting for the batch effect. Covariance within classification groups is maximised. Plot only includes samples with results across all three platforms, included for multi-platform analysis.

The pair-wise correlations between each of the three datasets were used to calculate an appropriate weight for each dataset in the design matrix. The pair-wise correlations were similar between each pair, RNA-seq-proteomics, 0.69, RNA-seq-cytokine, 0.67 and proteomics-cytokine, 0.64. These weights were incorporated into the design matrix.

A PLS-DA model incorporating the weighted design matrix was used to calculate the error rates when each of the first five components were considered. The error rates were similar when including any number of components in the model, see Table 5.3. The error rates suggest that the optimal number of components is one and that maximum distance should be used for the prediction calculation, as the maximum distance results in the lowest error rates for Component 1, when the error rates across the three platforms are combined.

5.3.6 Variable selection

Using the weighted design matrix and the maximum distance prediction measure, PLS-DA models were fitted using different numbers of variables. The optimal model included nine proteins from the MS dataset, five RNA transcripts and one cytokine, see Table 5.4 for features included.

5.3.7 Assessment of the multi-platform model

Correlation was calculated between the three platforms to see how well the DIABLO integration had integrated the different platforms. Figure 5.7 shows the Pearson's correlation values for each pair-wise comparison and PLS-DA biplots

Table 5.3: PLS-DA model is fitted to each of the three datasets to assess the global performance of the model and choose the number of components to include. The error rates, using three different prediction methods, when 1-5 components are shown for each of the three platforms. Maximum distance is the performance measure with the lowest combined error rates for Component 1.

	Maximum distance	**Centroids distance**	**Mahalanobis distance**
RNA-seq			
Component 1	0.130645	0.16129	0.16129
Component 2	0.130645	0.16129	0.166129
Component 3	0.135484	0.159677	0.153226
Component 4	0.146774	0.158065	0.154839
Component 5	0.125807	0.158065	0.133871
MS proteomics			
Component 1	0.145161	0.122581	0.122581
Component 2	0.15	0.127419	0.129032
Component 3	0.135484	0.124194	0.137097
Component 4	0.146774	0.125807	0.140323
Component 5	0.145161	0.129032	0.143548
MSD cytokine panel			
Component 1	0.232258	0.283871	0.283871
Component 2	0.232258	0.28871	0.285484
Component 3	0.233871	0.290323	0.287097
Component 4	0.243548	0.280645	0.290323
Component 5	0.243548	0.290323	0.309677

Table 5.4: List of features included in the PLS-DA multi-platform model and the original platform on which these features were measured. The final columns show if the feature selected was also selected for at least one of the signatures created in the individual platform experiments, either RNA-seq, MS proteomics or MSD cytokine panel. *Official names taken from the National Library of Medicine Gene Database for transcripts and from the UniProt database for proteins.[162, 287]. RNA-seq, ribonucleic acid sequencing; MS, mass spectrometry; MSD, Meso Scale Discovery.

Platform	Feature included in multi-platform model	Official name of feature included*	Feature also included in at least one of the models built in the single platform experiments
RNA-seq	FAM151B	Family with sequence similarity 151 member B	Yes
	ALPL	Alkaline phosphatase, biomineralization associated	**No**
	FSTL4	Follistatin like 4	**No**
	CAPN13	Calpain 13	**No**
	ARL4C	ADP ribosylation factor like GTPase 4C	**No**
MS proteomics	LBP	Lipopolysaccharide-binding protein	Yes
	HAUS3	HAUS augmin-like complex subunit 3	Yes
	SAA1	Serum amyloid A-1 protein	Yes
	SAA2	Serum amyloid A-2 protein	Yes
	A2GL	Leucine-rich alpha-2-glycoprotein	Yes
	IDH3A	Isocitrate dehydrogenase [NAD] subunit alpha, mitochondrial	Yes
	AACT	Alpha-1-antichymotrypsin	Yes
	CLUS	Clusterin	**No**
	GELS	Gelsolin	**No**
MSD cytokine panel	IL-6	Interleukin 6	Yes

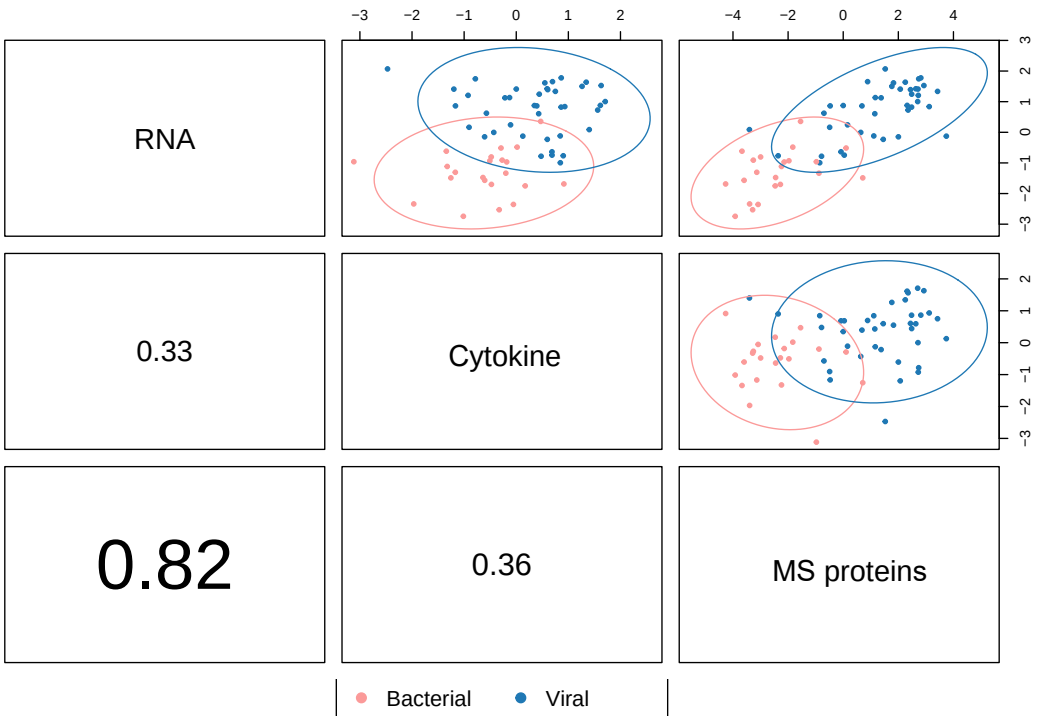


Figure 5.7: Assessment of final model for differentiating bacterial and viral LRTIs using DIABLO methods. The upper-right of the diagram shows PLS-DA biplots comparing the first component of each different platform in a pair-wise fashion. The numbers in the lower left corner indicate the Pearson's correlation coefficient between pairwise correlations of the different platform features used in the model.

showing how well each combination of platforms discriminates between bacterial and viral cases. The correlation between the RNA-seq platform and the MS proteomics platform was better than the correlation of the MSD cytokine panel with other platforms.

The heat map in Figure 5.8 shows the different features included and how the results for each feature differ between bacterial and viral groups. Most of the features in the signature have results which are increased in bacterial LRTIs relative to viral LRTIs. Three features have higher values in the viral

Table 5.5: PLS-DA model error rates for classifying LRTI cases as bacterial or viral, using the features associated with each of the three platforms. Error rates measured using maximum distance performance calculation.

	Bacterial	**Viral**
RNA-seq	0.1	0.19
MS proteomics	0.18	0.07
MSD cytokine panel	0.7	0.2

group, ADP ribosylation factor-like GTPase 4C (ARL4C), clusterin (CLUS) and gelsolin (GELS).

The loadings plot in Figure 5.9 shows the direction of the contribution of each feature. The features are ranked by the absolute value of their coefficients. IL-6 is the only cytokine included in the model. The most important protein is LBP and the most important RNA transcript is alkaline phosphatase, biomineralisation associated (ALPL).

The error rates for the model per platform were calculated, see Table 5.5. The error rates were lowest for classifying bacterial cases using the RNA-seq features, and the error rates were lowest for classifying viral cases using the MS proteomic features.

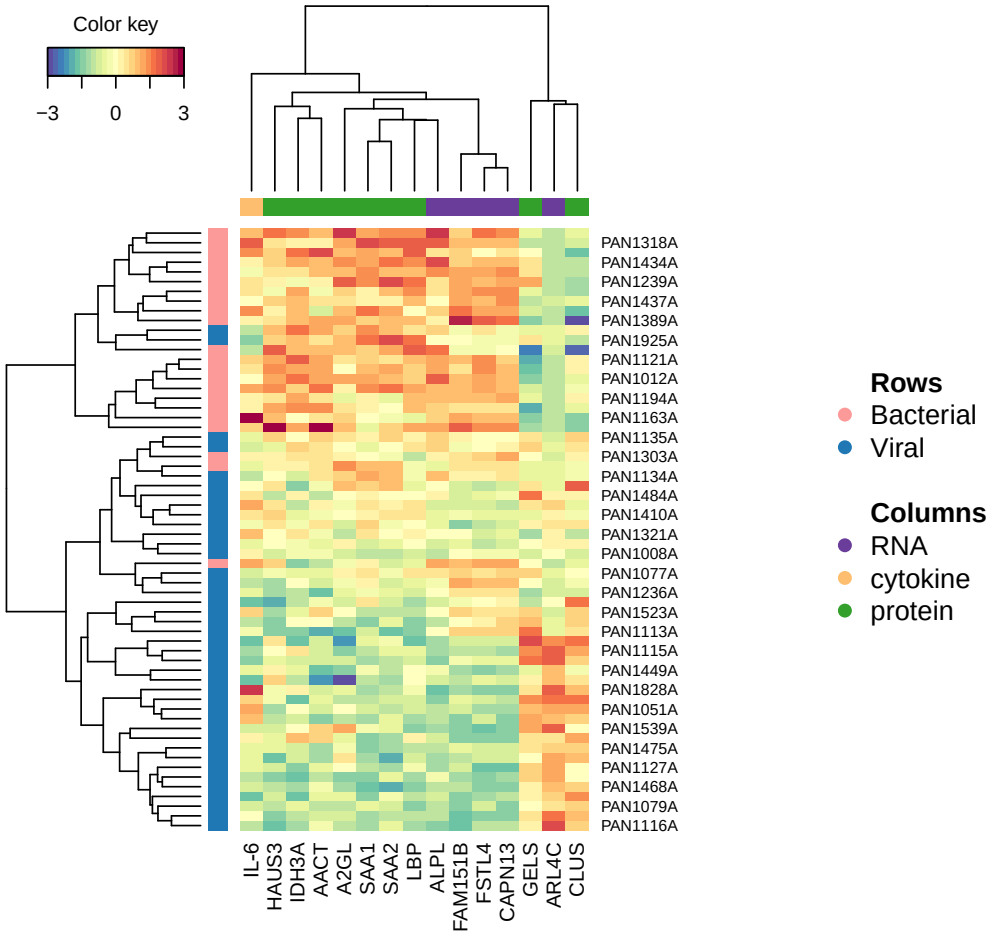


Figure 5.8: Heat map showing the variables included in the DIABLO multi-platform model to differentiate between bacterial and viral LRTIs, grouped by bacterial-viral classification on the y-axis. The columns represent the features included in the model ordered by correlation of results between individual features. The intensity of the colours represent the standard deviation from the mean, limited at +3 and -3 standard deviations.

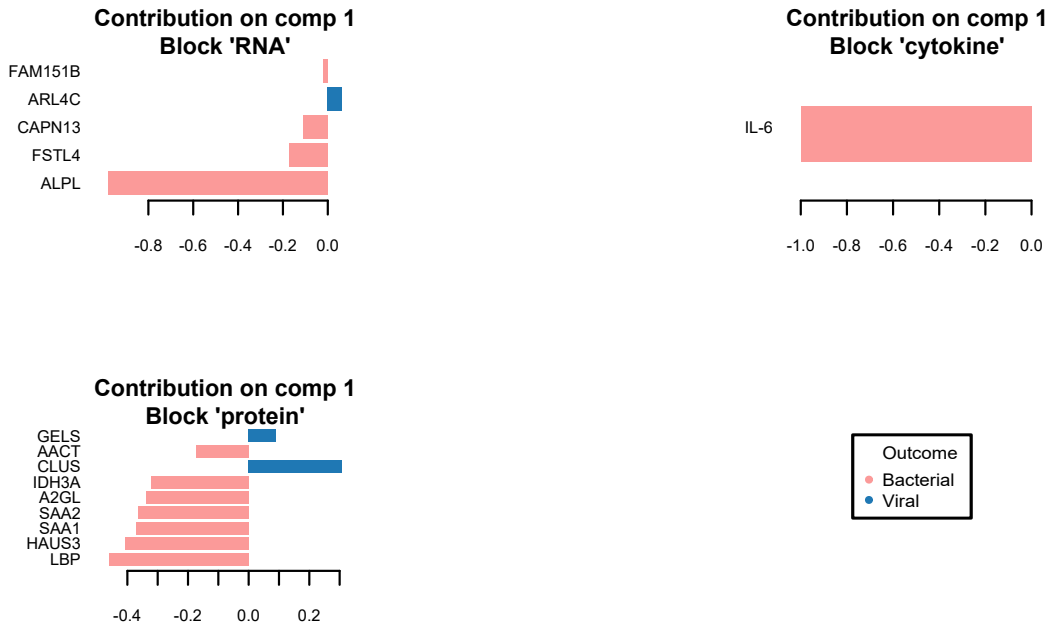


Figure 5.9: Loadings plot showing the most important features in the model differentiating between bacterial and viral LRTIs. The features are ranked in order of importance, according to the absolute value of their coefficients, with the most important values at the bottom. The features are divided by platform. The colours indicate the class in which the median expression value is the highest for each feature.

Table 5.6: The features included in the multi-platform model to differentiate between bacterial and viral LRTIs are shown here, together with the performance statistics in the training and test datasets.

Features included	**Genes**	ALPL, FSTL4, CAPN13, ARL4C, FAM151B			
	Proteins	**LBP**, **HAUS3**, **SAA1**, **SAA2**, **A2GL**, **IDH3A**, CLUS, **AACT**, GELS			
	Cytokine	IL-6			
		Training dataset		**Test dataset**	
		Reference - bacterial	**Reference - viral**	**Reference - bacterial**	**Reference - viral**
Predicted bacterial		20	5	9	1
Predicted viral		2	35	0	16
Accuracy (95% CI)		0.89 (0.78, 0.95)		0.96 (0.80, 0.99)	
Sensitivity		0.91 (0.71, 0.99)		1 (0.66, 1)	
Specificity		0.88 (0.73, 0.96)		0.94 (0.71, 0.99)	
Positive predictive value		0.8 (0.64, 0.90)		0.9 (0.57, 0.98)	
Negative predictive value		0.95 (0.82, 0.99)		1 (0.79, 1)	

Performance statistics

A confusion matrix was created for results when the model was applied to the training dataset. Table 5.6 shows that the sensitivity was 91% with a specificity of 88% in the training data. The model correctly predicted 91% (20/22) of bacterial cases. Figure 5.10 shows that the AUC was 96.2% with an a pAUC of 87.4%.

The model performance was then assessed in the test dataset. Table 5.6 shows that the sensitivity was 100% with a specificity of 94% in the test dataset. There

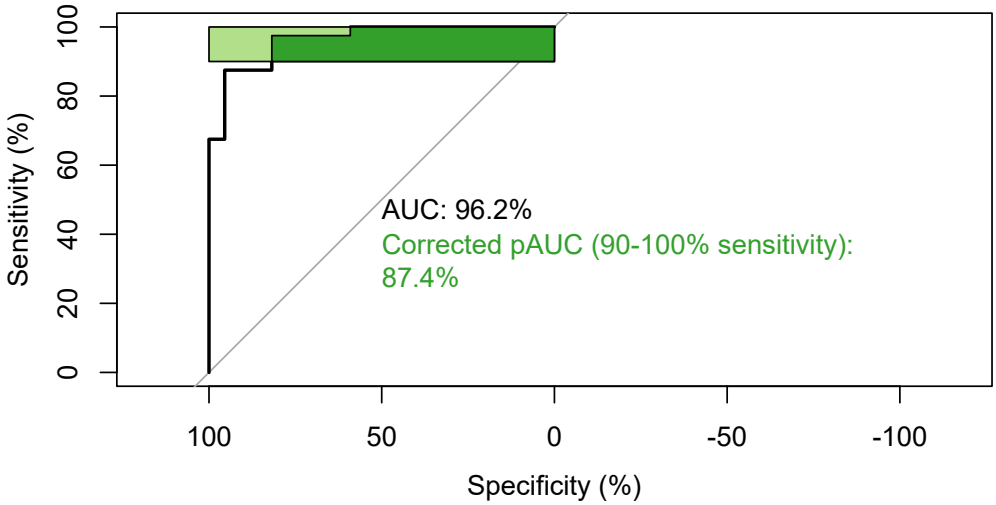


Figure 5.10: In the training dataset, receiver operating characteristic (ROC) curves for the multi-platform model used to differentiate between the acute bacterial and acute viral groups. Features from the RNA-seq, MS proteomics and MSD cytokine panel are included in this model. The area under the curve (AUC) presented for each with a partial AUC at sensitivity range from 90%-100% also shown.

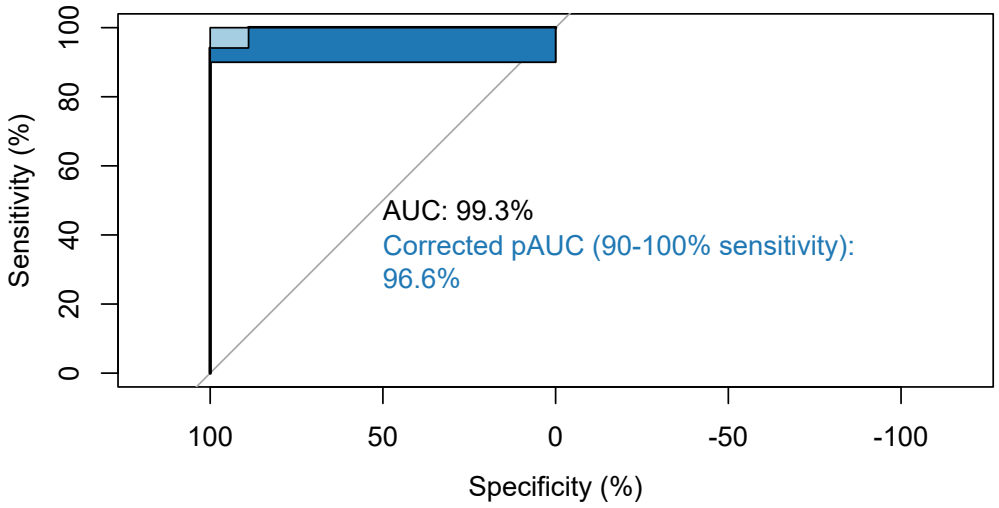


Figure 5.11: In the test dataset, receiver operating characteristic (ROC) curves for the multi-platform model used to differentiate between the acute bacterial and acute viral groups. Features from the RNA-seq, MS proteomics and MSD cytokine panel are included in this model. The area under the curve (AUC) presented for each with a partial AUC at sensitivity range from 90%-100% also shown.

were nine bacterial LRTI cases in the test dataset, the model correctly predicted 100% of these cases. Figure 5.11 shows that the AUC was 99.3% with an a pAUC of 96.6%.

5.4 Discussion

In this chapter, I have integrated data across three different platforms to create a multi-platform model which can differentiate between paediatric bacterial and viral LRTIs with greater accuracy than models created using the individual platforms, as outlined in the next section.

Multi-platform model

A similar multi-platform approach to this chapter has been reported in other studies aiming to improve the accuracy of classification of disease. Byeon *et al.* used a multiomic model incorporating proteins, lipids and metabolites to identify a set of 102 features to predict severity outcomes in COVID-19 with greater accuracy than other tests in routine use.[288] The authors used a machine learning approach as integrating several datasets is highly complex. Going from a complex model incorporating features from different platforms to a more simple test which can be used in routine practice is a challenge.

Outside of infectious disease, Martin-Hernandez *et al.* also used a machine learning approach in identifying a multiomic signature with a high prognostic value for adrenocortical carcinoma. The authors identified a nine-feature signature incorporating genes and microRNA which successfully predicted cancer patients at risk of poor outcomes. In high-resource settings, oncology might be an area where these resource-intense multiomic signatures are used more often. Due to the cost involved in treating long-term cancer patients, it may be easier to justify incorporating a resource-intense multi-platform test.

Features in multi-platform model

The multi-platform model incorporated features from three platforms, five transcripts from the RNA-seq experiment in Chapter 2, nine proteins from the MS proteomics experiment in Chapter 3 and one cytokine from the MSD cytokine panel in Chapter 4.

Of the RNA transcripts included in the model, only family with sequence similarity 151 member B (FAM151B) was selected for one of the models created to differentiate between bacterial and viral LRTIs, using the RNA-seq results in Chapter 2. There is little published literature related to FAM151B, with only three studies found when performing a MEDLINE search. FAM151B has been associated with retinal degeneration, methotrexate resistance and obesity and not linked to any infectious disease processes. Whether or not FAM151B is a useful biomarker for bacterial infections, will need to be confirmed by future studies.

Of the other transcripts included, I have highlighted two genes of interest. First, ARL4C encodes for a guanosine triphosphate (GTP)-binding protein which is part of the RNA synthesis process. ARL4C has a role in cancer invasion and tumour suppression.[289, 290] ARL4C has not been widely associated with infection, but Lin *et al.* included ARL4C in a four-protein signature they discovered for sepsis. ARL4C expression was found to be decreased, using single cell RNA-seq, in the sepsis patients. However, the numbers in this study were low, and the signature was found using a group of just five adults with Gram-negative sepsis. The authors did validate their findings using a knock-out mouse model.[291] The findings reported in this chapter are consistent with the Lin *et al.* study. In our LRTI cohort, ARL4C expression was relatively decreased in the bacterial group compared with the viral LRTI group. These findings point to ARL4C as a possibly useful biomarker, and it's role in bacterial infections should be examined in future work.

The second gene of interest, ALPL, encodes a member of the alkaline phosphatase

family of proteins that are not associated with a particular tissue. Tissue-non-specific alkaline phosphatases have a recognised role in bone mineralisation and have not broadly been linked to infection. However, one mouse study did suggest human tissue non-specific alkaline phosphatase as a treatment for sepsis. The mice that were given tissue non-specific alkaline phosphatase had improved survival when exposed to a bacterial antigen associated with Gram-negative bacterial, lipopolysaccharide.[292] Relative expression of ALPL was increased in the bacterial LRTI group in our study. ALPL may be part of the immune response to bacteria. One of the proteins included in the model, LBP is also important in the immune response to lipopolysaccharide. From the data in this chapter, we cannot be confident that ALPL is associated with Gram-negative bacterial infections, but the expression pattern of LBP and ALPL were similar across our cohort of children with LRTIs. This may suggest that a link between ALPL and Gram-negative infections is worth exploring. A gene which encodes for a protein with a similar function to ALPL, ACP3, was included in each of the four gene signatures evaluated in Chapter 2, and suggests that phosphatase proteins may have a role in the response to bacterial LRTIs in this cohort.

To my knowledge, the other highlighted transcripts have not been suggested as biomarkers for infectious diseases.

Of the nine proteins included in the multi-platform model, seven of these were included in the 12 proteins that were differentially abundant between bacterial and viral LRTIs in the MS proteomics results, Chapter 3. These proteins have been discussed in Section 3.4, and include acute phase reactants, SAA1 and SAA2, and proteins previously associated with bacterial infection, LBP, and

neutrophil function, AACT. Of the other two proteins, not previously discussed, clusterin and gelsolin were both relatively more abundant in the viral group.

Clusterin helps to remove misfolded extra-cellular proteins, and has been associated with regulating the inflammatory response in sepsis. In a cohort of patients with sepsis, clusterin was reported to bind to histone and reduce inflammation. Lower plasma levels of clusterin were observed in patients with sepsis.[293] Clusterin has been associated with down-regulation of NF-kappaB in a study looking at human neuroblastoma cell lines.[294] We have discussed the role of NF-kappaB in Section 2.4, as regulation of NF-kappaB was one of the highest enriched pathways in GSEA when pneumococcal LRTIs were compared with other bacterial LRTIs. In this chapter, clusterin was included in the multi-platform bacterial-viral model because it was relatively more abundant in viral LRTIs. Taken together, the findings presented in this chapter and the previously published data suggest that lower levels of clusterin are associated with bacterial infections.

Gelsolin is a calcium-activated, multi-functional protein associated with remodelling of cytoskeleton structure. Gelsolin interferes with actin function and is found in the cytoplasm of most cell types, and in the extracellular space.[295] Gelsolin has been suggested as a biomarker for sepsis. In a study of 91 adults admitted to the intensive care unit (ICU), gelsolin was lower in severe sepsis patients, compared with other ICU patients without sepsis.[296] However, gelsolin may be a more non-specific marker for inflammation, with another study of hospitalised adults reporting gelsolin to be decreased in myocardial infarction and acute liver injury, as well as sepsis patients.[297] In our cohort of children with

LRTIs, gelsolin was decreased in the majority of children with bacterial LRTIs relative to levels in viral infections. Gelsolin is a protein worth further study to see if these results can be replicated in other studies of paediatric infection.

Only one cytokine from the MSD panel was included in the final model, IL-6. IL-6 was also found to be significantly higher when bacterial and viral LRTIs were compared in the MSD experiment in Chapter 4. In Section 4.3.7, IL-6 was selected for inclusion in a three-cytokine signature that had moderate sensitivity for detecting bacterial LRTIs. IL-6 has been suggested as a useful biomarker for bacterial disease in several studies, as discussed in Section 4.4.[233, 265, 266]

The multi-platform model incorporating features from the three platforms performed better than the signatures from the individual platforms at classifying bacterial LRTIs. Clinically, a concern with a new diagnostic test of this kind would be missing bacterial infections. In the results presented so far in this thesis, the best performing single platform signatures in the test datasets, if measured in terms of missed bacterial infections, were the ten-gene and three-gene RNA-seq signatures, Section 2.3.5; the RNA-seq signatures both would have misclassified 11.1% (2/18) of the bacterial cases. In the test dataset in the current chapter, the multi-platform signature correctly identified 100% of bacterial LRTIs (9/9). It would be good to be sceptical of this result as the number of cases in the bacterial group was smaller than the other comparisons. However, this is a promising initial finding. The pAUC at the sensitivity range of 90%-100% was also higher in the multi-platform signature than any other signature in this thesis, at 96.6%.

Utility of multi-platform signatures

Integrating data across platforms has the potential to improve accuracy if assigning disease classification, as shown in this chapter. However, if the aim is to produce a new diagnostic test that can identify bacterial infections in children, it may be difficult to develop a test that incorporates genes and proteins which have been measured on different platforms. To my knowledge, there is no published data incorporating RNA transcripts and proteins in one model to differentiate between bacterial and viral infections.

However, maybe the aim does not have to be to create one single-platform new diagnostic test. Clinicians do not use one test in isolation to make a diagnosis, but rely on all the available data in making a clinical decision. In many settings, diagnostic decisions are made based on integrating host protein and pathogen RNA results. If the multi-platform signature reported in this chapter, or other multi-platform models, were to be validated across different settings, it should be possible to incorporate assays for a small number of proteins, and a small number of RNA transcripts in clinical laboratories. Asare-Werehene *et al.* suggested this kind of approach when classifying severity of COVID-19 cases. They used a protein assay to measure gelsolin and a separate cytokine panel; they combined these results to show that this multi-platform approach could predict outcome in COVID-19 cases with greater accuracy when compared with CRP and ferritin.[298]

Relevant to the results presented in this chapter, Jackson *et al.* took a multi-platform approach to develop a signature to differentiated between bacterial and viral infections in children. They used different proteomic platforms and

different immunoassay platforms to test smaller numbers of proteins to build a six-protein signature. This signature included six-proteins which could ultimately be measured on a Luminex immunoassay, this shows that even when starting with a multi-platform approach, a signature could be developed which can be measured using a single platform.[101]

In the result presented in this chapter, CRP was excluded when building the model, as CRP was used as part of the classification system. Incorporating CRP, or other biomarkers which are already known to be useful in classifying bacterial infections may increase the performance of this model further.

5.4.1 Limitations

As discussed in Section 2.4.4, there are limitations in this clinical study which affect the utility of the results including potential bias introduced by age differences between the different classification groups. The issue with age was adjusted for statistically across all of the three platforms before the data were used to build the model.

The number of participant results included in this chapter was lower than in other chapters, as only samples with a full set of results were included. The small number of bacterial cases in the test dataset, in particular, mean that this signature should be tested on a larger population to increase confidence in the results.

Across each of the platforms a small number of bacterial cases were consistently misclassified; it is difficult to determine whether this is due to the new models

misclassifying these cases, or whether these cases were actually misclassified using gold standard diagnostics and the semi-supervised approach used in Section 2.2.5.

5.4.2 Future work

The bacterial-viral multi-platform signature presented in this chapter was identified using one cohort of Nepali children. For this signature to be useful outside of this Nepali setting it would need to be tested in different populations.

The follow-up cohort study being undertaken at the same site, and described in Chapter 6, will allow for RNA-seq and protein measurements to be undertaken in a new cohort of Nepali children with various infections, including LRTIs. This would allow the signature to be validated in a different cohort at the same Nepali setting. If the protein and RNA results from this follow-up study supported the use of this signature, the signature would then need to be validated in a non-Nepali context, to assess the global utility of the signature as a diagnostic test.

5.4.3 Conclusions

The multi-platform signature identified in this chapter can classify LRTI cases with increased accuracy when compared with signatures built using individual platforms. It would be technically more challenging to incorporate features from different platforms into a diagnostic test, but the results presented in this chapter suggest that measuring a small number of genes and proteins could be used to create a clinically useful diagnostic test.

Chapter 6

Description of paediatric infections at Patan Hospital, Nepal, using routine diagnostics and additional molecular testing

6.1 Introduction

6.1.1 Relationship to other thesis sections

In this chapter, I report the initial results from a prospective cohort study enrolling children with signs of infection, which I refer to as the fever cohort. These children were enrolled at the same study site, Patan Hospital, as the LRTI cohort study described in Chapter 2. Cases were classified based on routine diagnostics, and then re-classified following the addition of results from molecular diagnostic panels.

This fever cohort includes children with respiratory and non-respiratory infections, expanding the breadth of diseases examined in the preceding LRTI cohort.

The results from the children with LRTIs in this fever cohort can be used in future work to validate the RNA, Chapter 2, and protein results, Chapters 3 and 4, from the LRTI cohort.

In Chapter 1, I discussed the causes of infection in Nepali children. The causes of paediatric infections have not been well described in many Nepali settings. In a study of paediatric invasive bacterial disease at the same Nepali hospital as these cohort studies, Carter *et al.* reported that *S.* Typhi and *S. pneumoniae* accounted for 44% of the pathogens cultured from blood, pleural fluid of CSF. The bacterial disease data from that study were collected between 2005 and 2013, just prior to the introduction of PCV10 in Nepal.[56] An updated report on the causes of paediatric invasive bacterial disease, post-PCV10 introduction, would provide useful information to Nepali policy makers.

Also in Chapter 1, I discussed routine culture-based diagnostic tests in paediatric infections, and the potential utility of antibody and antigen testing. Diagnosis of infections in Patan Hospital includes PCR testing for SARS-CoV-2 only, no other molecular testing is available for other infectious disease pathogens.

The utility of molecular testing is a main aim of this chapter, and is discussed in Section 6.1.3 below.

6.1.2 Consortium and overall project

The cohort study described in this chapter was undertaken as part of a large multi-centre study. The name of the consortium is Diagnosis and Management of Febrile Illness using RNA Personalised Molecular Signature Diagnosis (DIAMONDS). The consortium aims to discover and validate specific RNA signatures for different infectious and inflammatory diseases. These signatures will then be used to develop novel diagnostic tools to identify infectious and inflammatory conditions, using RNA platforms.

At the time of writing, there were 29 partner institutions across 13 countries involved in the consortium. Most of these sites are in Europe with three non-European sites, in Taiwan, The Gambia and Nepal. The non-European sites are important so that new diagnostic tools developed during the study can include pathogens relevant to non-European sites, and to show that any new diagnostic tools developed can also work in non-European settings.

6.1.3 Molecular testing

Molecular testing is used in many high-resource settings to augment traditional culture-based diagnostics. Molecular tests, in general, have the advantage of providing faster results, and having increased sensitivity compared with cultures.[70] In the wake of the COVID-19 pandemic, more settings have the facilities to perform molecular testing, including Patan Hospital. In the following sections, I highlight useful molecular targets in this context.

As discussed in Section 1.6, the detection of certain viruses in the upper respiratory tract has been associated with LRTIs. In large case-control trials of paediatric LRTIs, RSV, influenza, parainfluenza viruses and human metapneumovirus have higher detection rates in LRTI cases compared with healthy controls.[17, 54]

Detection of nucleic acid for a pathogen in blood can provide the diagnosis in infection cases but the results need to be interpreted in the clinical context.

Certain pathogens are more important in South Asian contexts, such as Nepal, than in other geographical areas. As outlined in Section 1.6.1, molecular testing for enteric fever-related *Salmonella*, *Rickettsia* species and *Leptospira* has the potential to augment diagnosis but these PCR tests have variable sensitivity.[91, 98, 99] These tests are not routinely used in many Nepali healthcare settings.

Molecular panels

Four molecular panels were evaluated in this cohort study, and the full list of molecular targets is discussed in Section 6.2.5.

Molecular testing can be useful in diagnosing important bacterial infections. Shah *et al.* reported results from molecular testing in a large cohort of European children admitted to hospital with febrile illnesses. I have focused on the Shah *et al.* results as this was a large multi-site cohort study across nine European sites which used the same molecular targets as two of the molecular panels used in the fever cohort presented in this chapter. Figure 6.1 shows the molecular targets tested. Using a similar classification to that used in this chapter, Section 6.2.3, the authors compared molecular results between bacterial and viral cases across several bacterial targets, *S. pneumoniae*, *N. meningitidis*, Group A *Streptococcus* and *E. coli* had significantly higher detection rates in the bacterial, compared with the viral, group. In the following sections, I discuss other evidence around the use of these four molecular targets in diagnosing bacterial infections. The levels of detection for the other bacterial targets in the study were not significantly different between the bacterial and viral groups.[76]

S. pneumoniae is one of the most common causes of pneumonia, and was the most commonly detected bacterial pathogen at Patan Hospital, prior to the introduction of PCV10.[56] A molecular test to diagnose *S. pneumoniae* infections would be useful in identifying acute infections and monitoring epidemiology of disease. Several promising gene targets have been reported, including the autolysin gene *lytA* which was highly specific for *S. pneumoniae* infection.[299] Rello *et al.* reported the detection of *lytA* gene targets in 62% of adult patients with confirmed or suspected pneumococcal pneumonia.[300] In a study of Nigerian children with febrile illness and healthy controls, 1,038 dried-blood spots were tested for *S. pneumoniae* using *lytA* molecular target. Nine out of 15 culture-

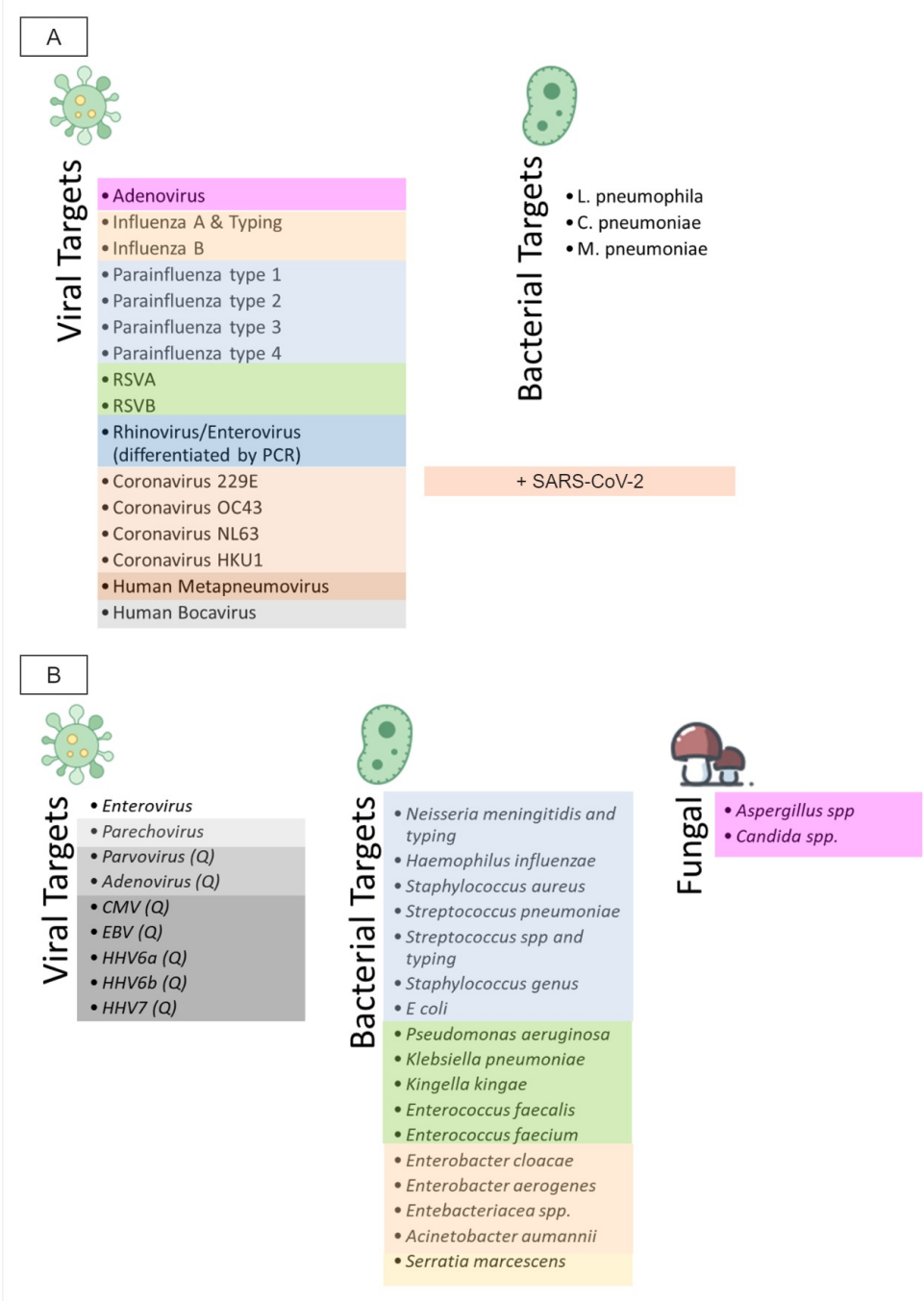


Figure 6.1: Targets tested on molecular panels at the laboratory of Micropathology Ltd. Panel A shows the targets included in the NxTag Respiratory Panel. Panel B shows the targets for the in-house blood molecular panel. RSV, respiratory syncytial virus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; CMV, cytomegalovirus; EBV, Epstein–Barr virus; HHV, human herpesvirus. Figure used with permission from Micropathology Ltd.

positive participants were also positive for *S. pneumoniae* on PCR, giving a sensitivity of 60% with a specificity of 99.7%. Only one healthy control sample was positive for *S. pneumoniae* on PCR (1.2%).[301]

N. meningitidis is an important cause of meningitis, and if symptoms of meningitis are present, CSF is the preferred sample type to diagnose meningitis. However, blood PCR may add value in cases where a CSF sample is not available or CSF does not provide an answer.[302, 303] In a study of adult patients with invasive meningococcal disease, blood PCR was 3.5 times more sensitive than blood culture in diagnosing disease.[304]

Group A *streptococci* are commonly associated with pharyngitis in children, however invasive group A streptococcal disease can have very poor outcomes if not recognised early.[305] Throat samples are commonly tested for group A *streptococci* using PCR, but when looking for invasive disease, detection of group A *streptococci* in blood is a potentially useful test. A case-control study of blood PCR from cord blood samples reported an odds ratio of 4.5 (95% CI 1.6-13) for serious bacterial infection in infants that were PCR positive for group A *streptococci*.[306]

The site of *E. coli* infections is commonly the urinary tract, but *E. coli* infections can be associated with sepsis or meningitis, particularly in young children.[307] Limited data were found on the correlation between *E. coli* PCR blood results and blood culture results. In a study comparing PCR positive blood samples with blood culture-confirmed *E. coli* infections, Lucignano *et al.* report an 83% correlation between blood culture and PCR results (five out of six *E. coli* infections were positive on blood culture and PCR).[308] In a study of adult

sepsis patients, 18 cases were blood culture-positive for *E. coli*, with 11 of these also positive on PCR. Taking blood culture as the gold standard, sensitivity of PCR for *E. coli* was 61% with 97.8% specificity.[309] In a case-control study of infants with serious bacterial infections, a cord blood sample positive for *E. coli/shigella* species was associated with early onset sepsis with an odds ratio of 2.6 (95% CI 1.6-4.4).[306]

E. coli PCR might be useful in settings where antibiotics have already been given. In a study where rats were intravenously challenged with *E. coli* PCR was positive in 62% of subjects 3.5 hours after challenge, compared with 54% positivity in blood cultures. When antibiotics were given, the positivity rate for PCR was 50% compared with 0% positivity in blood cultures.[310]

Of the blood viruses tested in the molecular panel in the Shah *et al.* paper, enterovirus was the only virus which had a significantly higher odds of detection in the viral group, compared with the bacterial group.[76] Enterovirus PCR testing has shown to be sensitive test when compared with cell culture in a study of different clinical samples, the majority of which were CSF samples, 288/322. Blood samples were not included in this study.[311] In a case-control study from Obiero *et al.*, cord blood samples were tested for a variety of molecular targets in neonates who had signs of neonatal sepsis in the first 48 hours of life, and compared with samples from healthy controls. Enterovirus was found to have a cumulative odds ratio of 9.1 (95% CI 2.3-37) for detection in the cases.[306] A review of enterovirus PCR diagnostics from Harvala *et al.* recommended that blood enterovirus PCR be tested in blood, in addition to CSF samples in patients with suspected enterovirus and neurological symptoms. However, this review

highlighted that evidence for enterovirus PCR use clinically was more robust for PCR in CSF samples.[312]

While there is evidence supporting the use of these four bacterial molecular targets in diagnosis, Shah *et al.* also compared the odds of detecting bacterial molecular targets in cases and controls; in this comparison no bacterial targets were statistically different between cases and controls.[76] The authors suggest that the clinical significance of a positive blood PCR result can be unclear. The clinical context of any diagnostic test needs to be considered; the clinical context will be reviewed in the cases re-classified based on molecular testing in this chapter.

Impact of molecular testing on clinical practice

Identifying a pathogen in a molecular test is more important if this identification can potentially alter the clinical management. Identification of viruses can be important in identifying outbreaks and for infection control measures. Increased identification of bacterial pathogens should allow physicians to focus limited healthcare resources on these potentially severe infections.[309, 74]

6.1.4 Diagnostics in dengue fever

Dengue fever diagnostics are discussed here as there was a large outbreak of dengue virus infections during enrolment to this cohort study.

Routine testing for dengue fever at Patan Hospital involves NS1 antigen, IgM and IgG dengue antibody testing. NS1 testing provides a quick answer early in

the dengue fever infection but questions have been raised about the specificity of NS1 testing. Iqbal *et al.* reported that 28.5% of NS1 positive results were false positives when taking PCR as the gold standard.[313] Amand *et al.* also reported higher NS1 positivity (80.9%) rates in dengue-suspected cases compared with PCR (68.1%). This could mean that the NS1 positive, PCR negative cases are false positives, or NS1 is a more sensitive test for dengue. Timing of the sample could be important with NS1 positivity found 1 to 12 days after symptom onset, compared with PCR positive samples from day 1 to day 8 after symptom onset.[314]

Dengue IgM antibodies can be detected from day 3 of illness and are present in over 90% of infected individuals in the second week of illness.[1, 44, 315] Dengue IgG can be detected at low levels at the end of the first week of illness. A very high IgG level, or paired testing showing a rise in IgG during acute infection, are accepted as confirmation of a recent dengue infection.[44, 315]

Molecular testing is considered the most sensitive and specific test available for dengue fever.[1] Molecular testing for dengue is not part of routine diagnostics at Patan Hospital but dengue PCR was part of the additional molecular screen described in this chapter.

Dengue serotype testing is useful in determining the serotypes circulating in a population. Changes in serotype distribution have the potential to lead to increased incidence of severe dengue cases. In Nepal, all four serotypes have been identified, with serotype 1 generally predominating during outbreaks.[51, 52]

6.1.5 Identification of children at risk of severe infections

Children with certain co-morbidities, such as congenital cardiac or chronic respiratory conditions, are known to be at higher risk of severe infections.[316] In children without known co-morbidities, it can be difficult to predict which children, at presentation to hospital, will go on to develop severe disease in different infectious presentations. Identifying cases of severe disease early improves outcomes.[317, 318] Correctly classifying severe cases is important in biomarker discovery studies.

It was not possible to make an accurate assessment of disease severity in the LRTI cohort described in Chapter 2, due to the clinical data collected. An assessment of severity in LRTIs enrolled to the fever study in this chapter will be made but this is outside of the scope of this thesis. See Appendix F, for a description of the LRTI severity measures that will be used in future analyses.

As several dengue cases were enrolled to this fever cohort, an assessment of dengue fever severity is included to describe the characteristics of the dengue group.

Severe dengue disease

After diagnosing a patient with dengue fever, it is important to identify which patients are at risk of severe disease. Patients with dengue fever are risk stratified on presentation to Patan Hospital, see Figure 6.2.

The 2009 WHO Dengue Case Classification system is used to stratify patients. Dengue cases without warning signs can be discharged from the Emergency

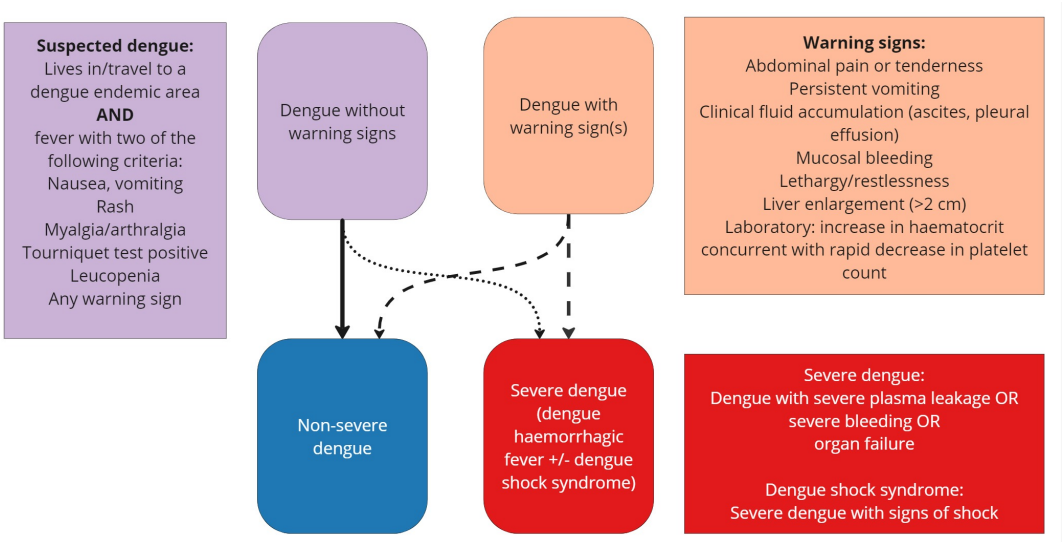


Figure 6.2: Classification and risk stratification system used for suspected dengue cases on presentation to Patan Hospital, Nepal. Adapted from World Health Organisation Dengue Case Classification 2009.[1]

Department. Cases with warning signs, or cases already classed as severe dengue require admission to hospital. A systematic review of factors associated with severe dengue found that all the warning signs shown in Figure 6.2 are associated with increased risk of severe dengue disease.[319, 320]

A biomarker which can predict severity of dengue early in the illness presentation could be used in conjunction with existing clinical classification systems and could add important clinical information.[320]

6.1.6 Hypotheses

In this chapter, I present the results from a prospective cohort study which recruited children with different infections, and the use of molecular diagnostics, in a tertiary level hospital in Nepal. Additional molecular diagnostic testing was

performed after the resolution of the clinical cases.

Two hypotheses are tested in this chapter:

1. In the majority of cases, routine diagnostic tests fail to identify the cause of paediatric infection.

2. To investigate whether molecular diagnostics can aid diagnosis in this setting.

Aims and objectives

Linked to hypothesis one - In the majority of cases, routine diagnostic tests fail to identify the cause of paediatric infection. I aim to describe the causes of fever in Nepali children with additional routine clinical investigations.

1. Using medical record data and the results of routine investigations at Patan Hospital, a cause of infection is assigned to cases where possible.

2. Cases are grouped into different diagnostic groups to classify the different causes of infection.

3. Define cases at risk of severe dengue infection and the outcomes in these high-risk cases.

Linked to hypothesis two - Molecular diagnostics can aid diagnosis in this setting. I aim to show that additional molecular diagnostic testing adds diagnostic information which has the potential to change clinical management.

1. Summarise the molecular pathogen results.
2. Describe the potential additional diagnostic benefit of molecular testing.
3. Identify cases where molecular results could have changed clinical management.
4. Identify cases of agreement and disagreement between routine diagnostics and molecular testing.
5. In the dengue participants, use additional molecular results to describe the:
   (a) Most useful diagnostic tests at different stages of dengue disease.
   (b) Additional value of dengue PCR testing.
   (c) Dengue serotypes circulating.

6.2 Methods

6.2.1 Study design

A prospective cohort study was undertaken at Patan Hospital, Lalitpur, Nepal. Children up to 14 years of age, presenting to hospital with signs of infection, were eligible for enrolment. A cohort of children presenting to Patan Hospital without signs of infection were also enrolled to act as healthy controls.

After enrolment, research samples were obtained at three study time points; as early as possible after presentation to hospital (time point 1), during the hospital admission for participants who had were admitted for over 24 hours (time point 2), and 6-12 weeks after admission (time point 3) when the participant had recovered.

Ethical approvals

Ethical approval to undertake the study was obtained from the Ethical Review Board of the NHRC (reference number: 1937). Institutional approval was obtained from the Institutional Review Committee (IRC) of Patan Academy of Health Sciences (IRC reference: drs2102161480). Approval was also obtained from the University of Oxford's, OxTREC (reference: 55-20).

Enrolment

Enrolment began on 1st September 2021. The planned study period was from September 2021 to August 2023. However, as recruitment was progressing well

at the site in Nepal and extra funding became available, the enrolment period was extended to September 2024.

Setting

Patan Hospital, Lalitpur, Nepal is a 640-bed teaching hospital. Patan Hospital serves a large urban population in the Kathmandu Valley and is also a referral centre for complex paediatric cases from other parts of Nepal. This was the same site as the LRTI cohort study described in previous chapters.

Population

Cases

Inclusion criteria Participants could be included if they fulfilled the following criteria:

- Any child, up to 14 years of age, presenting to Patan Hospital
- And who has one or more of the following:
  - Fever (temperature $\geq$ 38.0°C) and/or history of fever in the preceding 24 hours
  - Other signs of infection identified by the admitting physician
- Consent is obtained for blood samples to be obtained for research purposes

Fourteen years of age is the upper age limit for admission to the paediatric ward at Patan Hospital.

Exclusion criteria Participants were excluded from the study if any of the following criteria were met:

- Participants do not consent to research blood sampling.
- Preterm neonates, less than 37 weeks corrected gestational age.
- Blood sample for RNA not obtained.

Controls A cohort of children, also up to 14 years of age, presenting to Patan Hospital without signs of infection were enrolled to act as non-infectious controls for comparison in future analyses. Control participants included patients presenting to hospital for minor trauma, vaccination, elective surgery or routine health checks. Control participants were excluded if they had a febrile illness or vaccination within the preceding three weeks.

Informed consent

Consent prior to study procedures When potential participants were identified, parents or guardians were approached by trained research staff. Parents/guardians were given verbal and written information about the study, and they were given time to ask questions, read the information and decide if they wanted their child to be included in the study. Written assent was obtained from children 12-14 years of age.

Deferred consent In situations where the patient required emergency management or it was unsafe for the research staff to approach the participant, a

deferred consent approach was used. This approach allowed for research samples to be taken before informed consent was obtained.

As soon as possible after research samples were taken using the deferred consent approach, parents/guardians were approached and written informed was sought. If informed consent was given the participants were included in the study and the study procedures were continued. If informed consent was refused, study samples which had been obtained were destroyed and the participant was not included in any further study procedures.

Deferred consent is important in this study, as an RNA blood sample taken before any medical intervention is the most useful sample for developing RNA signatures. RNA samples taken after medical interventions risk introducing bias caused by RNA transcript changes due to the treatment given.

6.2.2 Study procedures

Figure 6.3 outlines the different study steps after a potential participant presents to Patan Hospital. At each time point three blood samples are taken, an RNA-stabilisation tube, ethylenediaminetetraacetic acid (EDTA) and serum. A nasopharyngeal swab sample is obtained at time point 1 only.

The RNA sample was deemed the most important sample to the primary study endpoint, and a minimum of one RNA sample was required for a child to continue in the study. The samples taken in order of preference were:

1. RNA – PAXgene® Blood RNA Tube, for stabilisation of intracellular RNA

2. EDTA

3. Serum

PAXgene tubes are used to preserve RNA. The commercial PAXgene tubes require 2.5 mls of blood to be added to each tube.[321] As obtaining adequate blood volume is often an issue in paediatric studies, the reagent in these commercial PAXgene tubes was aliquoted into smaller ribonuclease (RNase)-free tubes, to make versions of the commercial PAXgene tube which requires smaller volumes. These smaller RNase-free tubes required only 1ml of blood to be added.

Blood samples were collected in a syringe. The first 1 ml of blood was transferred directly to an RNA stabilisation tube. The next 2 mls of blood were transferred to the EDTA tube. The next 1ml of blood was transferred to the serum tube. Any remaining blood was divided between EDTA and serum tubes.Maximum blood volume obtained complied with European Union (EU) and WHO guidelines for blood loss associated with paediatric research.[322, 323] After samples were obtained, they were transferred to the laboratory within one hour.

A nasopharyngeal sample was also obtained from participants, at time point 1 only. The nasopharyngeal swab was placed in a cryovial with transport medium containing STGG.

Time point 1 When an eligible child presented to hospital, informed consent was obtained prior to research sampling, then research samples were obtained. Research blood samples were obtained at the same time as clinical samples, if possible, to reduce the number of blood draws that a child needed to undergo.

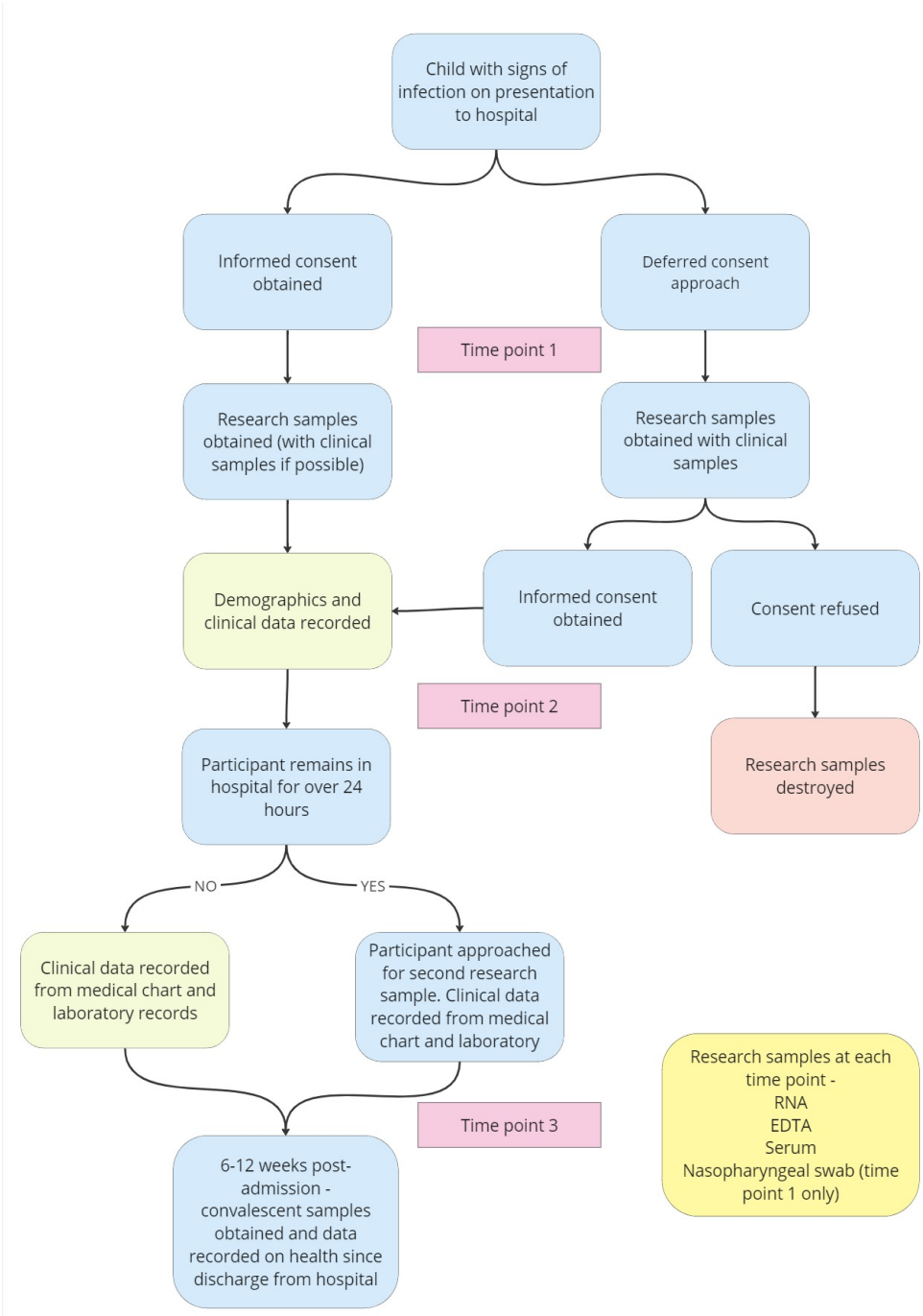


Figure 6.3: The study procedures when a potential participant in the fever cohort study presents to hospital. The steps to be followed with the deferred consent approach are shown and study procedures at each time point. RNA, ribonucleic acid; EDTA, ethylenediaminetetraacetic acid.

Demographic information and details of initial clinical presentation were also collected.

If the deferred consent approach was used, research blood samples were obtained with clinical samples. These research samples were stored and the patient's parent/guardian was approached as soon as it was appropriate to obtain informed consent.

Time point 2 If the participant remained in hospital for over 24 hours, they were approached to give a second set of research blood samples. This was an optional time point. This second time point involved the same research blood samples as time point 1, obtained 24-72 hours after admission, to identify for changes in biological markers, such as RNA expression or protein levels, as the course of illness progressed.

Medical chart review Regardless of whether time point 2 samples were obtained or not, the medical chart of each participant was reviewed 24-72 hours after admission, and again at discharge from hospital. Data were collected on investigations during admission, treatment given, level of care required and outcome of the illness. The laboratory system results were also reviewed and results recorded.

Time point 3 Following discharge, participants were asked to return to hospital 6-12 weeks after their admission. The same research blood samples as the previous time points were obtained again, following confirmation of ongoing consent. Data were collected regarding the condition of the participant since discharge from

hospital, including any illnesses or vaccinations since discharge.

6.2.3 Classification of cases based on routine investigations

Cases were classified into different diagnostic groupings based on a uniform system used across the consortium sites, see Figure 6.4. It is important for the RNA-seq experiment, and other future analyses, that cases be assigned to their correct classification.

The classification system used in this study is similar to the system used for the LRTI cohort. The differences in the classification system used for the cohort in this chapter reflect the more diverse group of participants that could be recruited to the fever cohort, including participants with other non-bacterial, non-viral infections, and cases which present with signs of infection but have a non-infectious cause.

The following steps were followed when classifying each participant:

1. Enrolled participants were first classified based on the clinical data only.
2. Results of routine investigations were reviewed and these were used to classify participants into their primary classification groups. The investigations and thresholds used for classification are listed in Table 6.1.

Cases were grouped into 11 different categories, as listed in Table 6.1. The most important classifications for analysis were the definite bacterial and definite viral groups. Cases were classified as definite bacterial if the participant presented

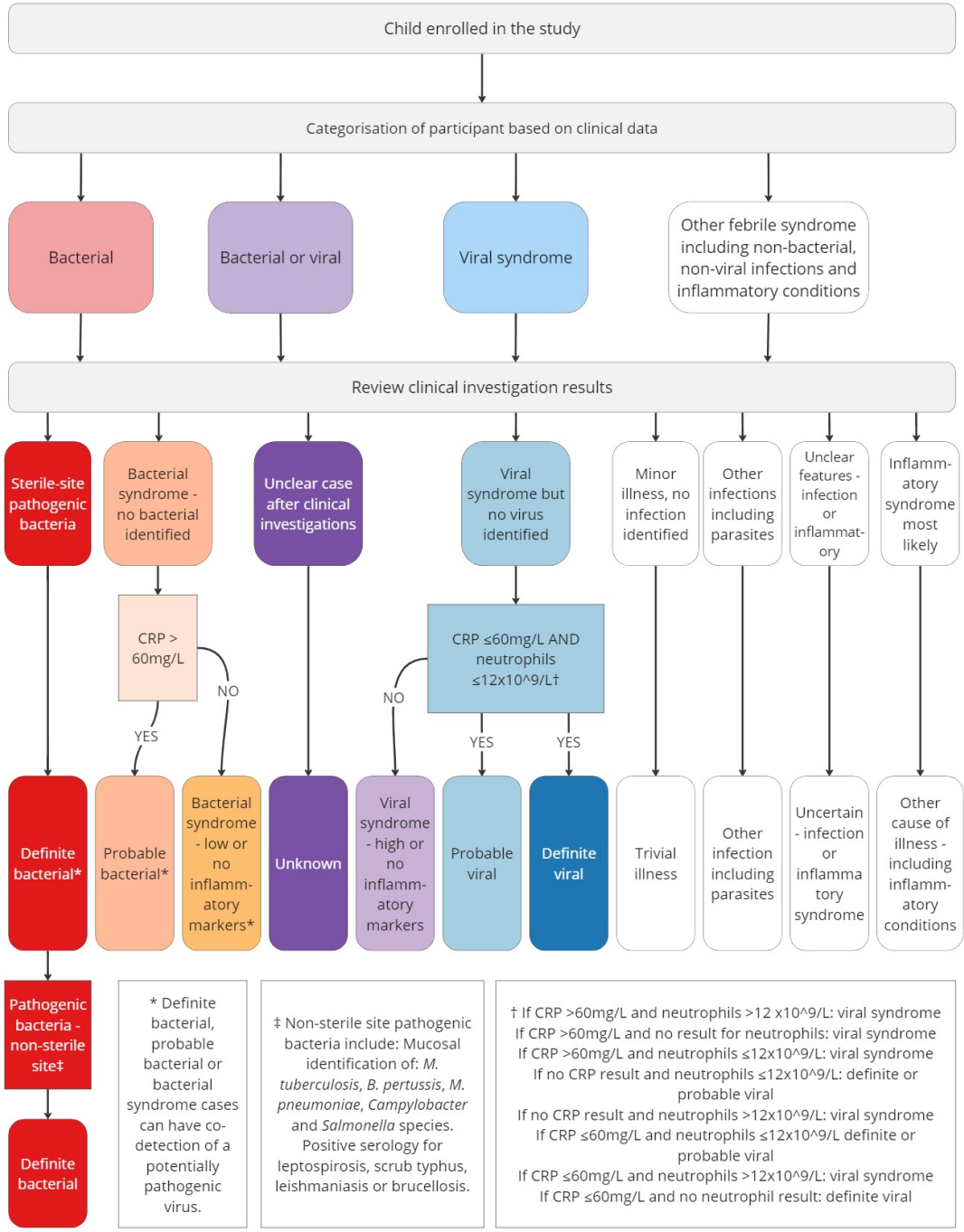


Figure 6.4: System used in the fever cohort study to group cases into their various classifications for further analyses. CRP, C-reactive protein.

Table 6.1: Classification groupings to be used in the analysis with the criteria used to classify the cases. CRP, C-reactive protein.

Classification	Criterion 1	Criterion 2	Criterion 3
Definite bacterial – sterile site	Clinically a bacterial syndrome	Sterile-site pathogenic bacteria	
Definite bacterial – non-sterile site	Clinically a bacterial syndrome	Non-sterile site pathogenic bacteria	
Probable bacterial	Clinically a bacterial syndrome	No pathogenic bacteria identified	CRP > 60mg/L
Bacterial syndrome - low or no inflammatory markers	Clinically a bacterial syndrome	No pathogenic bacteria identified	CRP ≤ 60mg/L
Unknown	Unclear features after clinical investigations	No pathogen which fits case identified	
Viral syndrome - high or no inflammatory markers	Clinically a viral syndrome	No pathogenic virus identified	CRP > 60mg/L OR Neutrophils >12x10^9/L OR no result for CRP or neutrophils
Probable viral	Clinically a viral syndrome	No pathogenic virus identified	CRP ≤60mg/L AND neutrophils ≤12x10^9/L
Definite viral	Clinically a viral syndrome	Pathogenic virus identified	CRP ≤60mg/L AND neutrophils ≤12x10^9/L
Trivial illness	Minor illness	No pathogen identified	
Other infections including parasites	A non-bacterial, non-viral pathogen identified	No bacterial or viral pathogen identified	
Uncertain – infection of inflammatory syndrome	Unclear whether clinical presentation fits with infective or inflammatory process	No pathogen identified	
Other cause of illness – including inflammatory syndromes	A non-infectious cause of illness most likely based on clinical investigations	No pathogen identified	

with a clinical syndrome which is commonly bacterial in origin, and a bacterial pathogen was isolated from a normally sterile site, including blood, pleural fluid, CSF or urine. Cases could also be grouped as definite bacterial if the bacterial pathogen was found in a non-sterile site and the case fulfilled the following criteria:

1. Pathogen identified by:

   (a) Mucosal detection of *M. tuberculosis*, *B. pertussis* or *M. pneumoniae* OR

   (b) Nucleic acid amplification test positive for *M. tuberculosis* (Gene Xpert *Mycobacterium tuberculosis* (MTB)/rifampicin (RIF) assay) OR

   (c) IgM positive for leptospirosis, scrub typhus, leishmaniasis or brucellosis

2. The clinical syndrome can be fully explained by the pathogen detected

3. No other pathogen is identified which explains the clinical picture

Cases were classified as definite viral if they had clinical presentation consistent with a viral syndrome and a likely pathogenic virus was identified, in the absence of the identification of pathogenic bacteria. The criteria used to classify other cases are described in Table 6.1. These cases were classified only using investigations which are part of routine care at Patan Hospital.

6.2.4 Sample processing

Once the samples arrived in the laboratory the RNA stabilisation samples were stored at room temperature for two hours before being transferred to a -80°C freezer, as per the manufacturers recommendations.[321]

When the EDTA sample arrived in the laboratory an aliquot of whole blood was taken from the sample. If volume allowed, an aliquot was transferred to a proteomic stabiliser tube (Smart Tube Inc.). The remaining EDTA sample was centrifuged, plasma samples were aliquoted, and a cell pellet sample was stored. All aliquots were frozen at -80°C.

Serum samples were allowed to clot and separate. Then the sample was centrifuged, and aliquots of serum stored at -80°C.

When the nasopharyngeal swab sample arrived in the laboratory an aliquot of STGG was taken and cultured for pneumococcal serotypes, using the Quellung method. This was part of another study running at the same site. A second aliquot of STGG was obtained and stored at -80°C.

6.2.5 Molecular testing

An aliquot of STGG from each nasopharyngeal sample was sent to Micropathology Ltd., University of Warwick Science Park, UK for molecular testing. A whole blood sample from the EDTA tube was also sent to Micropathology Ltd., UK for molecular testing.

An aliquot of plasma was tested using the Siemens Tropical Fever Core panel at

Patan Hospital. If a plasma sample was positive for dengue on PCR, this sample was tested using the Siemens Dengue Differentiation panel.

Total nucleic acid was extracted from the nasopharyngeal swab STGG media, and from the blood samples. Nucleic acid from the nasopharyngeal samples were analysed using the NxTAG™ Respiratory Pathogen Panel + SARS-CoV-2 (Luminex® Corporation). This panel detects 17 different viral targets and 3 bacterial targets.

Extracted nucleic acid from blood samples were analysed using a molecular diagnostic panel developed and validated at Micropathology Ltd. This panel allows the detection of 9 different viruses (9 viral targets), 16 bacteria (23 targets), and 2 fungal species (6 fungal targets). See Figure 6.1 for the full list of pathogen targets.

Molecular testing at Patan Hospital

The laboratory protocols for extraction of nucleic acid and use of the Tropical Fever Core Panel can be found in Appendix F.

Extraction of nucleic acid After the plasma sample was thawed, the sample was combined with lysis buffer and isopropanol. The mixture was transferred to an absorption column and centrifuged. An internal control was then added. The first wash buffer was added, and the solution was centrifuged. The liquid was then discarded and the second wash buffer was added to the absorption tube, and then centrifuged again. The liquid was discarded again, the absorption column was allowed to dry before RNase-free water was added. Following a final

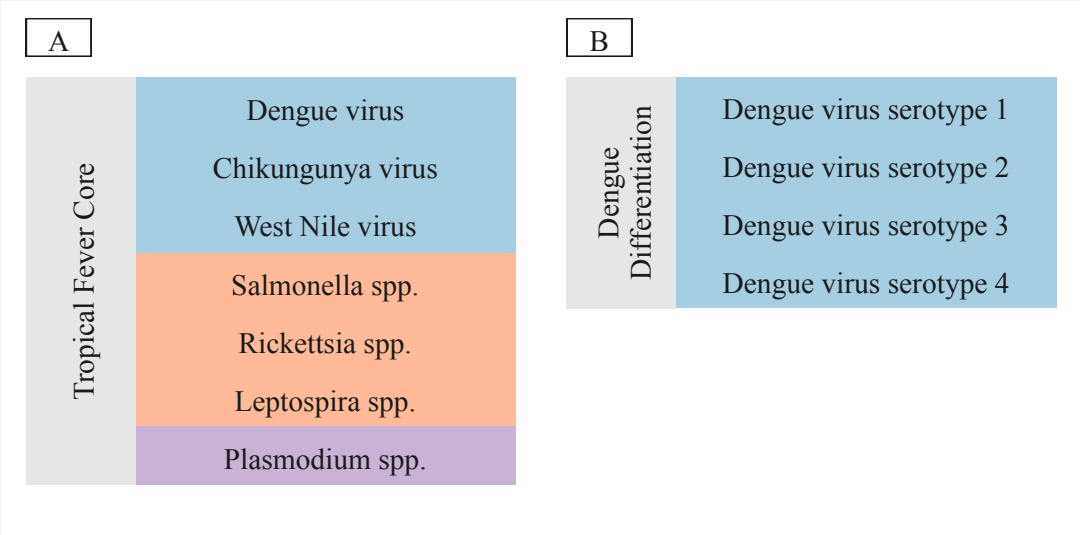


Figure 6.5: Targets tested on molecular panels at Patan Hospital. Panel A shows the targets included in the Siemens Tropical Core Panel. Panel B shows the targets included in the Siemens Dengue Differentiation panel. spp., species.

centrifuge step the extracted nucleic acid solution was either used immediately or stored at -80°C.

Tropical Fever Core Panel The Siemens Tropical Fever Core Panel is designed to identify DNA sequences associated with specific pathogens, using plasma or serum samples. This panel is recommended for research-use only currently. Targeted RNA is reverse transcribed to complementary DNA (cDNA). Using real-time PCR, the DNA molecules are then simultaneously amplified. If increased fluorescence is detected from the probe, this indicates the presence of the specific DNA sequences of interest.

The Siemens Tropical Fever Core Panel has seven pathogen targets: dengue virus, West Nile virus, chikungunya, *Rickettsia* species, *Salmonella* species, *Plasmodium* species and *Leptospira* species, see Figure 6.5. See Appendix F for gene targets.

The PCR master-mix was prepared by combining the primer-probe, enzyme and buffer supplied in the kit. Then 10 µl of master-mix was added to the wells on the PCR plate. Next, 10 µl of extracted nucleic acid, positive control, or negative control was added to each well. The plate was loaded into the PCR machine (BioRad CFX96) and the PCR program was run. Results were reviewed using the BioRad software.

Dengue Serotype Panel If a sample was positive for dengue on the Tropical Fever Core Panel, extracted nucleic acid from that sample was analysed using the Siemens Dengue Differentiation Panel. This panel uses the same principle as the Tropical Fever Panel except the targets are the four dengue serotypes. The steps undertaken to run the Dengue Serotype Panel are similar to the Tropical Fever Core Panel steps described above, see also Appendix F.

6.2.6 Classification of cases incorporating molecular testing

A subset of the additional molecular pathogen results was used to re-classify cases into new diagnostic groups. The molecular results selected were based on the evidence outlined in Section 6.1.3. Other molecular results were used for descriptive purposes.

Cases were re-classified using the algorithmic approach described here, and, later, the clinical details of cases were reviewed to see if the re-classifications made sense clinically.

Blood bacterial molecular results from the Micropathology in-house molecular panel used to re-classify cases were *N. meningitidis*, *S. pneumoniae*, *E. coli* and Group A *Streptococcus*. Bacterial molecular result from the Siemens Tropical Fever Core Panel used to re-classify cases were *Rickettsial spp.*, *Leptospira spp.* and *Salmonella spp.* If any of these bacterial targets were positive cases were re-classified as per Figure 6.6. Following re-classification, a new classification was added to account for co-infections. Cases were classified as bacterial-viral co-infection if both bacteria and a virus have been detected, and both pathogens could account for some of the features of the presentation.

Depending on whether the result came from a nasopharyngeal or blood sample, molecular viral results were used in different ways in re-classifying cases as definite viral.

Viral molecular blood results used to re-classify cases were dengue virus from the Tropical Core Fever Panel and enterovirus from the Micropathology in-house bacterial panel. The remaining viruses in the Tropical Core Fever Panel were not expected to be found in Nepal, chikungunya and West Nile viruses. The justification of only including enterovirus is outlined in Section 6.1.3. Other positive viral molecular results were used for descriptive purposes. The justification for including only these two viruses is explained in Sections 6.1.4 and 6.1.3.

If a case was positive for either of these viruses the case was re-classified as per Figure 6.7.

For nasopharyngeal results, cases were only re-classified if cases had a respiratory presentation; a respiratory presentation was when one of the following features was documented: cough, increased work of breathing, increased respiratory rate,

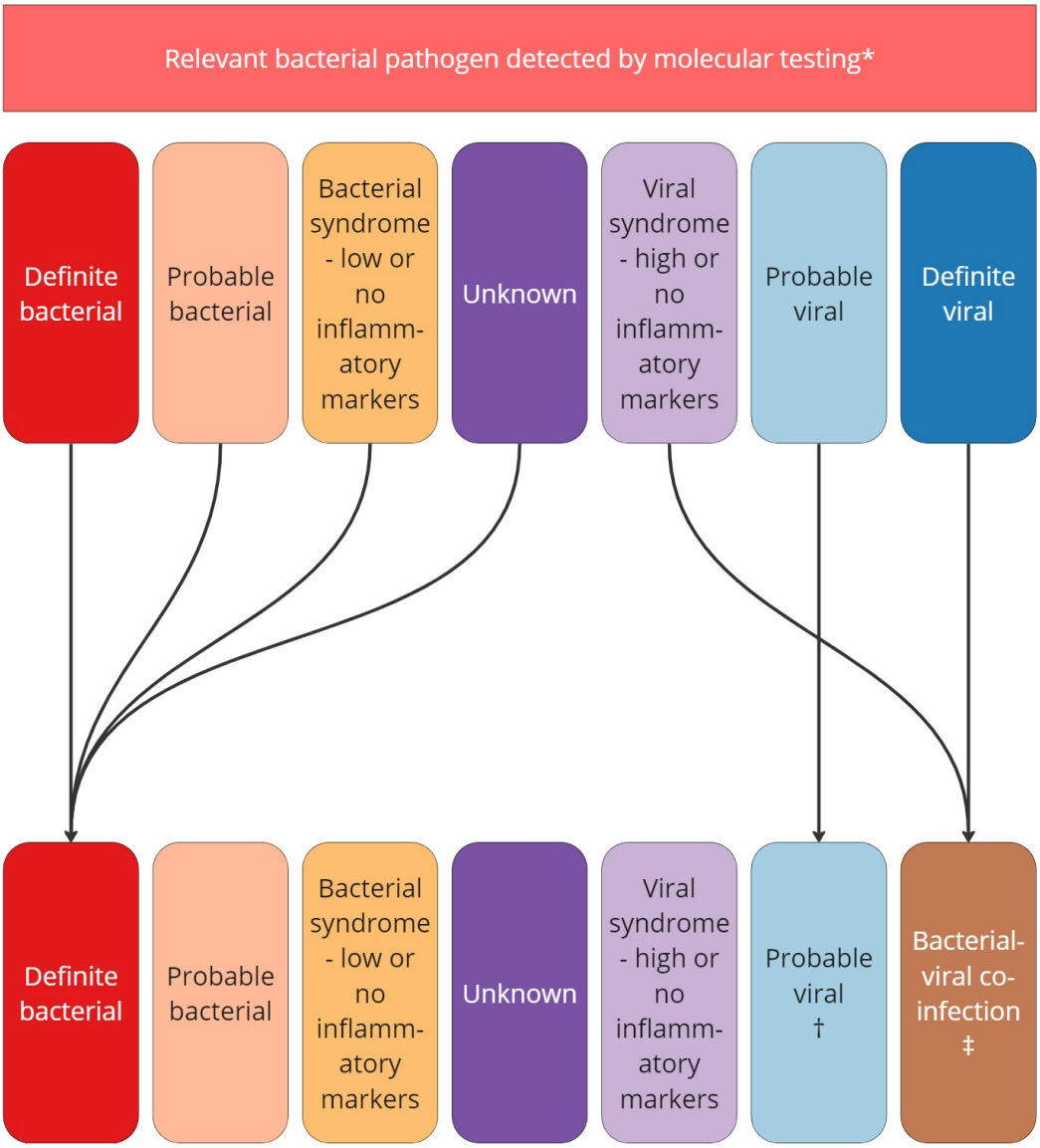


Figure 6.6: Re-classification of cases in the fever cohort study if one or more of the selected bacterial molecular targets were positive. *Cases were re-classified if *N. meningitidis*, *S. pneumoniae*, *Rickettsial spp.*, *Leptospira spp.* or *Salmonella spp.* were detected. † Probable viral infections were not re-classified but probable viral infections where a bacterial pathogen was detected were recorded as possible co-infections. ‡ Bacterial-viral co-infection is a new classification where both a potentially pathogenic virus and bacteria have been detected.

required supplementary oxygen or diagnosed as a respiratory tract infection. The viral nasopharyngeal targets used to re-classify cases were RSV (type A and B), influenza (type A and B), parainfluenza (types 1, 3 and 4) and human metapneumovirus. These are the same LRTI-associated viruses used to classify LRTIs in the previous LRTI cohort.

A new classification of bacterial-viral co-infection was also used. Definite bacterial cases were re-classified as bacterial-viral co-infections when a viral molecular pathogen was identified which could be contributing to the clinical presentation.

6.2.7 Co-infections

Co-infections were not part of the initial classification system which was used throughout the consortium sites. I have added the co-infection classification as I have an interest in this group as co-infections can create difficulties with clinical decision-making. To be included in this group it must be reasonable for both bacterial and viral pathogens to be contributing to the clinical picture.

6.2.8 Pathogen groups

Cases were also grouped together based on the individual pathogens detected. These pathogen groups were used to describe the causes of infection in this setting. These pathogen groups can be examined during future RNA-seq and protein analyses.

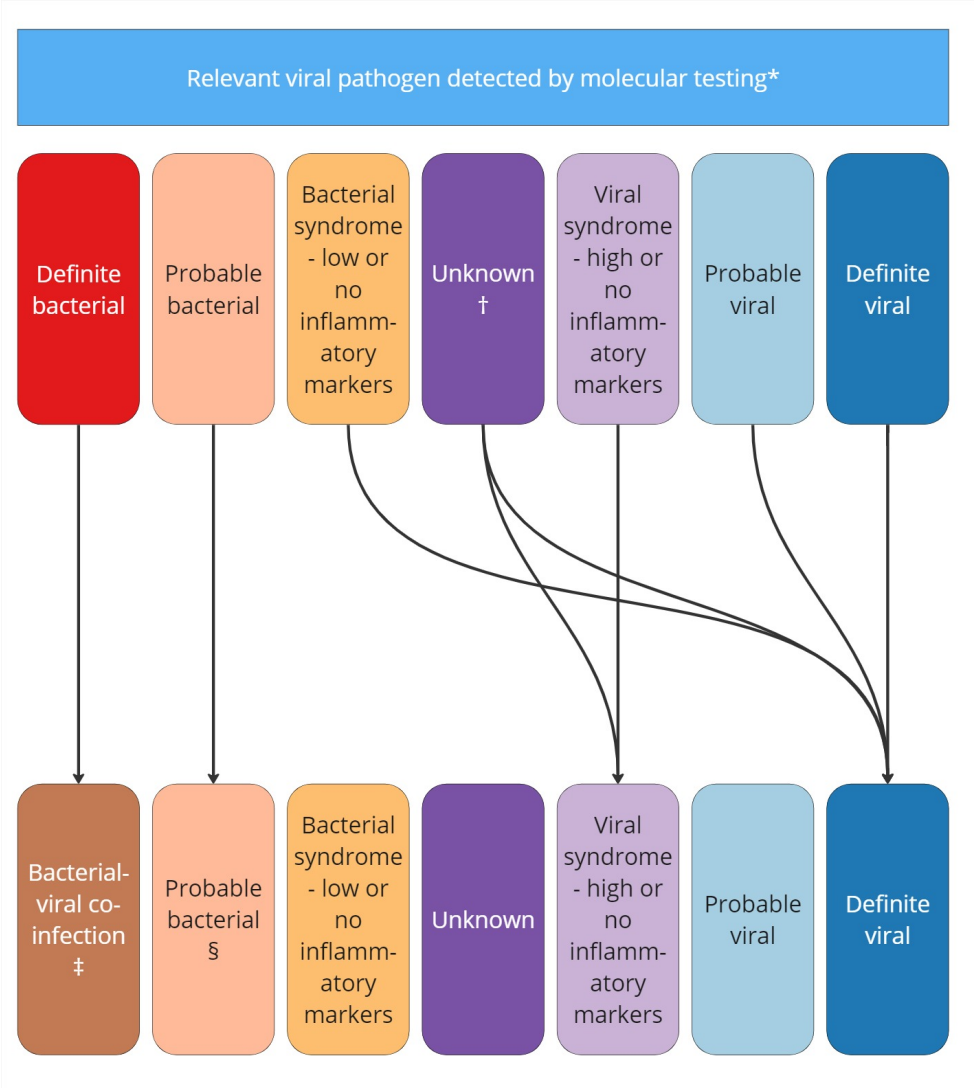


Figure 6.7: The system used to re-classify cases if a relevant viral molecular test was positive. * If enterovirus or dengue virus were detected in the blood, all cases were re-classified as per the figure. If RSV (A or B), influenza (A or B), parainfluenza (1, 3 or 4) or human metapneumovirus were detected, and the case was classified as a respiratory presentation then the case was re-classified. † Unknown cases could be re-classified as viral syndrome – high or no inflammatory markers or re-classified as definite viral cases depending on inflammatory marker levels. ‡ Definite bacterial cases were re-classified as bacterial-viral co-infection if a viral molecular pathogen was identified which could be contributing to the clinical presentation. § Probable bacterial cases where a viral pathogen was identified were not re-classified, but probable bacterial cases where a viral pathogen was detected were recorded as possible co-infections.

6.2.9 Severity

Dengue fever cases in the cohort were classified as severe if they met the WHO severe dengue criteria. I compared outcomes for dengue cases with and without warning signs, see Figure 6.2.

In the whole cohort, and within the dengue subgroup, I looked at cases which had poor outcomes, including death, paediatric intensive care unit (PICU) admission, change in disability level at discharge, or prolonged hospital stay. I looked for data points collected at admission which were associated with poor outcomes.

6.3 Results

6.3.1 Description of cohort

From 1st September 2021 to 19th April 2023, 574 participants were enrolled as cases. During the same period, 113 control participants, without signs of infection, were enrolled. Control participants were made up of children presenting to the hospital for non-infectious causes including minor injuries or for vaccinations.

Table 6.2 shows the number of participants enrolled, divided into their diagnostic classifications. The median age of the cases in the cohort was 3 years and 35.2% of the cohort were female. The control group had a median age of 4.5 years with 37% females. CRP, blood white cell count (WCC) and neutrophils were higher in the bacterial, compared with the viral groups, as expected. The duration of hospital stay was longest in the definite bacterial group.

The classifications in Table 6.2 were based on diagnostic tests ordered by the clinical team caring for each patient. The viral diagnostics were mostly based on two tests, NS1 antigen testing for dengue fever (n=85 positive cases), and SARS-CoV-2 PCR testing (n=13 positive cases). Many of the definite bacterial cases were urinary tract infections with urine culture positive for *E. coli* (n = 20 positive cases). There were 13 cases where blood, pleural fluid and/or CSF were positive for the likely causative pathogen including *S. pneumoniae* (n=4), *S.* Paratyphi A (n=4) and *S.* Typhi (n=3). Table 6.7 below shows a full list of pathogens assigned during routine investigations.

Table 6.2: Number of cases and healthy controls recruited to the cohort study, grouped by diagnostic classification. Demographics and inflammatory markers at hospital admission are presented. CRP, C-reactive protein; WCC, blood white cell count. Cases classified as, other cause of illness (n=2), other infection including parasites (n=4) and, uncertain - infection or inflammatory syndrome (n=3) were excluded from the table.

	Definite viral	Probable viral	Viral syndrome - high or no inflammatory markers	Unknown	Bacterial syndrome - low or no inflammatory markers	Probable bacterial	Definite bacterial	Total cases	Healthy controls
Number of cases	104	77	29	113	145	58	39	574	113
Age, years (median)	8.81	1.99	3.73	2.12	2.15	3.80	2.49	3.02	4.51
Female (%)	27.88%	32.47%	27.59%	32.74%	37.24%	46.55%	43.59%	35.19%	37.17%
Duration of hospital stay, days (median)	4	4	4	6	5	6	9.5	4	NA
CRP (mg/L)	5.00	9.00	61.20	6.10	14.00	90.05	43.35	15.30	NA
WCC x10^9/L	4.60	10.00	17.20	11.20	10.60	15.75	12.00	10.00	NA
Neutrophils x 10^9/L	2.45	5.59	12.88	6.34	6.23	11.20	7.77	5.75	NA

Pathogens identified

The 104 cases classified as definite viral using routine investigations included dengue fever cases, 84.6% (88/104), SARS-CoV-2 positive cases, 11.5% (12/104) and hepatitis A, 3.8% (4/104).

The 39 definite bacterial cases included 20 *E. coli*-positive cases from urine culture. One of the *E. coli* cases also had a positive blood culture for *E. coli*. There were seven culture-positive enteric fever cases (four *S.* Paratyphi A and three *S.* Typhi). Four cases were positive for *S. pneumoniae*. There were smaller numbers of other bacteria isolated, see Table 6.7.

6.3.2 Discharge diagnosis based on clinical syndrome

Cases were separately classified based on their clinical syndrome at discharge, as documented by the clinical team in the medical notes. Table 6.3 shows that 40.2% (231/574) of cases were LRTIs, 15.2% (87/574) of cases had a gastrointestinal diagnosis and 6.4% (37/574) of cases were primarily neurological. One per cent (6/574) of cases were given a sepsis diagnosis.

Table 6.3: Discharge diagnoses as documented by the clinical team, grouped by clinical syndrome. The number of cases in each group and the percentage that each group contributes to the overall cohort are also shown.

Discharge diagnosis - clinical syndrome	Number of cases	Percentage of cases (total cases = 574)
Lower respiratory tract infections	231	40.2%
Pneumonia	196	34.1%
Bronchiolitis	20	3.5%
Gastrointestinal	87	15.2%
Enteric fever - clinical or culture diagnosis	45	7.8%
Culture-confirmed enteric fever	7	1.2%
Neurological	37	6.4%
Meningitis	12	2.1%
Encephalitis	7	1.2%
Urinary tract infections	33	5.7%
Upper respiratory tract infections	29	5.1%
Fever without source	22	3.8%
Skin and soft tissue	7	1.2%
Sepsis	6	1%
Musculoskeletal	4	0.7%
Inflammatory	2	0.3%
Other	2	0.3%
Immunocompromised at presentation	6	1%

Outcome of hospital admission

Participants were classified as having a poor outcome if they required intensive care admission, their discharge disability level was lower than their baseline disability, or if they died.

Participants who did not require intensive care, were discharged well without a new disability and had a hospital stay of fewer than seven days were considered to have had a good outcome. See Table 6.4 for a comparison of baseline characteristics for those with and without poor outcomes. Note, participants who did not meet the criteria for a poor outcome but had a prolonged hospital admission, over 7 days, were not included in either the poor or good outcome groups. Four participants, 0.7%, unfortunately died during their hospital admission.

6.3.3 Additional molecular testing

The use of additional molecular panels depended on the availability of samples. In cases where a nasopharyngeal swab or an EDTA sample were not obtained additional molecular testing was not possible.

Samples from 498 participants were tested using at least one additional molecular panel. Samples from 53 participants were tested using all three molecular panels. At least one molecular target was positive in 72.1% (359/498) of samples. Of the cases with a respiratory presentation, 69.7%(249/357) had a nasopharyngeal sample tested for respiratory pathogens. Table 6.5 shows the number of samples tested using each molecular panel.

Table 6.5 shows that 78% of cases had a positive result using the Micropathology

Table 6.4: Cases divided by outcome. Cases with a poor outcome required ICU admission, had a decrease in their baseline disability level at discharge or died. Cases with a good outcome were discharged well without a new disability and had a hospital stay of fewer than seven days.

	Poor outcome n=48	**Good outcome n=434**
Final classification - additional molecular testing included		
Bacterial (definite and probable)	12.5% (n=6)	13.8% (n=60)
Definite bacterial	4.2% (n=2)	4.4% (n=19)
Viral (definite and probable)	39.6% (n=19)	46.1% (n=200)
Definite viral	29.2% (n=14)	35% (n=152)
Known comorbidity	2.1%	2.1%
Female	50%	31.6%
Age, years (median)	1.99	3.1
Antibiotics in the seven days prior to admission	25%	26.5%
CRP (mg/L)	26	14
WCC x10^9/L	13.2	9.3
Neutrophils x 10^9/L	7.81	5.48
Tachypnoea for age†	54.2%	18.7%
Fever with significant tachycardia‡	16.67%	9.7%
Received EPI vaccines§	97.9%	99.8%

* Classification of cases following additional molecular testing using selected targets.
† Documented by clinician at admission.
‡ Temperature over ≥ 38°C with tachycardia (< 1 year of age with heart rate ≥ 170 beats/min, 1-2 years of age with heart rate ≥ 160 beats/min, 2-5 years of age and ≥ 150 beats/min, ≥ 5 years of age with heart rate ≥ 130 beats/min.
§ Received at least one dose of vaccine as per the Nepali national Expanded Program on Immunization (EPI) programme.

Table 6.5: The number of samples tested using each of the additional molecular panels. Cases with any positive result are shown as a percentage of the number of cases tested using that panel. Cases with a positive result for a pathogen used in the re-classification process are shown as a percentage of the number of cases tested using that panel. Molecular targets used to re-classify cases, respiratory panel: RSV A and B, influenza A and B, parainfluenza 1, 3 and 4, human metapneumovirus; blood targets: *S. pneumoniae*, *N. meningitidis*, *E. coli*, *Rickettsia spp.*, enterovirus, dengue virus. PCR, polymerase chain reaction.

	Respiratory panel (Micropathology)	**Blood PCR (Micropathology)**	**Tropical Core Panel (Patan)**	**Total cases**
Number of cases tested	379	325	82	498
Any positive result (% of cases tested)	43.54%	78.15%	32.93%	72.09%
Positive result for a pathogen which could alter classification (% of cases tested)	25.33%	3.69%	32.93%	27.11%

Ltd. blood PCR panel, however, less than 4% of the blood PCR panel results could have altered the classification of cases. A quarter of the cases tested with the respiratory panel had a positive result which could have altered management. Using the Tropical Core panel, 33% of cases tested had a positive result for at least one of the molecular targets, and all of the positive results could have altered the classification of cases.

6.3.4 Re-classification of cases following molecular testing

As outlined in Section 6.2.6, a limited number of molecular results were used to re-classify cases. Following the addition of these results from the molecular panels to the clinical investigations, 16.9% (84/498) of cases were reclassified. Using routine investigations only, 28.7% (143/498) of cases had a definite diagnosis, bacterial or viral, compared with 45.6% (227/498) following additional molecular testing.

Table 6.6 and Figure 6.8 show that the greatest change in classification was in the definite viral group, with 77 extra cases assigned to a definite viral classification. The seven extra definite bacterial cases identified represented an 18% increase in the number of definite bacterial cases.

6.3.5 Pathogens used for re-classification

Viral pathogens. Eighty-four cases were re-classified to one of the definite groups following additional molecular testing, see Table 6.7. The respiratory panel provided the majority of the results used, 30 RSV cases, 25 influenza cases, 14 human metapneumovirus cases and 8 parainfluenza cases.

Six additional SARS-CoV-2 cases were also identified. All of these cases presented with respiratory symptoms, none of them had a documented test result for SARS-CoV-2 as part of their routine investigations.

Seven cases were enterovirus-positive in blood. Of these, four had a lumbar

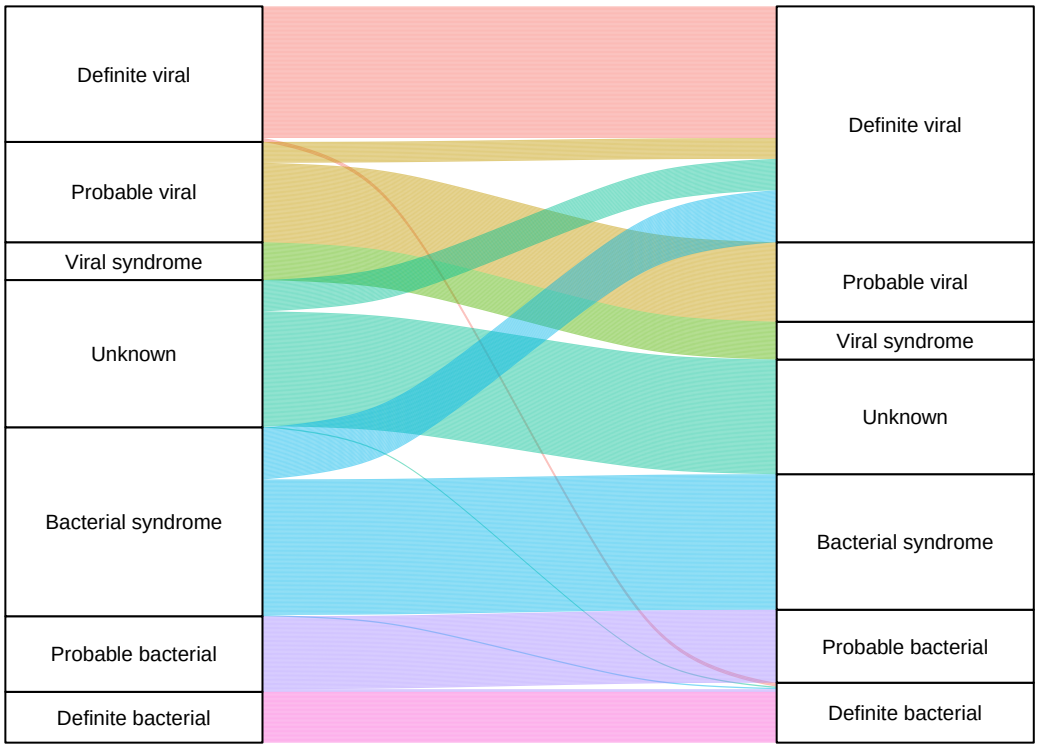


Figure 6.8: The classification of cases in the fever cohort study with routine investigations and the change in classification following additional molecular testing.

Table 6.6: Number of cases in each classification group using routine investigations, and number of cases in each group following re-classification of cases using molecular panels. The percentage change in the number of cases in each classification group is also shown.

n=498	Definite viral	Probable viral	Viral syndrome - high or no inflammatory markers	Unknown	Bacterial syndrome - low or no inflammatory markers	Probable bacterial	Definite bacterial
Classification post-clinical investigations	104	77	29	113	145	58	39
Classification post-molecular panels	181	61	29	88	104	56	46
Change to classification in each group after molecular testing (percent)	+74.04%	-20.78%	0%	-22.12%	-28.28%	-3.45%	+17.95%

puncture performed and three had a discharge diagnosis of meningitis. One case had persistent vomiting and was diagnosed with bacillary dysentery. Two cases presented with respiratory symptoms and were diagnosed as bronchiolitis and pneumonia. All seven cases were treated with antibiotics with two cases receiving 21 days of intravenous antibiotics.

Bacterial pathogens. Three cases were positive for *S. pneumoniae* on PCR. Two of these cases were also positive for *S. pneumoniae* on blood culture. The third, blood culture negative, PCR-positive *S. pneumoniae* case, presented with chest indrawing and difficulty breathing. This participant had a prolonged hospital stay, complicated by pleural effusion, and was treated for pneumonia.

Three additional *E. coli* cases were identified using PCR. None of these cases

Table 6.7: List of pathogens contributing to the definite viral and definite bacterial cases. The number of cases identified by routinely available investigations and additional molecular testing are presented. Several cases had more than one pathogen identified. *Routinely available investigations at Patan Hospital include PCR for SARS-CoV-2. †Source of the positive test was either the respiratory PCR panel (NxTAG), the blood PCR panel (Micropathology Ltd.), the Tropical Core panel (Siemens) or routinely available rests. PCR, polymerase chain reaction; IgM, immunoglobulin M, IgG, immunoglobulin G.

Pathogen	**Cases identified using routine investigations at Patan Hospital***	**Additional cases identified by molecular testing panels**	**Total number of cases**	**Source of positive test†**
Definite viral cases	104	77	181	
Dengue	88		88	Blood antigen/PCR (routine)
RSV		30	30	Respiratory PCR (NxTAG)
Influenza		25	25	Respiratory PCR (NxTAG)
SARS-CoV-2	12	6	18	Respiratory PCR (NxTAG)
Human metapneumovirus		14	14	Respiratory PCR (NxTAG)
Parainfluenza		8	8	Respiratory PCR (NxTAG)
Enterovirus		7	7	Blood PCR (Micropathology)
Hepatitis A	4		4	IgM/IgG hepatitis A (routine)
Definite bacterial cases	39	7	46	
E. coli	20	3	23	Urine culture/blood culture/blood PCR (routine/Micropathology)
S. pneumoniae	4	1	5	Blood culture/blood PCR (routine/Micropathology)
S. Paratyphi A	4		4	Blood culture (routine)
S. Typhi	3		3	Blood culture/CSF culture (routine)
Klebsiella species	3		3	Urine culture (routine)
M. tuberculosis	2		2	Gene Xpert (routine)
O. tsutsugamushi	2		2	Scrub typhus IgM (routine)
N. meningiditis		2	2	Blood PCR (Micropathology)
S. aureus	1		1	Pleural fluid culture (Routine)
Rickettsia species		1	1	Blood PCR (Siemens)

had a urine culture tested as part of their routine investigations. All three cases presented with a short history of fever. One case presented with respiratory symptoms, was diagnosed as pneumonia and treated with amoxicillin intravenously. The other two cases were also positive for dengue on NS1 and PCR testing. These cases were treated as dengue fever, not given antibiotics and had short hospital admissions.

Two cases of *N. meningitidis* were identified. One of these cases presented with altered consciousness and CRP 327 mg/L. No CSF result was documented but the child was diagnosed as meningitis and treated with 14 days of antibiotics. The second case presented with respiratory symptoms including cough and wheeze, CRP 51 mg/L, was diagnosed as occult bacteraemia and treated with seven days of antibiotics.

One case was positive for *Rickettsial spp.* on PCR. This case was also positive for dengue virus on NS1 and PCR testing. This case presented with respiratory symptoms, was diagnosed as dengue fever with an exacerbation of asthma, did not receive antibiotics, and had a one-week stay in hospital.

Of the other bacterial pathogens which could have been used to re-classify cases, there were no positive cases of Group A *streptococcus* or *Mycoplasma pneumoniae*.

Of the seven culture-positive enteric fever cases, two had a plasma sample available and were tested using the Tropical Fever Core Panel, both of these samples were negative for *Salmonella spp.*

Molecular results not used for re-classification

Respiratory panel The NxTAG respiratory panel includes pathogens which were not strongly associated with LRTIs in the case-control studies discussed previously. These pathogens were not used to re-classify cases. These viruses include bocavirus, adenovirus, non-SARS-CoV-2 coronaviruses, rhinovirus and enterovirus. Figure 6.9 shows the number of LRTI cases per month, and the number of these viruses detected in these LRTI cases per month. Apart from one peak of adenovirus in mid-2022, detection of these viruses does not appear to be temporally-associated with peaks of LRTI admission. For comparison, Figure 6.10 shows the viruses used for re-classification of cases detected per month, with the number of LRTI cases.

There were no positive results among the three bacterial targets in the NxTAG respiratory panel, *Mycoplasma pneumoniae*, *Legionella pneumophila* or *Chlamydia pneumoniae*.

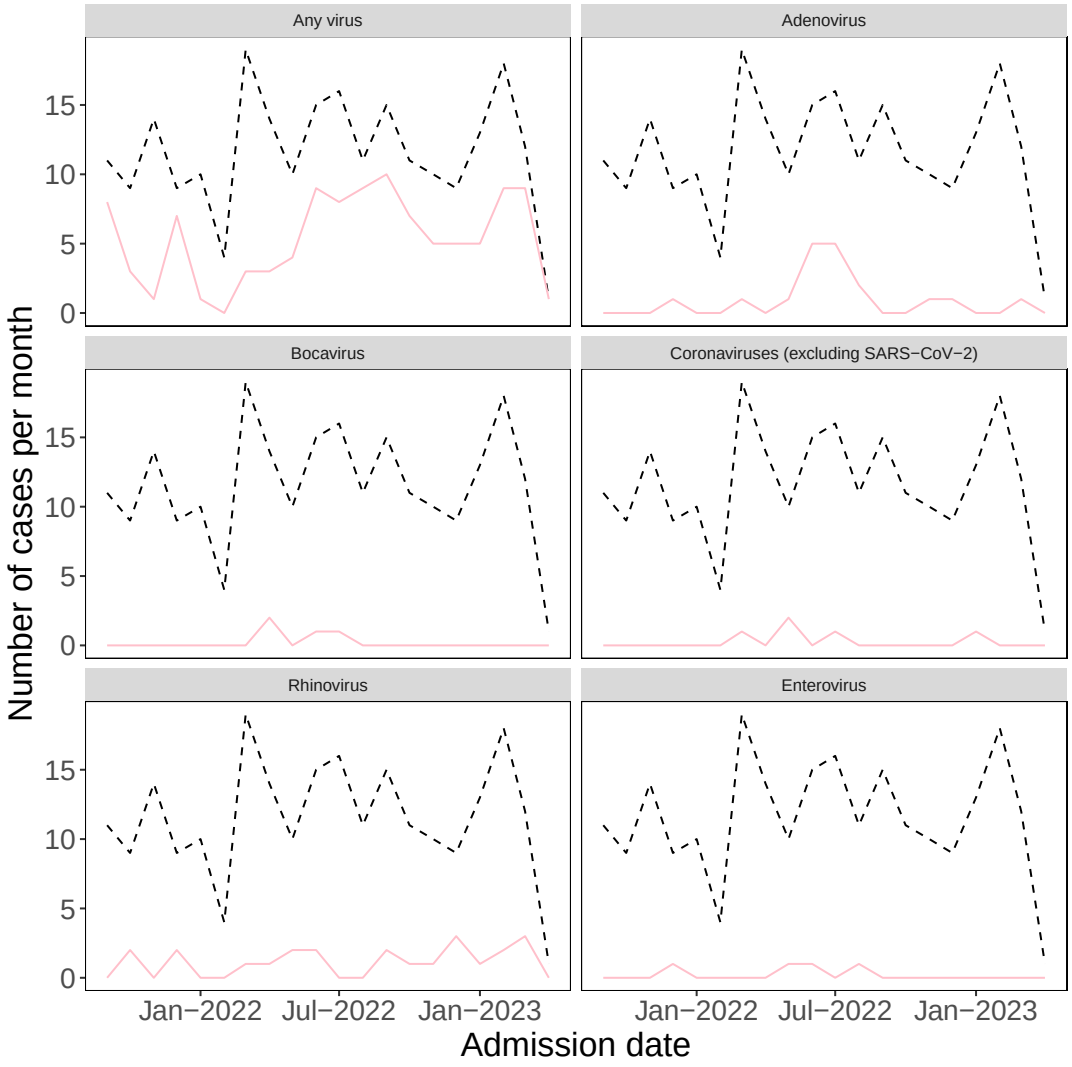


Figure 6.9: Number of lower respiratory tract infections admitted per month is shown by the dotted line in each graph. The number of each virus detected per month in nasopharyngeal samples from study participants is shown by the red solid line. The solid red line in the top-left panel shows the number of LRTI cases where any virus has been detected. SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

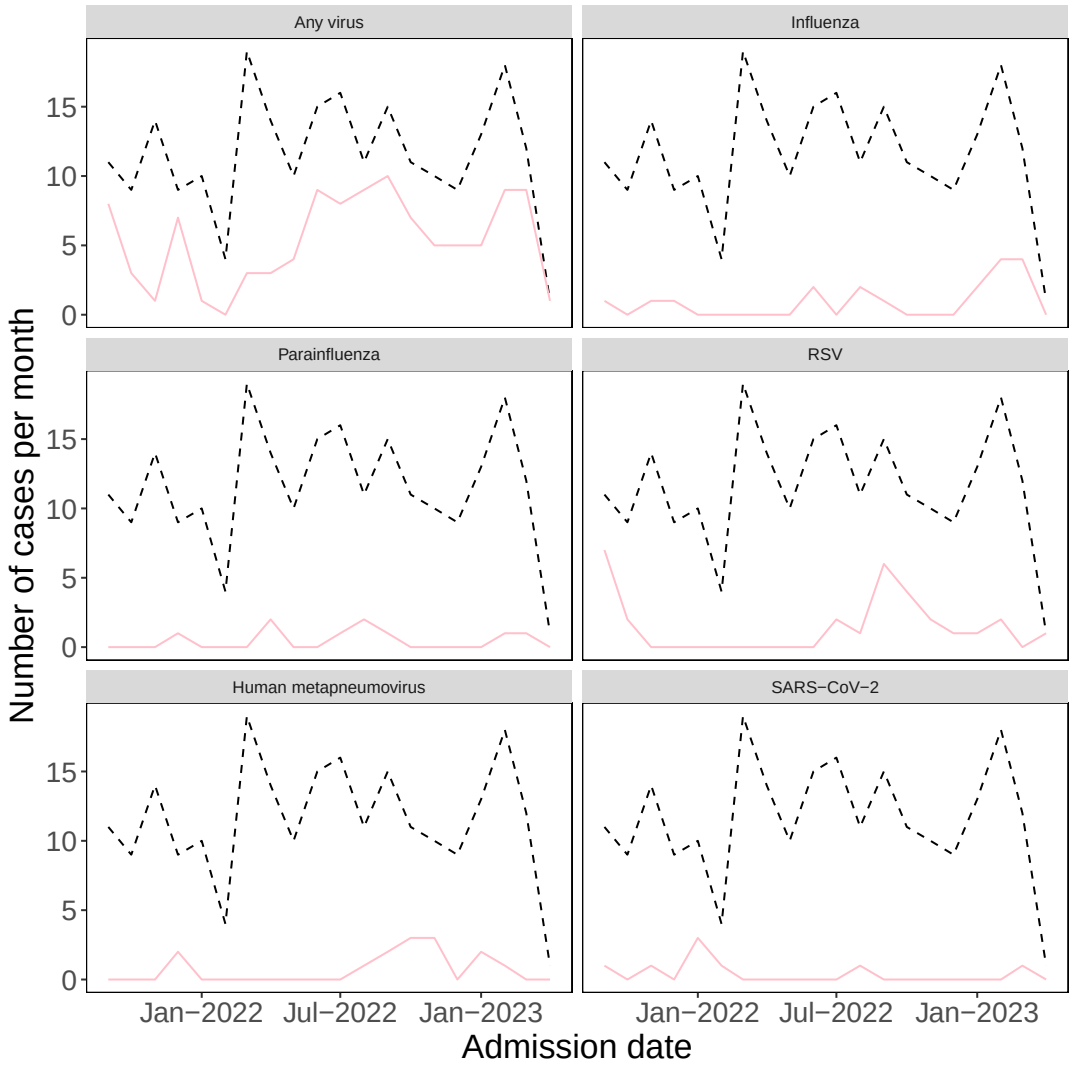


Figure 6.10: Number of lower respiratory tract infections admitted per month is shown by the dotted line in each graph. The number of each virus detected per month is shown by the red solid line. The solid red line in the top-left panel shows the number of LRTI cases where any virus has been detected. Viruses considered to be associated with LRTIs for classification of cases are included. RSV, respiratory syncytial virus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Blood molecular targets The in-house Micropathology Ltd. blood molecular panel was used to test 325 participant samples. Only four targets on this panel were used to re-classify cases, *S. pneumoniae*, *N. meningitidis*, *E. coli* and enterovirus. The full list of positive results is shown in Table 6.8. *S. aureus* was the most commonly detected bacteria, 4.31% (14/325) of samples. Human herpesvirus 7, 48% (156/325), Epstein–Barr virus (EBV), 45.2% (147/325), and human herpesvirus 6B, 33.2% (108/325), were each detected in at least a third of the samples tested.

The Siemens Tropical Fever Core Panel was used to test 82 samples. Dengue PCR was detected in 27 samples, this is discussed in relation to other dengue diagnostics below. *Rickettsial spp.* was positive in one sample. All samples were negative for chikungunya, West Nile virus, *Salmonella spp.*, *Plasmodium spp.* and *Leptospira spp.*

6.3.6 Potential management changes using molecular diagnostics

From the original classification using routine diagnostics, 100% (39/39) of the definite bacterial cases were treated with antibiotics, and 16.3% (17/104) of the definite viral cases were given antibiotics.

Of the 498 cases that underwent additional molecular testing, 84 cases had their initial classification changed from one of the probable/unknown groups to a definite bacterial/viral group. Three additional cases which were previously classified as definite viral, had a bacterial pathogen detected.

Table 6.8: Results from the Micropathology blood molecular panel. All positive results are shown, along with a list of pathogens with no positive results. **Enterobacter spp.* were detected by *Enterobacteriaceae* 16S PCR (n=6), two of which were also positive on *rpoB* gene analysis. †Staphylococcal testing, 12 samples positive on PAN16s rRNA, two additional samples were positive using targeted PCR for S. aureus.

n=325	Number of PCR positive cases	Percentage positive
Bacterial targets		
Staphylococcus aureus†	14	4.31%
*Enterobacter spp.**	6	1.85%
Acinetobacter baumannii	4	1.23%
Klebsiella pneumoniae	3	0.92%
Escherichia coli	3	0.92%
Streptococcus pneumoniae	3	0.92%
Neisseria meningiditis	2	0.62%
Pseudomonas aeruginosa	2	0.62%
Enterococcus faecalis	1	0.31%
Enterococcus faecium	1	0.31%
Viral targets		
Human herpesvirus 7	156	48%
Epstein–Barr virus	147	45.23%
Human herpesvirus 6B	108	33.23%
Adenovirus	7	2.15%
Enterovirus	7	2.15%
Cytomegalovirus	3	0.92%
Parechovirus	3	0.92%
Fungal targets		
Aspergillus spp.	5	1.54%
Candida spp.	3	0.92%

Pathogen targets with no positive result: *Haemophilus influenzae*, *Kingella kingae,* Group A *Streptococcus,* Group B *Streptococcus, Enterobacter cloacae, Serratia marcescens,* human herpesvirus 6A, parvovirus

Of the 80 cases re-classified as definite viral following additional molecular testing, 82.5% (66/80) of these cases were given antibiotics. Most of these cases were children with respiratory symptoms and detection of a pathogenic virus in their nasopharyngeal sample. A proportion of these cases potentially did not require antibiotics.

Two cases, who had enterovirus identified on blood PCR, were given 21 days of antibiotics to treat bacterial meningitis. Enterovirus was the likely cause of the meningitis symptoms, and whether the PCR result would have changed clinically management would have depended on how confident the clinician was that there was not a bacterial co-infection.

Of the seven cases confirmed as definite bacterial by molecular testing, four cases were given at least seven days of antibiotics. Two cases were treated with a third-generation cephalosporin, one of these cases was also treated with amikacin and acyclovir. The bacterial pathogen identified, antibiotics given, and duration are shown in Table 6.9. Three cases were re-classified as definite bacterial but were not given antibiotics, and were discharged well from hospital. These three cases would have likely been given antibiotics if the PCR results were available, but given that each participant also have a positive PCR result for dengue fever, the significance of the bacterial PCR results is unclear.

6.3.7 Co-detection of pathogens

There were nine cases where a bacterial and a viral pathogen were both detected, and both pathogens could have accounted for some of the features of the clinical

Table 6.9: Cases re-classified as definite bacterial following additional molecular testing. The pathogen(s) identified in each case, the duration and type of antimicrobial given are listed, along with the diagnosis and outcome at discharge are presented. PCR, polymerase chain reaction.

Pathogen identified by additional molecular testing	**Antibiotic given**	**Duration of antimicrobial treatment**	**Discharge diagnosis**	**Outcome**
N. meningitidis	Third generation cephalosporin, amikacin, plus aciclovir	21	Meningitis	Discharged alive, without sequelae
N. meningitidis	Ampicillin	7	Viral induced wheeze with occult bacteraemia	Discharged alive, without sequelae
E. coli plus human metapneumovirus PCR positive	Ampicillin	8	Pneumonia	Discharged alive, without sequelae
S. pneumoniae	Third generation cephalosporin, cloxacillin	19	Pneumonia with pleural effusion	Discharged alive, without sequelae
Rickettsia spp. plus dengue (PCR)	None documented	0	Dengue fever without warning signs, asthma exacerbation	Discharged alive, without sequelae
E. coli plus dengue (PCR)	None documented	0	Dengue fever without warning signs	Discharged alive, without sequelae
E. coli plus dengue (PCR)	None documented	0	Dengue fever with warning signs, non-specific abdominal pain, severe vomiting	Discharged alive, without sequelae

presentation; these cases were classified as bacterial-viral co-infections, see Table 6.10.

There were 14 cases where two or more potentially pathogenic viruses were detected, see Table 6.10. The importance of potential co-infections is discussed in Section 6.4.2.

6.3.8 Dengue cohort

One hundred and one participants tested positive for dengue virus using PCR, NS1 antigen, dengue-specific IgM or IgG. One participant had their dengue testing performed outside of Patan Hospital and was excluded from the analyses concerning different dengue testing below.

The proportion of participants who underwent dengue testing depended on the time of study enrolment, with most participants during the dengue outbreak having dengue testing performed. Research PCR testing was carried out for all participants with a sample available and enrolled during the dengue outbreak period, 1st August to 15th December, 2022.

Diagnostic tests compared

As part of routine care, NS1 testing was carried out on 100% (100/100) of samples, and 85% (85/100) of tests were positive. As part of the additional molecular panels, PCR testing was carried out on 64% (64/100) of samples in the dengue cohort, and 42.2% (27/64) of samples were PCR-positive.

Of the 64 participants that had results for PCR and NS1 testing, one case was

Table 6.10: Cases where two potentially pathogenic pathogens were detected. The bacterial pathogen is listed first in the bacterial-viral co-detections. In the viral-viral co-detections, the pathogens are listed in random order. No cases had three or more potentially pathogenic pathogens detected. HMPV, human metapneumovirus; SARS-CoV-2, Severe acute respiratory syndrome coronavirus 2; RSV, respiratory syncytial virus.

Pathogen one	Pathogen two	Number of cases
Bacterial-viral co-detection		
Escherichia coli	Dengue	2
Escherichia coli	Parainfluenza	2
Escherichia coli	HMPV	2
Escherichia coli	SARS-CoV-2	1
Rickettsia spp.	Dengue	1
Orientia tsutsugamushi	Dengue	1
Viral-viral co-detection		
Dengue	RSV	3
Influenza	HMPV	2
Dengue	SARS-CoV-2	2
Parainfluenza	HMPV	1
RSV	HMPV	1
RSV	SARS-CoV-2	1
Parainfluenza	SARS-CoV-2	1
Influenza	SARS-CoV-2	1
Dengue	Influenza	1
Influenza	Parainfluenza	1

PCR positive and NS1 negative, whereas 32 cases were NS1 positive and PCR negative. Taking PCR cases to be true positive, gives a sensitivity of 96.3% (95% CI 81% - 99.9%) with specificity of 13.5% (95% CI 4.5% - 28.8%) for NS1 testing. In Section 6.4.3, I discuss whether it is appropriate to classify NS1-positive and PCR-negative cases as false positives.

Timing of dengue testing relative to duration of illness

Twenty-six participants had results for all four dengue tests, assuming PCR to be the gold standard test, positivity rates for the other tests compared with PCR were 96.2% (25/26) for NS1, 11.1% (3/26) for IgM and 3.8% (1/26) for IgG.

The duration of dengue fever will alter the positivity rates of the different tests. Figure 6.11 shows how the positivity rates of the different tests varied with duration of symptoms. In the first two days of illness, NS1 and PCR have similar positivity rates, however, from day three onwards the proportion of PCR-positive samples decreases. NS1 samples remain positive into the second week of illness. The number of participants with test results from the second week of illness was small, 6% (6/100).

Dengue serotyping

The samples that were dengue virus positive on PCR, and had sufficient volume of extracted nucleic acid, were tested using the Dengue Differentiation panel for dengue serotype targets. Serotype one was the most commonly detected, 66.7% (12/18), followed by serotype three, 16.7%, (3/18) and serotype two 11.1% (2/18)

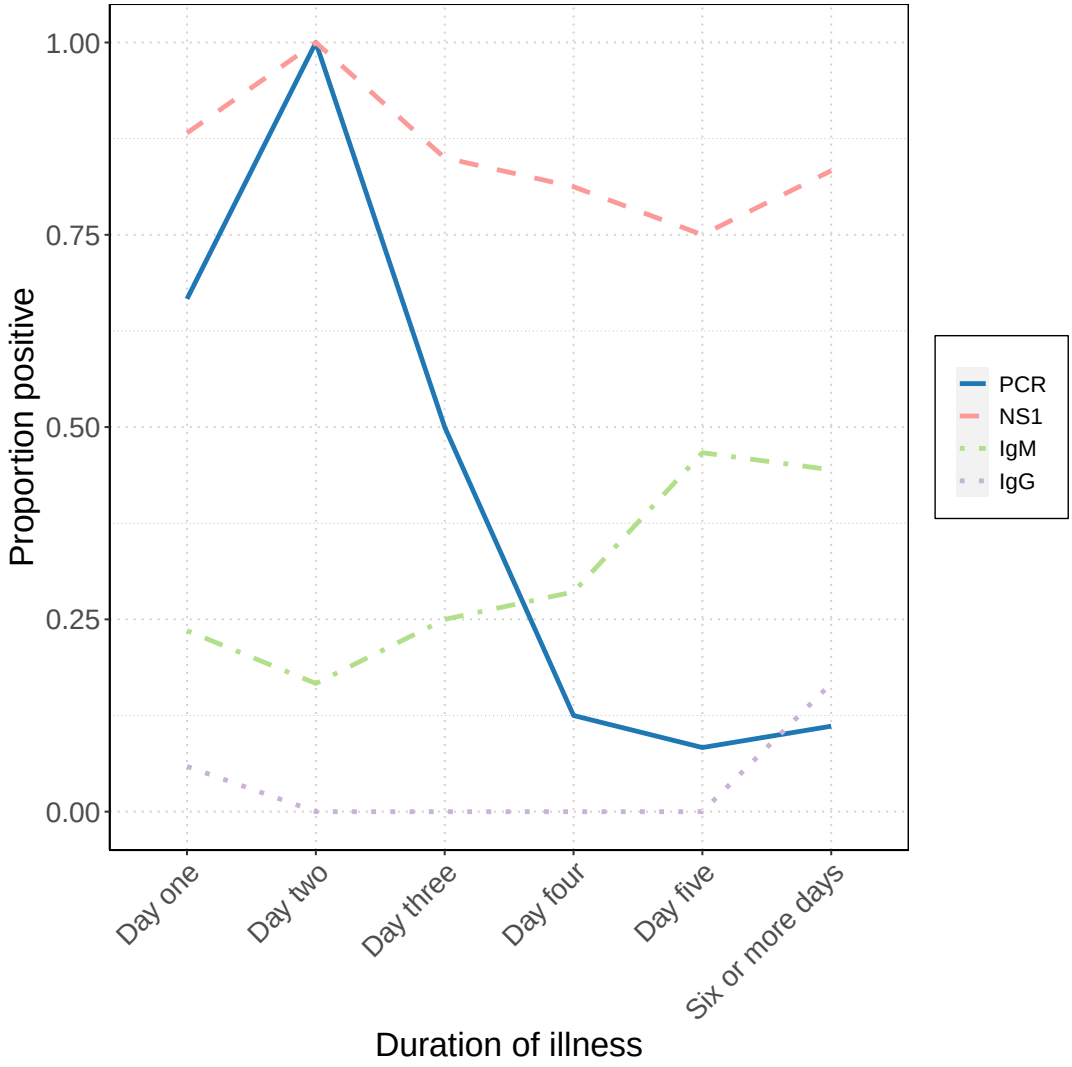


Figure 6.11: Proportion of positive test results for dengue virus among the participants with a diagnosis of dengue. The proportion of positive tests for dengue polymerase chain reaction (PCR), non-structural protein antigen (NS1), dengue-specific immunoglobulin M (IgM) and dengue-specific immunoglobulin G (IgG) are presented. The x-axis represents the number of days of illness when samples were taken for testing.

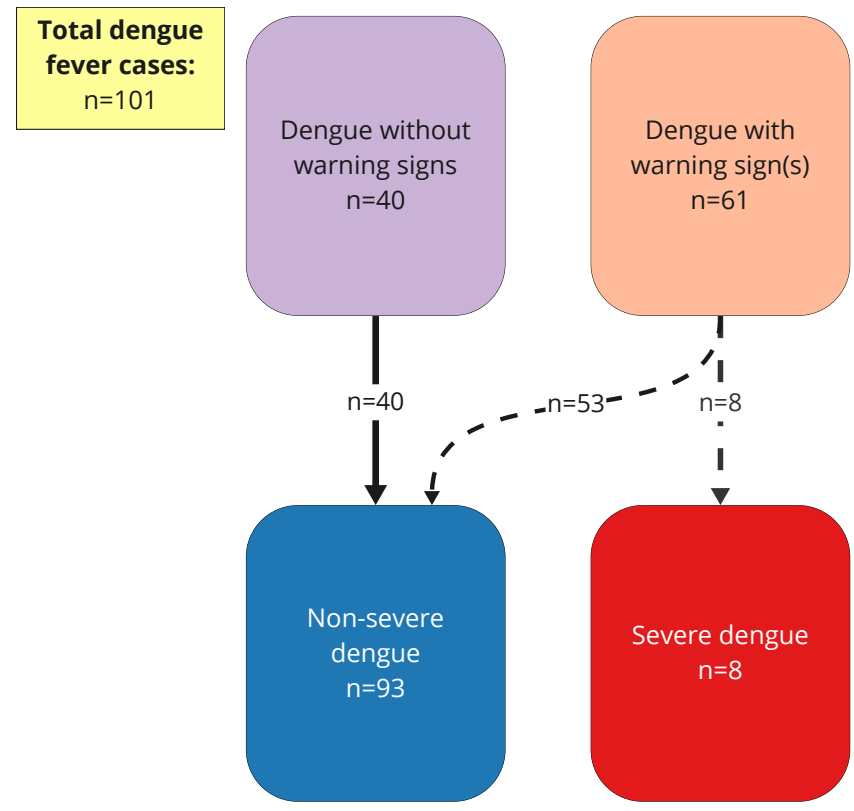


Figure 6.12: Dengue fever cases enrolled in the fever cohort study, divided by warning signs at admission and by severity at discharge.

of samples tested. Serotype four was not detected, and one sample which was positive for dengue PCR was negative for all four serotypes.

Severe dengue fever

Of the 101 dengue fever cases, 39.6% (40/101) of cases presented without warning signs and 59.4% (61/101) of cases had warning signs for potentially severe disease on presentation to hospital. On discharge, 7.9% (8/101) of cases were classified as severe dengue, with five cases requiring ICU admission. Figure 6.12 shows that all of the cases of severe dengue presented with warning signs.

Table 6.11: Number and type of samples collected at each study time point in the fever cohort, and samples collected from healthy controls. PAXgene tubes are used for RNA-stabilisation tube. EDTA, ethylenediaminetetraacetic acid.

	Time point 1	**Time point 2**	**Time point 3**	**Controls**
PAXgene	574	17	154	113
EDTA	351	1	113	12
Serum	182	1	62	7
Nasopharyngeal sample	391	NA	NA	12

6.3.9 Lower respiratory tract infection diagnoses

LRTIs make up 40.2% (231/574) of the overall cohort. Table F.3 shows the number of LRTI cases grouped by diagnostic classification, with demographics, duration of hospital stay and blood laboratory values at admission. These cases can be used as a validation cohort for the results presented in previous chapters. Of the LRTI cases, 13.4% (31/231) had a poor outcome. For demographics of the LRTI group and measure of severity, see Appendix, Section F.2.2.

6.3.10 Samples collected at each time point

For future analysis, it is important to understand how many, and what types of samples, are available. Table 6.11 shows that 26.8% (154/574) of participants had at least one research sample taken at time point 3, and the number of control participants is 19.7% (113/574) of the number of cases.

Molecular results from convalescent samples

Samples from time point 3, convalescence, were also sent for molecular panel testing using the Micropathology in-house blood panel. Results from paired samples, at time point 1 and time point 3, for the same participant were available in 53 cases. Table 6.12 shows the proportion of positive samples at each time point, in paired results only. Very few samples were positive for any bacterial or fungal targets.

The proportion of Epstein–Barr virus positive samples was the same at time point 1 and time point 3, with 56.6% (30/53) positivity. No cases of glandular fever were diagnosed clinically.

Human herpesvirus 7 had similar positivity rates at time point 1 and time point 3, 52.8% (28/53) and 45.3% (24/53) respectively. Human herpesvirus 6B also had similar positivity rates at time point 1 and time point 3, 28.3% (15/53) and 24.5% (13/53) respectively. There were low positivity rates for other viral targets.

Of the molecular targets used to re-classify cases into their diagnostic groups, enterovirus was the only positive target at time point 3. This positive result came from a participant who was also enterovirus positive at time point 1. This participant was diagnosed with meningitis, and the positive time point 3 sample was taken 14 weeks after initial presentation to hospital.

Table 6.12: Cases with paired results using the Micropathology Ltd. blood panel at time point 1 and time point 3. The percentage of positive cases at each time point is shown, with the number of positive cases in brackets. The molecular targets with no positive results, at either time point, are shown in the box below the table.

n=53	Time point 1 - percentage positive	Time point 3 - percentage positive
Bacterial targets		
Pseudomonas aeruginosa	1.9% (1/53)	3.8% (2/53)
Escherichia coli	1.9% (1/53)	0%
Staphylococcus genus	0%	1.9% (1/53)
Enterobacter spp.	0%	1.9% (1/53)
Viral targets		
Epstein–Barr virus	56.6% (30/53)	56.6% (30/53)
Human herpesvirus 7	52.8% (28/53)	45.3% (24/53)
Human herpesvirus 6B	28.3% (15/53)	24.5% (13/53)
Enterovirus	1.9% (1/53)	1.9% (1/53)
Adenovirus	1.9% (1/53)	0%
Cytomegalovirus	1.9% (1/53)	0%
Fungal targets		
Candida spp.	0%	1.9% (1/53)

Pathogen targets with no positive result at either time point: *Neisseria meningiditis, Haemophilus influenzae*, *Streptococcus pneumoniae, Kingella kingae, Klebsiella pneumoniae,* Group A *Streptococcus,* Group B *Streptococcus, Enterococcus faecalis, Enterococcus faecium, Enterobacter cloacae, Aspergillus spp., Serratia marcescens, Acinetobacter baumannii,* human herpesvirus 6A, parvovirus, parechovirus

6.3.11 Description of samples sent for RNA sequencing

A subset of participants from the cohort were selected for an RNA-seq experiment, the results of which were not available at the time of writing. These samples are

described further in Appendix F.

6.4 Discussion

In this cohort study, 574 children with signs of infection were recruited. Using routine investigations, 28.7% of cases were classified as either definite viral or definite bacterial infections. Samples from this cohort underwent extensive additional molecular testing. Following the addition of these molecular panel results, 45.6% of cases were classified as definite viral or definite bacterial infections.

Molecular panels allow for the detection of multiple pathogen targets with one test. The benefits of using molecular panels include that smaller sample volumes are required to test multiple targets, reduced laboratory time required and less cost than performing multiple single target tests.[74] The set of targets in a molecular panel needs to be relevant to the population where the test is being carried out.

The most commonly identified pathogens were dengue virus and respiratory viruses such as RSV, influenza and SARS-CoV-2. *E. coli* was the most commonly detected bacterial pathogen, mainly in urine culture. Smaller numbers of other important bacterial pathogens were identified including *S. pneumoniae*, *S.* Paratyphi A, *S.* Typhi, *N. meningitidis* and *M. tuberculosis*.

A large number of dengue cases were identified, mainly using NS1 antigen testing as part of routine investigations. Additional PCR testing was positive early in dengue fever disease, but PCR positivity dropped quickly after the first three

days of symptoms.

6.4.1 Causes of paediatric infections at Patan Hospital, Nepal

A comprehensive effort was made to describe the causes of fever at Patan Hospital, Nepal, using routine diagnostics and additional molecular testing.

The most common causes of infection varied throughout the enrolment period. This cohort was recruited during the latter part of the SARS-CoV-2 pandemic. An outbreak of dengue fever also occurred during the enrolment period. For these reasons, higher numbers of cases were classed as definite viral than would have been expected during a different period.

A large number of dengue fever cases were enrolled during the outbreak in the latter part of 2022. Most of these cases were classified as non-severe, and this was to be expected as the last large dengue outbreak in the Kathmandu Valley was in 2019, and this may have been the first dengue infection for several of these children.[35]

Additional molecular testing identified several respiratory viruses linked to LRTI cases. This was the first description of the causes of viral LRTIs in this cohort since the COVID-19 pandemic. RSV and influenza remain the most common pathogens. It is important to have data on rates of RSV infection in Nepal, especially as new RSV vaccines become available in the coming years. High rates of RSV detection among children admitted with LRTIs will add data when deciding on the cost-effectiveness of RSV vaccines.

As expected, few cases were culture-positive for a bacterial pathogen. Enteric fever pathogens were the most commonly identified bacteria in blood cultures which highlights that this is still an important disease in Nepal. The introduction of a typhoid conjugate vaccine to Nepal's routine immunisation programme in 2022 will hopefully reduce the burden of enteric fever in Nepali children.[324]

The next most common bacterial pathogen detected in blood was *S. pneumoniae*. A PCV10 vaccine was introduced to the Nepali immunisation schedule during 2015.[325] An effectiveness study, following PCV10 introduction, reported that vaccine serotype carriage was reduced in children younger than 2 years of age with pneumonia but also found increased carriage of non-vaccine serotypes.[143] Higher valent vaccines may need to be considered if non-vaccine serotypes are becoming more important in children in Nepal, and continued surveillance of pneumococcal serotypes is needed.

Describing the changing causes of infections in Nepali children is important when deciding on future public health interventions.

The percentage of children with a poor outcome was 8%, including those who required PICU admission or had a prolonged hospital stay. Female participants made up 35% of the overall cohort, however, 50% of those with a poor outcome were female. Differences in outcomes based on sex have been reported previously. In a Spanish study of 2,609 patients admitted to PICU, males were more likely to be admitted to PICU, 57.5% male compared with 42.5% female, however, females had worse outcomes after PICU admission, with an OR of 1.55 of in-hospital death for females compared with males.[326]

6.4.2 Utility of molecular testing

During the COVID-19 pandemic, the capability to perform PCR testing became available at Patan Hospital. So far this has mainly been used for SARS-CoV-2 testing. In this chapter, I have described how we have successfully run molecular panels at Patan Hospital.

Whether expanded molecular testing becomes available clinically at Patan Hospital will depend on financial constraints and how clinically useful these tests could be. The cost of PCR testing has been the main barrier to implementation in low-resource settings. The Nepali government designated Patan Hospital a site for SARS-CoV-2 testing and this allowed the infrastructure to be put in place. However, ongoing costs are involved with expanding the molecular testing programme including staff costs, PCR testing kits and laboratory consumables.

If the hospital authorities decide to expand the molecular testing programme at Patan Hospital, it would be important to know which tests are most useful clinically. As reported in this chapter, molecular testing could have a big impact on the number of cases with a confirmed diagnosis.

Most of these additional diagnoses came from a small number of molecular targets. When testing respiratory samples, RSV and influenza were present in significant numbers of LRTI cases. In blood samples, the number of diagnoses identified using molecular testing was much lower but the value of not missing bacterial infections means these positive results are important.

Enterovirus and *N. meningitidis* were identified in the blood in a small number of meningitis cases. Had these tests been available to the clinical team these

results may have been important in guiding management. In these meningitis cases, it would be more useful to look for a pathogen in the CSF. A molecular testing panel to test CSF for pathogens including enterovirus and *N. meningitidis* would be important in clinical decision-making.

Care must be taken in interpreting molecular test results, 72% of the participants had a positive result for at least one molecular target. Several of the viral targets had very high positivity levels in the blood, including EBV and human herpes virus 6B and 7. The clinical utility of these results is limited, especially considering that similar positivity rates were observed in the convalescent samples, 6-12 weeks after the acute illness.

The use of many of the bacterial targets for assigning diagnosis is not supported by good evidence. Most of the bacterial targets were not used to re-classify cases in this cohort. A small number of bacterial targets, with evidence to support their use, were used to re-classify cases in this study. Even when using these bacterial targets, the full clinical picture should be taken into account.

Most clinicians would agree that *N. meningitidis* is the likely pathogen when detected in the blood of a child with a presentation consistent with bacterial meningitis. However, other cases where blood PCR was used to re-classify cases were less straightforward. Two participants presented with a febrile illness and were positive for dengue on NS1 antigen testing, but also positive for *E. coli* on the blood PCR panel. These participants were diagnosed with dengue fever and were treated with supportive care and discharged. The management choices made may have been different if the clinicians treating these patients had access to the molecular results, but the management changes may not have been beneficial to

the participants.

The number of pathogen targets searched for increases the chance of detecting multiple potential pathogens in the same sample. The detection of two potential pathogens does not mean a co-infection is present, and using routinely used diagnostics it can be impossible to confidently assign diagnoses of co-infections.

Several LRTIs were assigned as a definite viral infection in this cohort based on the detection of a virus in their nasopharyngeal sample. While this is useful in assigning cases in biomarker studies, the detection of a viral pathogen does not rule out a bacterial co-infection. This means that even if these molecular results were available to the clinical team, the proportion of these cases receiving antibiotics may still be high.

6.4.3 Dengue diagnostics

During the dengue outbreak, NS1, IgM and IgG testing was carried out as part of routine investigations; dengue PCR testing was part of the additional molecular testing panels. Dengue PCR did not seem to add much in terms of cases missed with routine diagnostics, only one additional case was identified by PCR.

The duration of symptoms was important for the PCR positivity rates. In the first two days of symptoms, PCR and NS1 positivity rates were similar, however from day three onwards PCR positivity rates fell sharply, while NS1 positivity rates remained over 75%, even if measured in the second week of symptoms. NS1 tests remaining positive longer into illness than PCR has been reported previously, however, the drop-off in PCR positivity was earlier than expected.

PCR positivity was reported to peak between day two and day five of illness.[314] The earlier decline in PCR positivity reported here limits the utility of dengue PCR in this setting.

While NS1 seems to be a more useful test in this setting, the continued positivity of NS1 into the second week of illness raises questions about how long NS1 tests remain positive after acute infection. In one participant, NS1 was positive 15 days after the onset of symptoms. Persisting NS1 positivity could lead to an incorrect diagnosis if a child presents with a different illness following a recent dengue infection.

PCR testing for dengue serotypes provides information about the serotypes which were present during the outbreak. This is useful public health information and would be more useful if obtained in real-time. The team at the Molecular Laboratory at Patan Hospital aim to use the dengue serotype PCR kits to test samples during future outbreaks.

6.4.4 Limitations

Due to limitations in the samples obtained, additional molecular testing was not possible for several participants. However, most participants had at least some additional molecular testing, and even with this molecular testing more than half the cohort were still without a definite pathogenic diagnosis. Cases without a confirmed diagnosis are of less utility in future biomarker analyses using these samples.

Classification of cases into their diagnostic groups depends on gold standard

testing. Culture-based methods have low sensitivity, so we had to use other methods to classify cases. Detection of a virus in the nasopharynx is suggestive of the cause of LRTI but we cannot be sure that the virus detected is the pathogen. We are possibly misclassifying some of these cases. This highlights the need for improved diagnostics in febrile illness.

As with the LRTI cohort described in previous chapters, the relatively low number of definite bacterial cases identified will complicate the picture when trying to look for biomarkers of bacterial infection. Biomarkers which can confidently identify bacterial infections are important clinically, and low numbers of bacterial cases are not ideal for the discovery or validation of bacterial biomarkers.

6.4.5 Future work

This cohort study is still recruiting and the samples from this study will be used for several different analyses. Samples have already been sent for RNA-seq. Analysis of these results will include looking for new biomarkers related to dengue and severity of disease.

I want to repeat the RNA and protein work described in earlier chapters. Samples from the cohort in this chapter can be used to validate the results described in the earlier RNA and protein chapters of this thesis.

The DIAMONDS consortium has begun pilot studies of new host immune-based, point-of-care tests based on the biomarker discovery within the consortium. The Nepal site will be part of this pilot study and recruitment of cases to this cohort will continue with the aim of testing new point-of-care diagnostic tests.

6.4.6 Conclusions

A small number of molecular tests were identified which could be useful to diagnose common and important causes of infection. The efforts to describe the causes of fever at Patan Hospital help to make the samples collected more useful for biomarker studies using RNA and protein platforms.

Chapter 7

Discussion

In this thesis, I have shown that blood samples using RNA and protein platforms can be used to distinguish between different causes of LRTIs in children. I presented novel RNA and protein signatures to differentiate between bacterial and viral LRTIs. I have then integrated data across three platforms to identify a multi-platform signature which can recognise bacterial infections with high sensitivity.

I also present the most comprehensive description to date of the causes of fever in children at Patan Hospital. I have identified a small number of molecular tests which could provide useful diagnostic information in this setting.

In the following sections I address how the results presented in this thesis address each of the three original hypotheses proposed in Section 1.12.

7.1 Differentiating between bacterial and viral lower respiratory tract infections

The first hypothesis tested in this thesis was that differences in whole blood RNA expression and plasma protein abundance can be used to differentiate between different causes of LRTIs in children.

With each of the three platforms used to differentiate between bacterial and viral LRTIs, signatures of varying accuracy were developed. The main measure of interest in creating these signatures was the sensitivity for detecting bacterial LRTIs, as detecting bacterial infections is important clinically. The top-performing single-platform signatures were identified using the RNA data.

Even though the protein signatures, from the MS proteomics data and the MSD cytokine panel, had lower sensitivity for detecting bacterial infections these signatures could still be useful clinically; especially as it would be easier to measure a small number of proteins than to develop a multi-platform test. Two RNA signatures were presented which could identify bacterial infections with reasonably high sensitivity, 89% in the test dataset, and specificity of 100%. The question is are these results good enough to consider bringing these RNA signatures forward for further evaluation? To answer this, I compare these signatures to results from studies with similar aims.

Two large cohort studies identified different RNA signatures to distinguish between bacterial and viral infections. The RNA signatures presented in the Mahajan *et al.* and the Herberg *et al.* papers are high-sensitivity signatures discovered in different high-income settings. Both studies reported over 90% sensitivity for detecting bacterial infections in children.[100, 103] The transcripts identified in these published RNA signatures were different to the ones included in the signatures in this thesis. This may be due to the setting of the studies. The pathogen profile was different, cases of enteric fever, for example, would be rare in European or American settings but more common in this Nepali cohort. Genetic factors can also have an impact on RNA gene expression.[184, 183]

Any new diagnostic test would ideally be applicable across different settings. It is important to discover and validate new diagnostic tests in LMICs, like Nepal. Previous studies have shown the potential utility of RNA signatures in low-resource settings.[104, 327]

While the RNA signatures were the best performing individual platform-based

tests, the most accurate signature overall was the multi-platform signature incorporating genes and proteins from both the MS proteomic results and the cytokine panel. It would be more complex to take this kind of multi-platform test forward for development as a useful diagnostic test, but the performance of the multi-platform signature is good enough to warrant further study. An expert consensus paper defined the aims for any new diagnostic test to differentiate between bacterial and viral infections to be a sensitivity of over 90% with a specificity of at least 80%.[85] The multi-target signature meets those minimum thresholds with a sensitivity of 100% and specificity of 94%.

In this thesis, I have focused on signatures which can differentiate between bacterial and viral infections as this is an important clinical question to answer. However, in certain situations, it may be more useful to classify cases into a broader range of possible outcomes, including inflammatory syndromes, which can mimic severe bacterial infections, and non-bacterial, non-viral infections such as malaria. Habgood-Coote *et al.* recruited children with different infectious and inflammatory presentations to a large cohort study and used a machine learning approach to identify a 161-transcript signature which can differentiate between 18 different disease conditions including different bacteria, common viruses, inflammatory conditions such as Kawasaki disease and malaria.[328] While this more complicated gene signature would be more useful in cases with complex presentations, there is still a place for a highly sensitive test to identify bacterial infections.

The ideal scenario is that these RNA signatures could be converted to an affordable, easy-to-use, point-of-care diagnostic test. If these new diagnostic tests

are developed and validated exclusively in high-income populations, then their utility in other settings may be limited. In low-resource settings, a point-of-care diagnostic test to identify bacterial infections has the potential to have more of an impact than in similar high-resource settings. This is because improved diagnostic tests mean that more limited healthcare resources can be targeted to those who need them most.

Developing new signatures in this Nepali setting is also a potential limitation of this study. Since this was a single-site cohort study the signatures produced in this setting may not be applicable to other cohorts. The follow-up study described in Chapter 6, will allow validation of these results in another cohort of Nepali children. And as discussed above it is important to discover and validate new diagnostics in a variety of different settings, including LMICs.

7.2 Pathophysiology of lower respiratory tract infections in children

The second hypothesis to be tested in this thesis was that, in children with different causes of LRTIs, whole blood RNA expression and plasma protein abundance levels can be used to improve our understanding of the pathophysiology of disease.

Different pathways of interest were over-enriched when GSEA was carried out on RNA-seq results comparing gene expression across LRTI groups. For example, GSEA was performed comparing the acute pneumococcal group results with acute

samples from other bacterial LRTIs; the list of significantly different pathways showed that IL-1, IL-6, IL-8 and NF-kappaB were enriched in the pneumococcal group compared with other bacterial LRTIs. A small number of genes that were common to all these pathways were also elevated in the pneumococcal group in our data; genes which encode for the proteins TLR6 and LILRA2 had higher gene expression in the pneumococcal group compared with other bacterial samples. The potential mechanisms as to why these genes might be increased in pneumococcal infection were discussed in Section 2.4, TLR6 and LILRA2 may have important roles in pneumococcal infections.[186, 187, 189] TLR6 activates NF-kappaB signalling and NF-kappaB regulates many aspects of the innate and adaptive immune system and is a key mediator of the inflammatory response.[329] TLR6 levels were highest in the pneumococcal group but TLR6 levels were higher across the entire acute bacterial group when compared with gene expression in the acute viral group. The levels of TLR6 expression in the viral group were similar to levels in the convalescent samples. Another protein which regulates NF-kappaB was included in the multi-platform signature to differentiate between bacterial and viral infections; clusterin down-regulates NF-kappaB and was found to be more abundant in the viral LRTI group compared with the bacterial group.[294] Decreased levels of clusterin have been reported in patients with bacterial sepsis.[293] These results combined would suggest that regulation of NF-kappaB is different in bacterial and viral infections, linked to changes in levels of TLR6 and clusterin. These proteins may be useful as biomarkers, or potential therapeutic targets to alter the inflammatory response in LRTIs.

Another example of a finding which may improve our understanding of the

pathophysiology of LRTIs in children comes from the MSD cytokine panel study. In Section 4.3.5, I reported that three cytokines were increased in acute bacterial samples compared with samples from viral LRTIs, IL-6, G-CSF and IL-15. Previous work in a mouse model reported that IL-15 was important in the defence against *Salmonella* Typhimurium.[272] With *Salmonella* Typhi being a relatively common bacterial pathogen in this setting, this may help explain the higher IL-15 levels in the bacterial group. I discuss IL-15 further, including the previous report suggesting a link between IL-15 and bronchiolitis severity in Section 4.4.

7.3 Describing the causes of fever using routine testing and additional molecular diagnostics

The third hypothesis investigated in this thesis was that the addition of molecular testing to routine investigations would provide useful diagnostic information when describing the causes of fever in Patan Hospital. This hypothesis was tested in the second cohort study, presented in Chapter 6, which enrolled children with all signs of infection. Using routine investigations, 28.7% of cases were given a definite classification, bacterial or viral. Following the use of a select number of molecular results, 45.6% of cases were classified as definite bacterial or definite viral. This is a substantial rise in the number of cases that potentially could have had their diagnosis altered if these molecular test results were available to clinicians. However, the utility of these extra molecular testing depends on whether these additional molecular classifications had the potential to improve

patient management.

Many of these additional molecular diagnoses were from respiratory viruses detected in the nasopharynx. As discussed in Section 1.4, the detection of a pathogenic virus in the nasopharynx is often the likely cause of LRTIs.[17, 54] However, children can carry viruses in their upper airways without symptoms, and the detection of a virus does not rule out a bacterial co-infection.[17] Knowing that a virus is present in the nasopharynx may not change clinician decisions around management and antibiotic use.

Detecting which viruses are associated with acute respiratory illnesses in Nepali children would also provide useful public health information. Tracking the epidemiology of respiratory viruses in Nepal would provide important information about outbreaks and seasonality of respiratory infections. Also, information regarding viral LRTI aetiology will be important when making national decisions regarding vaccines, especially as new RSV vaccines become available.[30]

The additional blood molecular panel results identified a small number of important bacterial infections. *N. meningitidis* was identified in two meningitis cases. Identifying the cause of meningitis is important in guiding management in seriously ill children. Three cases had a positive blood molecular test for *S. pneumoniae*. Knowledge of the pathogens causing pneumonia can help to guide management. Two of these cases were also culture positive for *S. pneumoniae* in the blood, but molecular testing could have provided results in a shorter time frame.

The number of additional cases identified from the blood molecular panels was much lower than the respiratory panel. Despite this, the blood molecular panel

may be a more important test to clinicians, as identifying bacterial infections is more important clinically than assigning a pathogenic cause to LRTIs.

The results presented in Chapter 6 identified a small group of molecular targets that would have the most impact on clinical management of cases, including targets for serious bacterial infections like *S. pneumoniae* and *N. meningitidis*. A respiratory panel including targets for RSV and influenza would also provide useful information for public health officials and clinicians, with the caveat that detection of a respiratory virus does not rule out bacterial co-infection.

7.4 Future directions

The classification signatures discovered in this thesis would need be validated in separate cohorts. At time of writing, samples from the fever cohort study described in Chapter 6 have been sent for RNA-seq. This will allow for validation of the RNA-seq results from Chapter 2, in a new cohort of Nepali children. The fever cohort has recruited participants with a wider range of infections, so validation of the identified bacterial-viral differentiation signatures can be carried out across a broader range of infections.

As well as being a validation cohort for the RNA-seq results presented in this thesis, the follow-up fever cohort can act as an interesting discovery cohort. The large number of dengue cases recruited to this cohort could allow for discovery of useful signatures for dengue fever.

In Chapter 6, a small number of molecular tests were identified as potentially useful clinical investigations. The setting for this cohort study, Patan Hospital,

acquired molecular testing capabilities during the COVID-19 pandemic. The data presented in this thesis highlight tests which would be useful to include in a focused panel to diagnose infectious diseases in children in the Kathmandu Valley. These data may be used by the paediatric clinical team at Patan Hospital when discussing resources for diagnostic testing with hospital authorities.

RNA or protein based diagnostic signatures may become part of routine clinical practice in the future. Thinking about how to use these tests, and where they would be most useful is important. A test that can detect bacterial infections with a high sensitivity would be useful on presentation to hospital, as in the cohorts presented in this thesis. However, an easy-to-use, point-of-care test to identify bacterial infections would also have utility in pre-hospital settings. A test which could differentiate between many different infections may be better placed as a diagnostic tool in children who are more unwell and requiring admission to hospital.[328]

After deciding on where to use a new test, it is also important that the test is acceptable to the people using it, both patients and medical professionals. The cost involved would be important in any setting where healthcare resources are limited, in settings like Nepal the patients or their families contribute to the cost of tests performed. Any new test would need to be affordable to the population. Medical professionals then need to have confidence in the test and the results need to be easy to interpret. The barriers to implementation are not insurmountable, and hopefully new diagnostics to classify infectious diseases will become available in clinical practice in the near future.

7.5 Concluding remarks

There is room for improvement in assigning aetiology in paediatric infections. The different host immune signatures presented in this thesis, and the clinically useful molecular tests, could improve our ability to confidently assign the cause of infection. This is important when making clinical decisions regarding how to manage sick children.

Clinicians use several different parameters including history-taking, epidemiological data and clinical examination, along-side diagnostic investigations to make a diagnosis. The different diagnostic avenues explored in this thesis have the potential to work together in clinical practice. An RNA or protein signature may be able to identify bacterial infections, and molecular testing could be used to identify the specific bacteria causing disease. Diagnosing the cause of illness in children is often about putting all the different pieces of the puzzle together. I hope that the work presented in this thesis, and the post-doctoral work required to validate these findings, will make it a little easier for clinicians to fit the pieces of the diagnostic jigsaw together.

Appendix A

Appendix - Systematic Review

To help guide the work in this thesis, I systematically reviewed evidence related to novel biomarkers in paediatric viral LRTIs. Due to the large number of papers found, a subset of results related to severity of RSV infections was analysed and presented here. Biomarkers in severity were an initial interest of my work, however, the data available in the LRTI cohort study described in Chapter 2 did not allow for an adequate assessment of severity. Work to discover biomarkers of severity in the fever cohort described in Chapter 6 will take place when RNA-seq and protein platform results are available for this cohort.

I used the results of the systematic review to help decide which cytokines to include in the cytokine panel in Chapter 4, and these results have also informed the evidence presented in Chapter 1 and Chapter 4.4. This review will be included in a separate publication to this thesis. The methods are briefly outlined here.

A.0.1 Defining the research question

With the help of the librarian at the Department of Paediatrics, University of Oxford, the research question for the review was defined as: Which host genes/genetic factors or host biomarkers predict the aetiology and severity of viral pneumonia?

Genes and genetic factors were included as another colleague was interested in reviewing this area, and we combined the two topics to allow us to work together.

The search terms included related to five areas:

1. Disease – Terms related to LRTIs/pneumonia/bronchiolitis.
2. Viruses – viruses known to be associated with respiratory infections
3. Biomarker list – a list of potential biomarkers, including broad terms such as cytokines and antibodies and specific names of known biomarkers.
4. Diagnosis/severity terms including hospitalisation/intensive care
5. Platform – proteomics/transcriptomics/RNA-seq/metabolomics etc.

A.0.2 Databases searches

Ovid Medline, Ovid Embase, Web of Science, Scopus and Cochrane Library databases were searched. The initial search was performed on 24th July 2022, and repeated on 7th December 2022. The result will be updated again before the future publication of this work.

A.0.3 Deduplication and screening

Results were deduplicated using Microsoft Endnote. The de-duplicated results were uploaded to the DistillerSR platform and a deduplication check was performed again using the deduplication algorithm on DistillerSR.

Initially, the titles were screened, followed by screening of abstracts. Abstracts were streamed into different groups depending on the study type. Studies involving human subjects were kept separate from studies related to animals and *in-vitro* studies. Studies were also streamed if the studies were related to a particular pathogen, RSV, influenza, parainfluenza, human metapneumovirus or other viruses. Studies which compared different pathogens, or measured severity were also grouped together. I was assisted in this process by colleagues at the Oxford Vaccine Group. Two researchers reviewed each article before it was excluded. Conflicts were decided by review by a third member of the team. The full-texts of the included studies then reviewed, focusing on the biomarker results presented in each study.

The key results from each study were recorded in a table. The key points recorded from each study included the population of the study, which biomarkers were measured, what platform was used to measure the biomarkers, the results from each biomarker, whether a control group was used and whether the results were significant or not in the study. Results were plotted to summarise the changes in cytokine levels across different levels of clinical severity of RSV infection.

RSV infections were classified as mild if they could be managed outside of hospital. RSV infections were classified as moderate if they required hospital

admission. RSV infections were classified as severe if they required invasive ventilation or intensive care admission.

A.0.4 Results

Following deduplication 6,297 references were included for title screening. Figure A.1 shows the number of studies screened and included. These results relate to primary research which enrolled human children with RSV infection, measured biomarkers and included a severity measure. A total of 100 studies were included for full-text review. Included in the 100 studies were 43 studies which measured cytokines, 29 studies which measured RNA transcripts and 22 studies related to different white blood cell markers.

As we are interested in cytokine studies to help inform which biomarkers to include in the MSD cytokine panel in Chapter 4, results related to cytokine studies are summarised here.

Figure A.2 shows the most commonly measured cytokines across the studies. IFN-gamma and IL-10 were the most commonly measured cytokines. Table A.1 summarises the 24 studies which measured IFN-gamma and the results related to the severity of RSV disease. Thirteen of the IFN-gamma studies reported statistically significant results. Figure A.3 shows IFN-gamma levels across different severity grades of RSV infection. Generally, in severe clinical RSV infections levels of IFN-gamma are lower than in moderate severity RSV infections.

Figure A.4 shows whether the IFN-gamma results were associated with more or

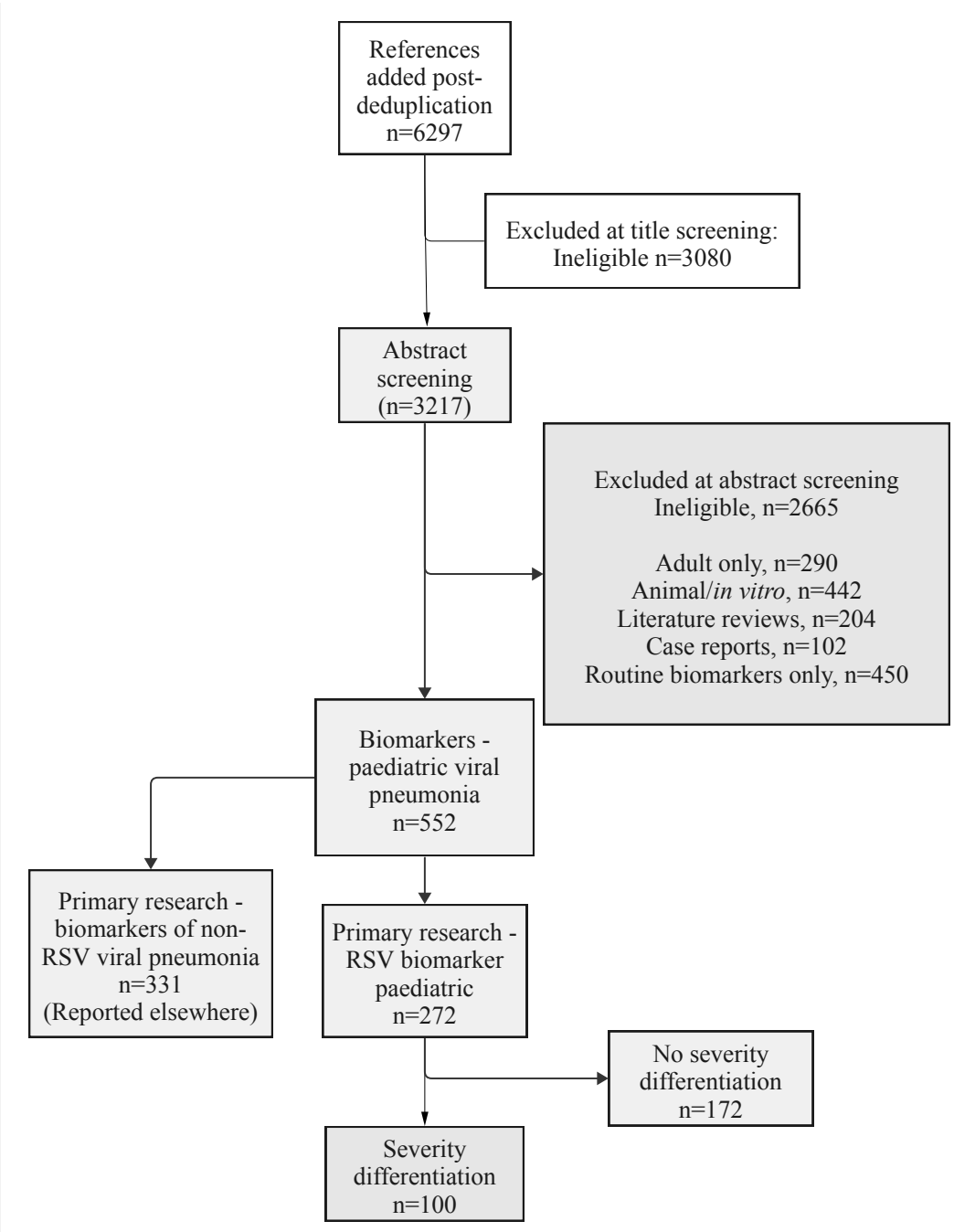


Figure A.1: Consort diagram showing the number of articles post-deduplication and the number of articles included after each screening step. Articles which reported results of primary research involving paediatric participants with RSV infections, and where the authors had reported differences in biomarker levels related to severity of clinical disease were included.

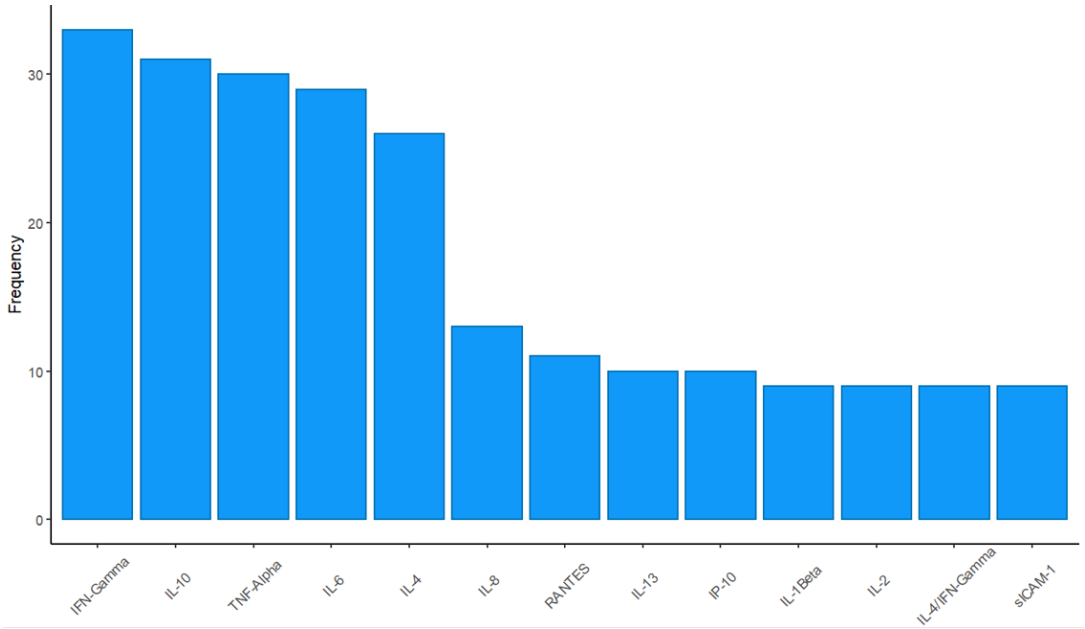


Figure A.2: Each bar represents the number of studies which measured a specific cytokine of the studies included for full-text review in the systematic review. Articles which reported results of primary research involving paediatric participants with RSV infections, and where the authors had reported differences in biomarker levels related to severity of clinical disease were included.

less severe disease. Across the included studies, as the severity of RSV disease increased, IFN-gamma decreased.

IL-10 was the second most-measured cytokine. Table A.2 shows the studies where IL-10 was measured and the main result when IL-10 was measured across different grades of severe RSV disease. Figure A.5 shows that IL-10 levels were highest in moderate disease when upper respiratory tract samples were measured, and IL-10 levels were lower in both mind and severe RSV disease. One of the included studies measured IL-10 in blood and reported that IL-10 levels decreased as the severity of RSV disease increased.

There was heterogeneity in the age and setting of enrolment of the population,

Table A.1: Summarising the studies which measured IFN-gamma in children with RSV LRTIs, and a severity measure was included in the study.

Study	Number of RSV Participants	Control Group	Severity Measure	Sample Type	Key Result	Statistically Significant
Alonso Fernández, 2008	196	No	Supplemental O2 required	Plasma	No difference - mild versus moderate disease	No
Bacharier, 2013	123	No	Severity score, lowest SpO2, duration of hospital stay	PBMCs	No difference - moderate versus severe disease	No
Bennett, 2007	63	No	Duration of supplemental O2	Nasal wash	Negative correlation - IFN-gamma lower in participants with longer duration of supplementary oxygen	Yes
Bont, 1999	50	Yes	Moderate diseae = Hospitalised, Severe diseaes = Mechanical ventilation	Plasma, supernatants of PHA stimulated blood cultures	IFN-gamma higher in moderate versus severe disease	Yes
Brandenburg, 2000	95	No	Supplemental O2 required, ventilation support	Plasma	No difference - mild versus severe disease	No
Caballero, 2015	418	No	SpO2 at enrolment	Nasopharyngeal aspirate	IFN-gamma higher in mild versus severe disease	Yes
Cepika, 2008	12	Yes	Days of wheezing, duration of supplemental O2, lowest O2 saturation, maximal RR	Blood	No correlation between IFN-gamma and disease severity	No
Chen, 2002	29	Yes	Supplemental O2 required	Serum	No difference - moderate versus severe disease	No
Diaz, 2012	49	No	Severity score	Plasma	No difference - moderate versus severe disease	No
Ferolla, 2018	51	No	ICU admission	Nasopharyngeal aspirate	IFN-gamma higher in moderate versus severe disease	Yes
Garcia, 2012	19	No	Duration of hospital stay, duration of supplemental oxygen	Nasal wash	Negative correlation - IFN-gamma lower in participants with longer duration of supplementary oxygen or longer duration of hospital stay	Yes
Garofalo, 2001	86	Yes	Mild = URTI, Moderate = Non-hypoxic, Severe = Hypoxia (SpO2 <95%)	Nasopharyngeal aspirate	IFN-gamma highest in moderate disease (versus mild or severe disease). No correlation between duration of supplementary O2 and IFN-gamma level	Yes
Hassan, 2008	58	Yes	Moderate = Not meeting severe criteria, Severe = High PCO2 (>50mmHg) or ICU Admission or Mechanical ventilation or SpO2 <85% during admission	Serum	IFN-gamma higher in moderate versus severe disease	Yes
Lee, 2007	30	No	Severity score	Blood/PBMCs	No difference - mild versus moderate versus severe disease	No
Legg, 2003	28	No	Mild = URTI, Moderate = bronchiolitis	Nasal wash	IFN-gamma higher in mild versus moderate disease	Yes
Mella, 2013	57	No	Mild = Non-hospitalised, Moderate = Hospitalised	Nasopharyngeal aspirate	IFN-gamma higher in mild versus moderate disease	Yes
Moreno-Solís-2, 2015	45	Yes	Severity score or supplementary O2 required	Nasopharyngeal aspirate, serum	No difference - mild versus moderate disease	No
Piedra, 2017	79	No	SpO2 at enrolment	Nasopharyngeal aspirate	Positive correlation - IFN-gamma higher when SpO2 at enrolment is higher (higher IGN-gamma in milder disease)	Yes
Pinto, 2006	42	Yes	Mild = Non-hospitalised, Moderate = Hospitalised	Blood/PBMCs	IFN-gamma higher in mild versus moderate disease	Yes
Siefker, 2020	54	No	Moderate = Hospitalised, Severe = Admitted to ICU	Nasopharyngeal aspirate	No difference - moderate versus severe disease	No
Tabarani, 2013	478	No	Mild = Non-hospitalised, Moderate = Hospitalised, Severe = ICU	Nasophayngeal wash	Detection level of IFN-gamma too low for comparison	No
Taveras, 2019	84	Yes	Mild = Outpatient, Moderate = Hospitalised	Nasophayngeal wash	No difference - mild versus moderate disease	No
Taveras, 2020	219	No	Mild = Outpatient, Moderate = Hospitalised, Severe = ICU	Midterbinate swab	IFN-gamma highest in mild disease versus moderate or severe disease	Yes
Twaites, 2018	30	No	Moderate = Hospitalised, Severe = Mechanical ventilation	Nasal sampling (nasorption)	IFN-gamma higher in moderate versus severe disease	Yes

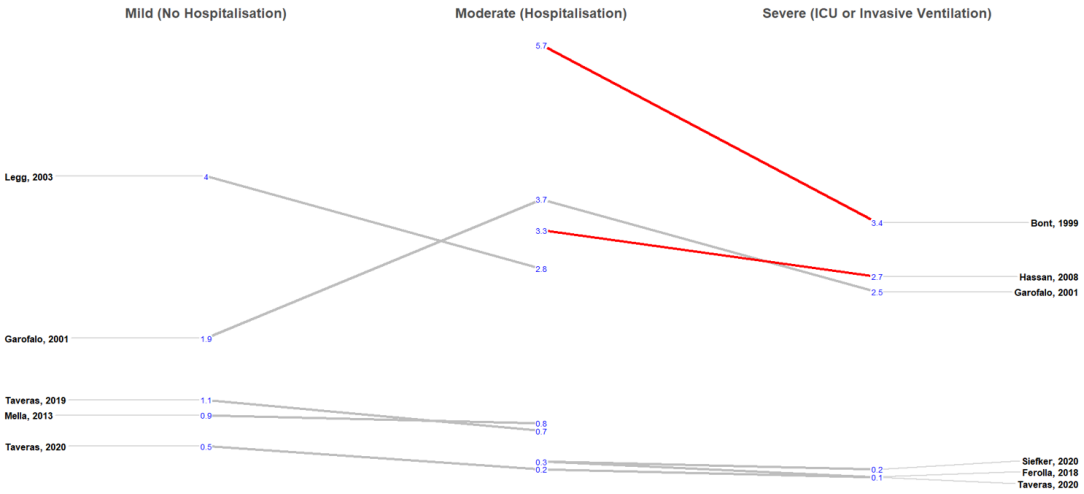


Figure A.3: IFN-gamma levels in studies of children with RSV infections. Mild disease is classified as when the infection can be managed at home. Moderate disease is classified as requiring hospital admission. Severe disease is classified as requiring invasive ventilation or intensive care admission. Articles which reported results of primary research involving paediatric participants with RSV infections, and where the authors had reported differences in biomarker levels related to severity of clinical disease were included.

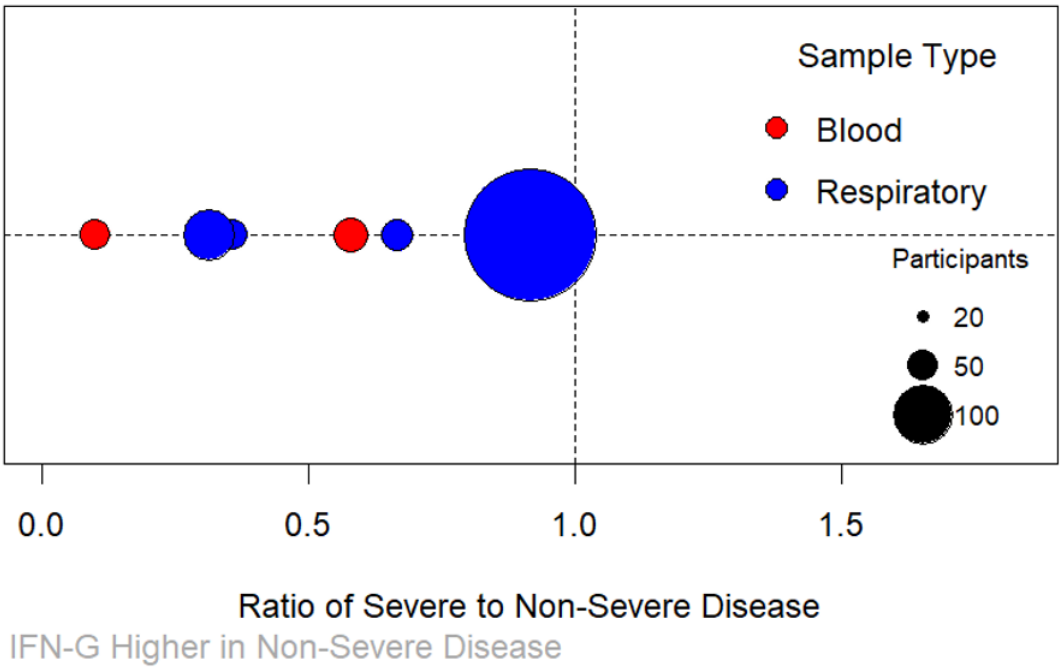


Figure A.4: IFN-gamma levels in included studies of children with RSV infections. The levels of IFN-gamma levels in severe disease were compared with levels of IFN-gamma in non-severe disease (mild or moderate classifications). A ratio under one indicates that IFN-gamma is higher in non-severe than severe disease. The grey lines represent studies where cytokine levels were measured in samples from the upper respiratory tract. The red lines represent studies where cytokine levels were measured in blood samples. Articles which reported results of primary research involving paediatric participants with RSV infections, and where the authors had reported differences in biomarker levels related to severity of clinical disease were included.

Table A.2: A table summarising the studies which measured IL-10 in children with RSV LRTIs, and a severity measure was included in the study. The key finding related to this severity measure is also reported here.

Study	Number of RSV participants	Non-RSV control participants	Severity Measure	Sample type	Key Result regarding IL-10 levels	Statistically Significant
Bennett, 2007	62	No	Duration supplementary oxygen		Negative correlation between increasing IL-10 levels and duration of oxygen symptoms, -0.28	Yes
Brand, 2013	50	No	Treatment level*	Nasopharyngeal_aspirates	Plasma levels of IL-10 highest in the group not requiring hospital admission. IL-10 levels from nasopharyngeal aspirates highest in the group requiring hospital admission and lower in the group requiring ICU.	No
Brandenburg, 2000	95	No	Severe RSV if one of these criteria: pCO2> 6,6kPa, O2sat <90% or artificial ventilation	Plasma	No difference in mild compared with severe RSV disease	No
Diaz, 2012	49	Yes	Severity score	Plasma	No difference in severity scores in mild compared with severe RSV disease	No
Fan, 2018	89	Yes	Duration of symptoms	Serum	Severity scores increased in mild RSV disease compared with moderate/severe disease	Yes
Ferolla, 2018	51	No	ICU Admission	Nasopharyngeal_aspirates	No difference when comparing moderate (hospital admission) with severe disease (ICU) admission	No
Garcia, 2012	19	No	Severity score/length of stay	Nasal_wash	No significant correlation between IL-10 levels and severity score/length of hospital stay	No
Legg, 2003	28	No	URTI is classified as mild compared with bronchiolitis which is classified as moderate	Nasal_wash	No difference between levels in mild and moderate RSV disease	No
Pinto, 2006	42	No	Hospitalisation compared with non-hospitalised	PBMCs	No difference in levels between hospitalised and non-hospitalised RSV cases	No
Sarkar, 2016	56	No	Severity score	Nasopharyngeal Aspirate	Levels increased in severe compared with moderate RSV disease	Not recorded
Schuurhof, 2012	25	No	Treatment level required - mild, treated at home; moderate, required hospitalisation; severe, required ventilation support	Nasopharyngeal Aspirate	Levels highest in moderate disease (requiring hospitalisation), levels lower in the mild group (treated at home) and the severe group (requiring intensive care)	Yes
Tabarani, 2013	478	No	Treatment level required - mild, treated at home; moderate, required hospitalisation; severe, required ICU	Nasophayngeal_Wash	IL=10 levels undetectable in most samples	No
Twaites, 2018	30	No	Hospitalised is classed as moderate and requiring mechanical ventilation is classed as severe	Nasorption	Levels increased in moderate compared with severe disease	No
Moreno-Solís-2, 2015	45	Yes	Severity score, supplemental oxygen required	NPA	No difference in levels between mild and moderate disease	No
Alonso Fernández, 2005	196	No	Supplemental oxygen required	Plasma	Levels increased in moderate compared with mild disease	Yes
Alonso Fernández, 2009	196	No	Duration of supplemental oxygen	Plasma	Positive correlation between duration of oxygen therapy and IL-10 levels	Yes
Mella, 2013	57	No	Non-hospitalised is classed mild compared with hospitalised which is classed as moderate	NPA	No difference in levels between mild and moderate disease	No
Viera, 2010	30	No	Severity score	NPA	Positive correlation between severity score and IL-10 levels	Yes

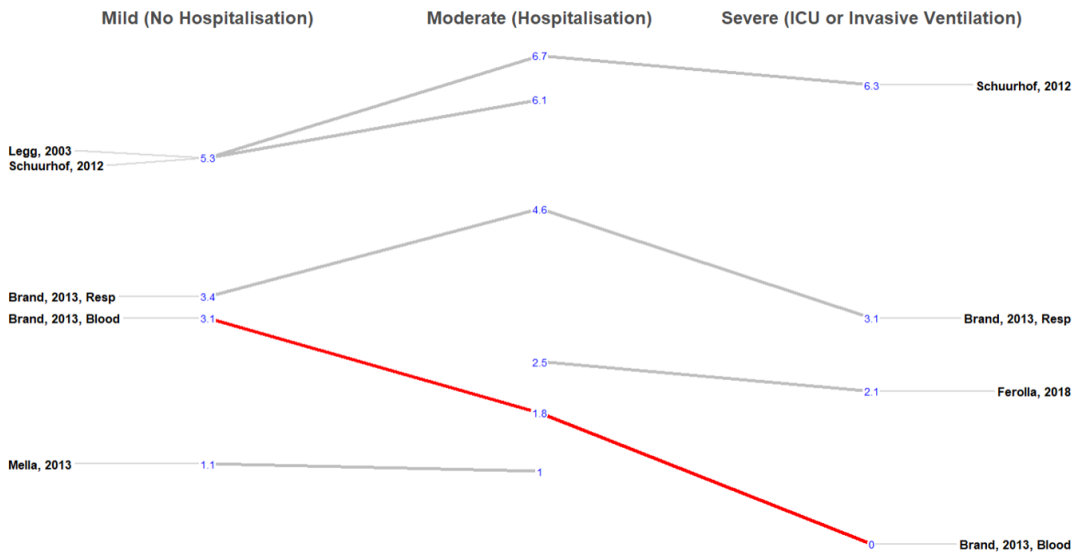


Figure A.5: IL-10 levels in studies of children with RSV infections. Mild disease is classified as when the infection can be managed at home. Moderate disease is classified as requiring hospital admission. Severe disease is classified as requiring invasive ventilation or intensive care admission. Articles which reported results of primary research involving paediatric participants with RSV infections, and where the authors had reported differences in biomarker levels related to severity of clinical disease were included.

and heterogeneity in the type of control group used and the platform used to measure the biomarkers. The heterogeneity of the study methods meant that it was not feasible to do a formal meta-analysis.

A.0.5 Discussion

As severity of LRTI infections was not a focus of this thesis, these results were not reported in the main body of the thesis.

This brief summary of the systematic review highlighted commonly measured cytokines in RSV infections. IFN-gamma levels were higher in less severe RSV infections compared with severe RSV infections requiring ICU or ventilation.

When IL-10 was measured in respiratory samples, levels were highest in moderate disease and fell in the severe group. Similarly, IFN-gamma levels were generally lower in severe RSV disease compared with moderate disease.

Lower cytokine levels in the most severe disease could be explained by a immunoparesis or an infection-induced immune suppression. This was suggested by Leahy *et al.* as a mechanism to explain a similar pattern of IL-15 cytokine levels in children with RSV infections. Leahy *et al.* suggested that immunoparesis explained why IL-15 levels were highest in moderately severe RSV infections and decreased in severe RSV bronchiolitis.[252] This infection-related immunosuppression is described in another virus, measles. A prolonged cytokine imbalance is one factor associated with this immunosuppression following acute measles infection.[330]

Other results from this systematic review will be published separately. A brief

summary of this review was included as the papers included in this review helped to inform decisions regarding the cytokines used in the MSD cytokine panel in Chapter 4.

Appendix B

Appendix - Linked to Chapter 2, Transcriptomics

B.1 Methods

B.1.1 Processing count data in R

Detailed methods used for processing the RNA-seq data from Chapter 2.

1. Load packages: readr, edgeR, limma, org.Hs.eg.db, gplots, ggplot2, RColorBrewer, NMF, dplyr, PCAtools, sva, variancePartition, BiocParallel, EnhancedVolcano, caret, mlbench, fgsea
2. Load count files and merge two count files from different experiment dates
3. Annotate the merged count file
4. Change GenBank accession numbers to Entrez IDs, using org.Hs.eg.db

Bioconductor package

5. Applying list of synonyms of gene names from NCBI-Gene database

6. Removed NA values from Entrez ID list – non-coding genes

   (a) Not found in NCBI Gene database, or org.Hs.eg.db

   (b) E.g. U47924.2, Z82206.1, RF02277, RF02209, RF00477, RF00001, AB015752.1, ABBA01006766.1, AC005230.1, AP001273.1, BX255925.3, CR382285.2, L34079.1

7. Parse the metadata

   (a) Add data points that are useful for analysis, e.g., combine all convalescent samples into one group for comparison with acute samples.

   (b) Merge the time point and diagnostic group columns.

   (c) Add batch column to represent which RNA-seq experiment each result belongs to.

8. Convert the data frame of raw counts to DGEgene list object. This is an object used by edgeR to store count data. It has a number of slots for storing various parameters about the data.

9. Various methods of filtering were trialled to exclude the majority of low-count genes. These methods included:

   (a) Using the filterByExpr package in edgeR – this package takes into account the average CPM and the smallest group sizes, 7 is the smalled group in the acute samples, bacterial-viral co-infections.

(b) Filtering cut-off of count >5 across 3 or more samples

(c) Filtering by CPM >1 across >/=20% of the samples (CPM>1 equates to mean count of approx. 20).[165]

10. Normalisation, to minimise systematic technical effects

(a) Use the calcNormFactors function to generate a set of normalisation factors between libraries

(b) Calculate the weighted trimmed mean of the log-expression ratios or M values

(c) One sample used as a reference, and the trimmed mean of M values (TMM) factors are calculated for each of the other samples

(d) TMM normalisation factors will be included in the model used for differential expression

(e) Use mean difference plots to show the bias in the data due to systematic bias before and after TMM normalisation

11. Quality control plots

(a) Bar plot of library sizes for each sample, in millions, to see if any big differences in library sizes across the samples.

(b) Box plots of log2 counts per million plotted to look for outlying samples when log transformed data.

(c) PC biplots

(d) Interactive MSD plots created using glimma package. Data explored for outliers and factors which may cause a bias, and need to be

adjusted for in the model used for differential expression analysis.

12. Batch correct. Using removeBatchEffect function in the limma package in R, corrected for the batch effect caused by the RNA-seq experiment occurring in two batches on different dates.

13. Repeated PC biplots post-batch correction.

    (a) Looking for outliers

    (b) Identify patterns in the data

    (c) Exclude outlier samples

14. Build design matrix to include covariates of interest, include primary classification group, and confounding factors (age and batch in this case) and using age as intercept

15. Voom transform the data.

    (a) Transform the count data in log CPM while taking into account the mean-variance relationship.[165] Takes into account the TMM normalisation factors previously calculated.

    (b) Use duplicateCorrelation function with sample ID as the block, as we have acute and convalescent samples from the same participant, but we don't have complete pairing of samples. This will give estimated weights for the regression fit.

    (c) Perform voom transformation a second time, this time incorporating the estimated correlation in the estimation of the weights.[331]

16. Fit a linear model to the voom transformed data to test for differential gene expression using limma, using lmFit.

17. Make the contrast matrix, use the makeContrasts function to make a contrast matrix of the comparisons of interest.

18. Apply the contrast matrix to the fit object to get the results of the comparison of interest

19. Use the eBayes function to perform empirical Bayes shrinkage on the variances, and estimate the moderated t-statistic and the associated p-values. p-values were adjusted using the BH method.

20. Extract a table of the differential expression results from the fit object.

21. Make volcano plot of comparisons of interest.

To identify bacterial and viral samples:

1. K-means clustering can identify unlabeled clusters using an unsupervised machine-learning approach.

    (a) Plot the within-cluster variance for different starting cluster numbers to define the ideal number of clusters

    (b) Use the k-means function in the stats package to plot each sample on a cluster plot

    (c) Add the cluster assignment for each sample to the metadata

    (d) Compare the cluster assignment with the previous LRTI classification label

(e) Use the labelled data to identify bacterial and viral clusters

(f) Identify probable bacterial cases that cluster within the bacterial clusters

(g) Use the definite bacterial, definite viral and probable bacterial cases which cluster with other definite bacterial samples for model training.

To identify genes for a signature to differentiate between bacterial and viral LRTIs:

2. Compare acute bacterial samples with convalescent results from the same participants.

3. Compare acute viral samples with convalescent results from the same participants.

4. Take the DEGs between bacterial acute and convalescent groups, and between viral acute and convalescent groups, combine these two lists of genes.

5. Use this subset of genes to look for which genes are differentially expressed between bacterial and viral genes.

Signature for bacterial compared with viral

1. Re-do the initial DE analysis steps:

(a) As co-infections contain bacterial element we want to identify these as bacterial so these were added to the bacterial cohort

   (b) Test for DGE between bacterial cohort (definite bacterial, probable bacterial and co-infections) and convalescent samples from these participants.

   (c) Keep the same list of DEGs between the definite viral group and convalescent samples from these participants

2. Combine the bacterial and viral DEGs and use this as the list of potential genes for the signature.

3. Split the data into train and test datasets with a 70:30 split

4. Re-do DE analysis steps as above, with the training set only, including only DEGs identified from the two comparisons above, acute bacterial samples compared with convalescent samples from the same individuals, and acute viral samples compared with convalescent samples from the same individuals.

Signature approach one:

1. The partial least squares approach was used to train the model, with 10-fold repeated cross-validation, repeated three times. Genes with an importance of <0.0015 were excluded.

2. Re-train with set of high importance genes.

   (a) Use plsr function

   (b) And use RMSE to identify ideal number of features

   (c) Use recursive feature elimination to select the most important genes, and rank using RMSE.

3. Look for highly correlated genes. Build correlation matrix and exclude the gene with lower importance when a gene pair has a correlation coefficients >0.9. Create final model with optimal number of genes.

4. Test the accuracy of the model in the training data

Signature approach two:

1. The partial least squares approach was used to train the model, with 10-fold repeated cross-validation, repeated three times. Genes with an importance of <0.0015 were excluded.

2. Forward selection using PLSR identified the optimal number of genes to take forward

3. Look for highly correlated genes. Build correlation matrix and exclude the gene with lower importance when a gene pair has a correlation coefficients >0.9.

4. Re-train with set of high importance, low-correlation genes

   (a) Use recursive feature selection to identify the optimal number of genes in the re-trained model

   (b) Create final model with optimal number of genes

5. Test the accuracy of the model in the training data

Signature approach three:

1. Build linear model with just the intercept, add additional genes using forward selection.

2. Limit the signature to 50 genes.

3. The the optimal number of genes/50 genes forward for re-training.

4. Re-train the model using PLS approach to rank genes in the signature in order of importance.

5. Select the optimal number of genes using their importance and loadings.

6. Look for highly correlated genes. Build correlation matrix and dxclude the gene with lower importance when a gene pair has a correlation coefficients >0.75, threshold of 0.75 used in this final model due to the larger number of genes included.

Testing the signatures:

1. Re-do DGE analysis for the test data, and test the accuracy of the model(s) in the test dataset.

Pathway analysis:

1. Re-do DGE analysis with the groups changed to reflect pathogen classification. The three main pathogen groups are the RSV, influenza and pneumococcal groups. The pneumococcal group is made up of samples with a definite and probable pneumococcal classification.

2. First compare acute and convalescent samples across the pathogen groups, then compare acute samples from each pathogen group with other viral/other bacterial samples.

3. Take the results of DGE analysis and use these for pathway analysis.

Using GSEA pathway analysis:

1. Rank the full gene list in order of log-2-fold-change for each comparison
   (a) Pneumococcal acute versus convalescent
   (b) RSV acute versus convalescent
   (c) Influenza acute versus convalescent
2. Perform GSEA analysis looking for over-enrichment of gene sets in two databases:
   (a) GO-BP MSigBP
   (b) Immunological MSigDB
   (c) Plot the gene sets with the highest and lowest NES.
   (d) Create box plots to visualise the key genes in any interesting gene sets.
3. Next, use the subset of DE genes from the acute versus convalescent comparisons for pneumococcal, RSV and influenza. Repeat steps 1 and 2.
4. The function used for GSEA performs the following steps:
   (a) Validates the input, checks for duplicate genes and sorts the gene list
   (b) Performs GSEA, using fgsea pachage the enrichment scores for each pathway are calculated using the GO file containing the gene set data, compared with the ranked log-2-fold-change results. Removes pathways with an adjusted p-value over 0.05.

(c) Filters out redundant pathways where two pathways are very similar.

(d) Enrichment classification, the pathways are classified as up- or down-regulated.

(e) A plot is created to visualise the top ten enriched pathways for both up- and down-regulated genes.

B.2 Results

Figures related to the processing of data using R.

B.2.1 Filtering

Figure B.1 shows the density of the log2CPM values for each sample, before Figure B.1a and after filtering using the filterByExpr function in R, Figure B.1b.

B.2.2 Quality control plots

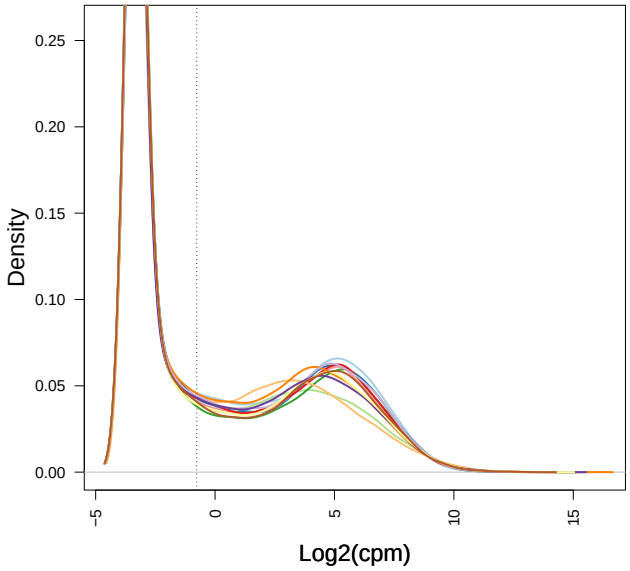
Figure B.2 shows a bar plot of the library size for each sample in millions.

The following figures show normalised counts per million. Figure B.3 shows box plots for all of the samples in the cohort pre-normalisation. Figure B.4 shows box plots for all of the samples in the cohort post-normalisation.

PCA was carried out and the following PC biplots were used for quality control.

Figure B.5 shows the first and second PC plotted against each other after batch correction, coloured by year of the RNA-seq experiment.

(a) Raw



(b) Filtered

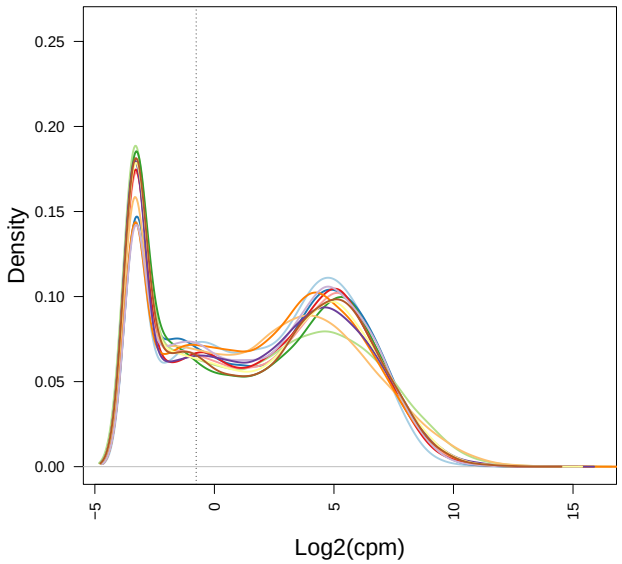


Figure B.1: Filtering using filteredByExpr function in EdgeR. Each line represents a sample. Panel (a) shows the raw density data for each sample in log2CPM. Panel (b) shows the density for each sample following filtering.

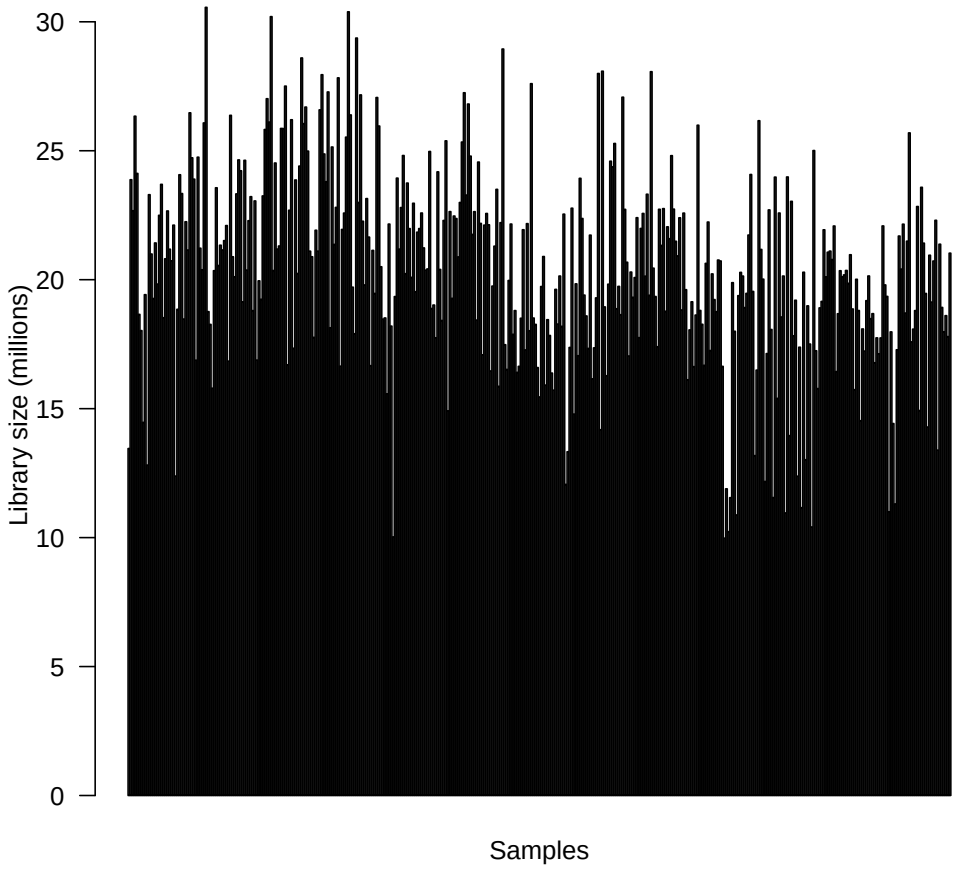


Figure B.2: Library size for each sample, expressed in millions of reads. Median library size for cohort, 20,348,397. CPM, counts per million

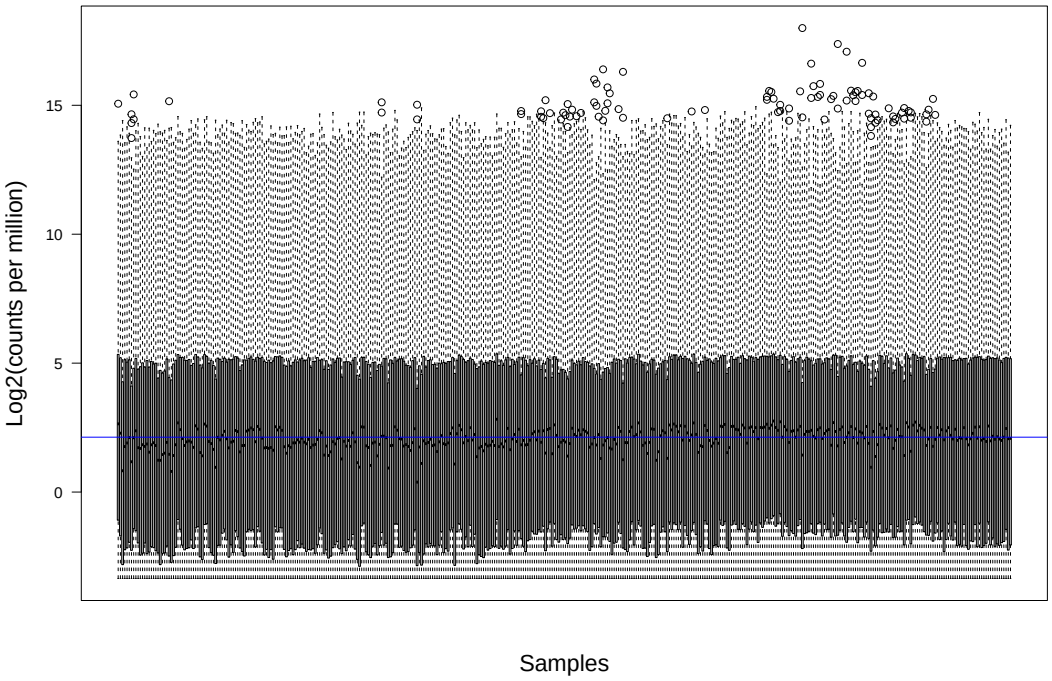


Figure B.3: Results from the RNA-seq experiment showing box plots of the log2-counts per million of each sample in the LRTIs cohort pre-normalisation.

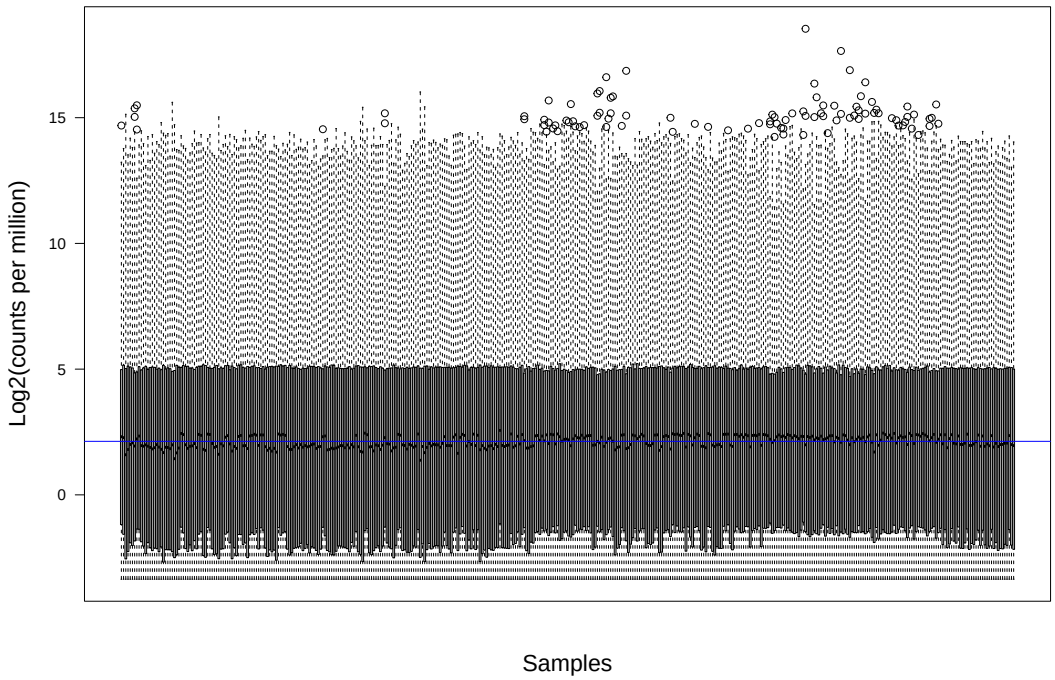


Figure B.4: Box plots of log2 counts per million for all samples included in the RNA-seq results, post-normalisation, blue horizontal line is median log(CPM) value.

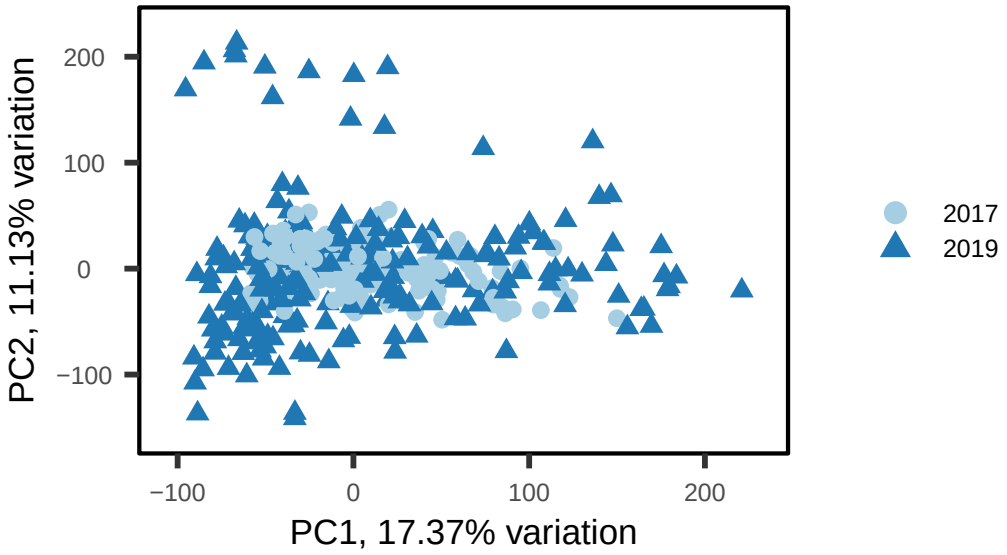


Figure B.5: PCA in the RNA-seq experiment. PC biplot, post batch correction using the removeBatchEffect function and coloured by the year of experimental batch.

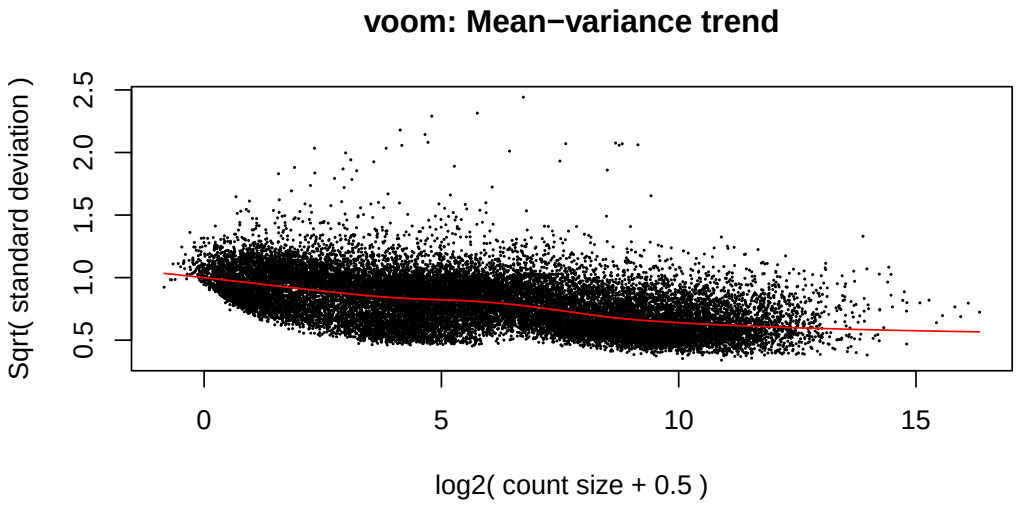


Figure B.6: In the RNA-seq experiment, the mean-variance plot post-voom transforming data is shown.

Figure B.6 shows the mean-variance trend post-voom transforming the RNA-seq gene count data.

B.2.3 Differential gene expression analysis

The Venn diagram in Figure B.7 shows the differential genes when all bacterial samples, definite and probable groups are compared with definite viral samples, and when the definite bacterial group only are compared with viral samples. The majority of the genes in the two comparisons overlap.

Figure B.8 shows the differentially expressed genes when the definite bacterial group was compared with the definite viral group.

The volcano plot in Figure B.9 shows the small number of DEGs when the acute viral group are compared with acute samples from the co-infection group. Figure

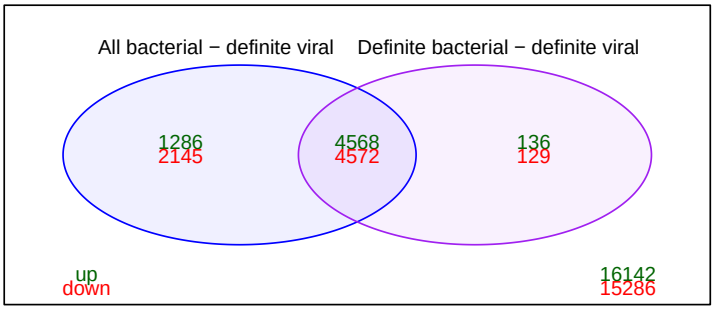


Figure B.7: Venn diagram showing the differential genes when all bacterial samples, definite and probable groups are compared with definite viral samples, and when the definite bacterial group only are compared with viral samples.

B.10 shows the DEGs when the acute bacterial group are compared with the acute co-infection group.

Pathway analysis - Pneumococcal acute compared with convalescent

Immunologic signature gene sets Using the immunologic signature gene sets, 1,314 pathways were found to have significantly changed enrichment scores when acute pneumococcal samples are compared with convalescent samples from the same participants. Following removal of redundant pathways, 1,032 significant pathways remained. Figure B.11 shows the pathways with the highest and lowest NES.

The top over-enriched pathway contains genes down-regulated in adults following the administration of an influenza vaccine. Other over-enriched pathways of

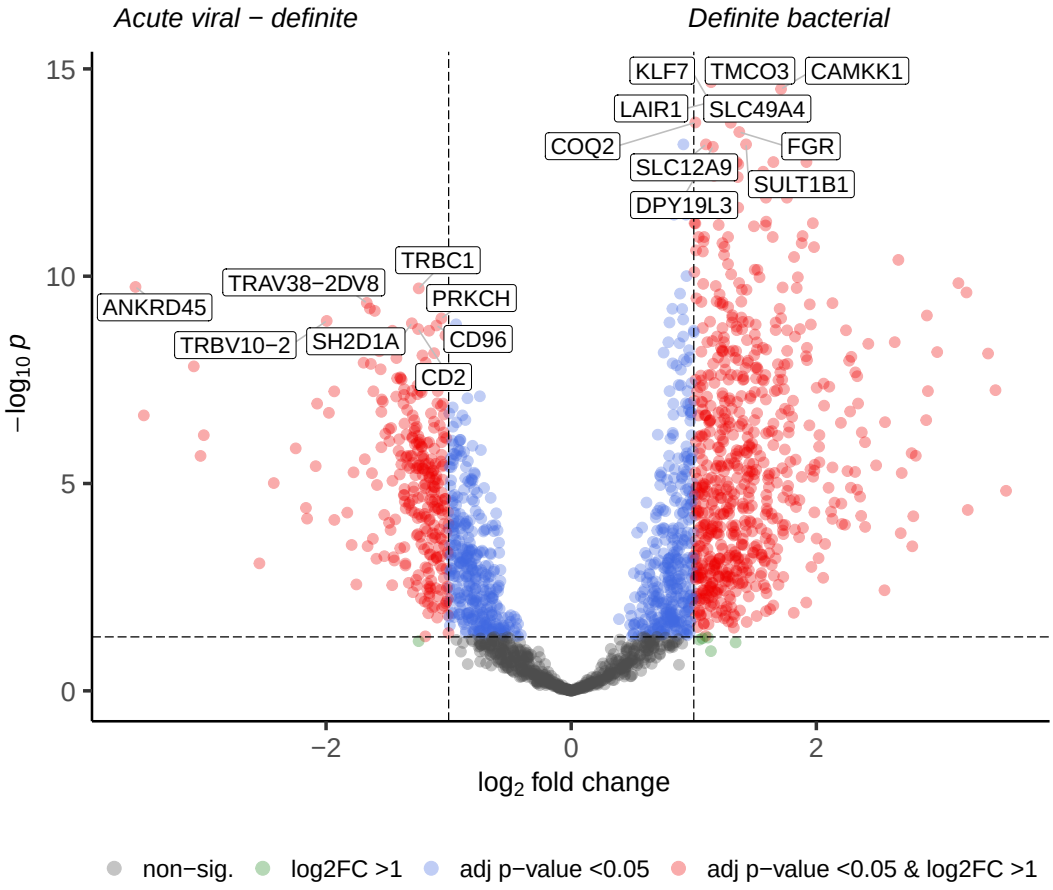


Figure B.8: Volcano plot showing differentially expressed genes when the acute definite bacterial group only are compared with the acute definite viral group, using the subset of genes that were differentially expressed when acute bacterial samples were compared with convalescent bacterial samples and acute viral samples were compared with convalescent viral samples.

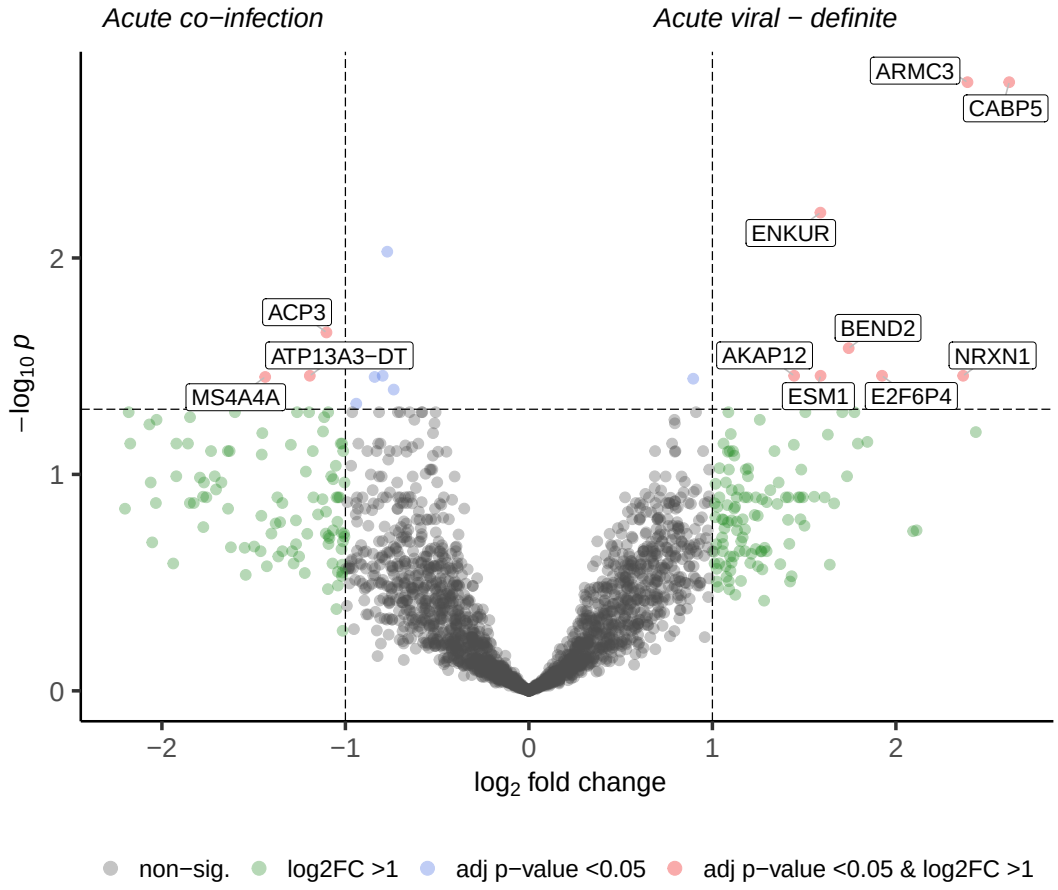


Figure B.9: Volcano plot showing the differentially expressed genes when the acute viral group are compared with acute samples from the co-infection group, DGE analysis using the subset of genes that were differentially expressed when acute bacterial samples were compared with convalescent bacterial samples and acute viral samples were compared with convalescent viral samples.

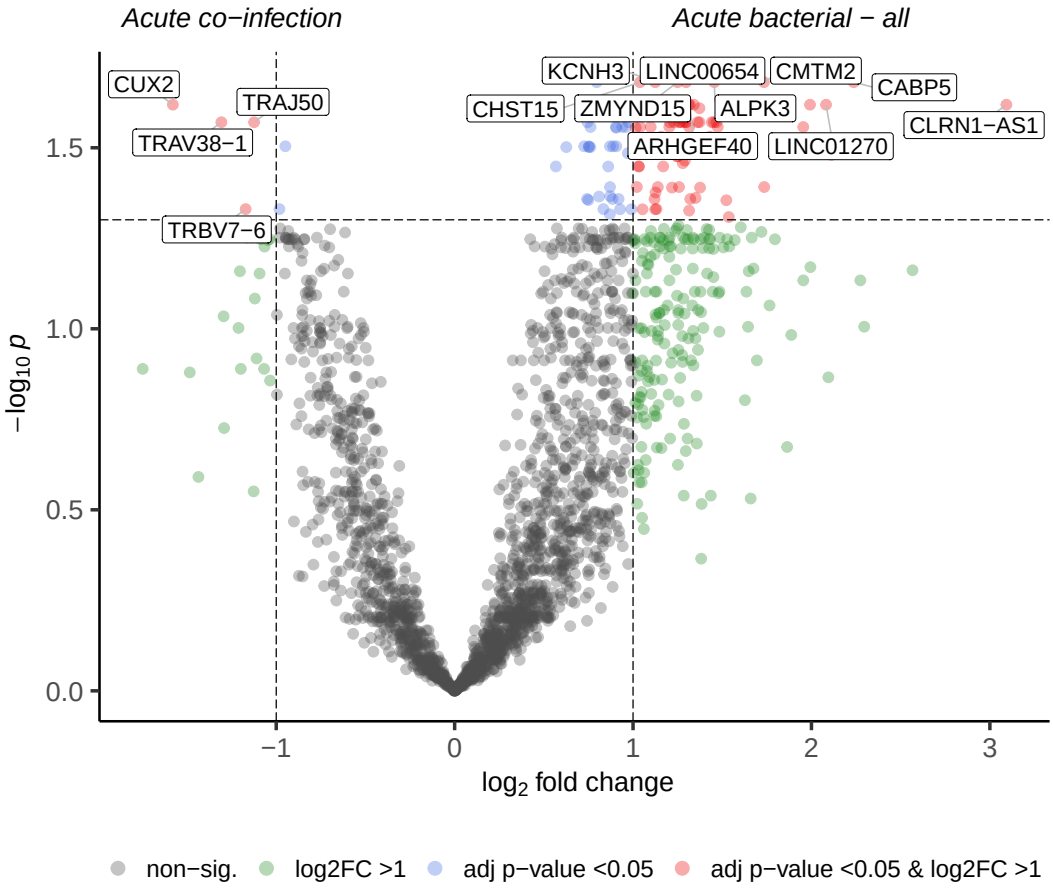


Figure B.10: Volcano plot showing the differentially expressed genes when the acute bacterial group are compared with acute samples from the co-infection group, DGE analysis using the subset of genes that were differentially expressed when acute bacterial samples were compared with convalescent bacterial samples and acute viral samples were compared with convalescent viral samples.

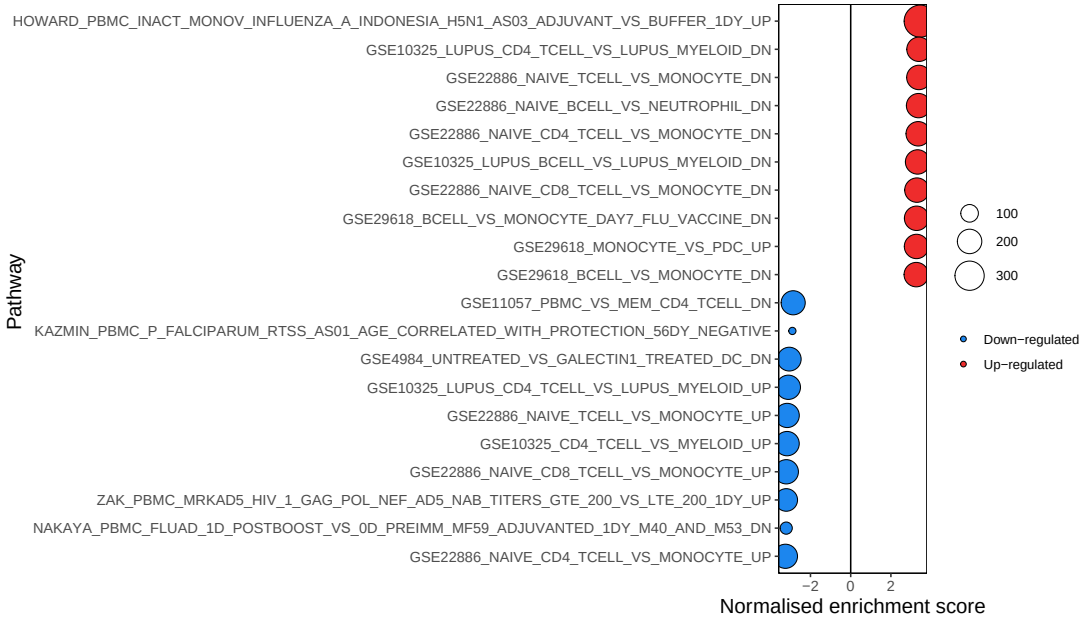


Figure B.11: Gene-set enrichment analysis results when acute pneumococcal samples were compared with convalescent samples. Using the immunologic signature gene sets, the ten pathways with the highest and lowest normalised enrichment scores are shown. The size of each circle represents the number of genes in the pathway.

interest include genes over-expressed in monocytes and neutrophils compared with other inflammatory cells.

Pathway analysis - Pneumococcal acute compared with other bacterial samples

In GSEA, when acute pneumococcal samples are compared with other, non-pneumococcal, bacterial samples the over-enriched pathways include NF-kappaB signalling, Figure B.12 and chemokine production, Figure B.13.

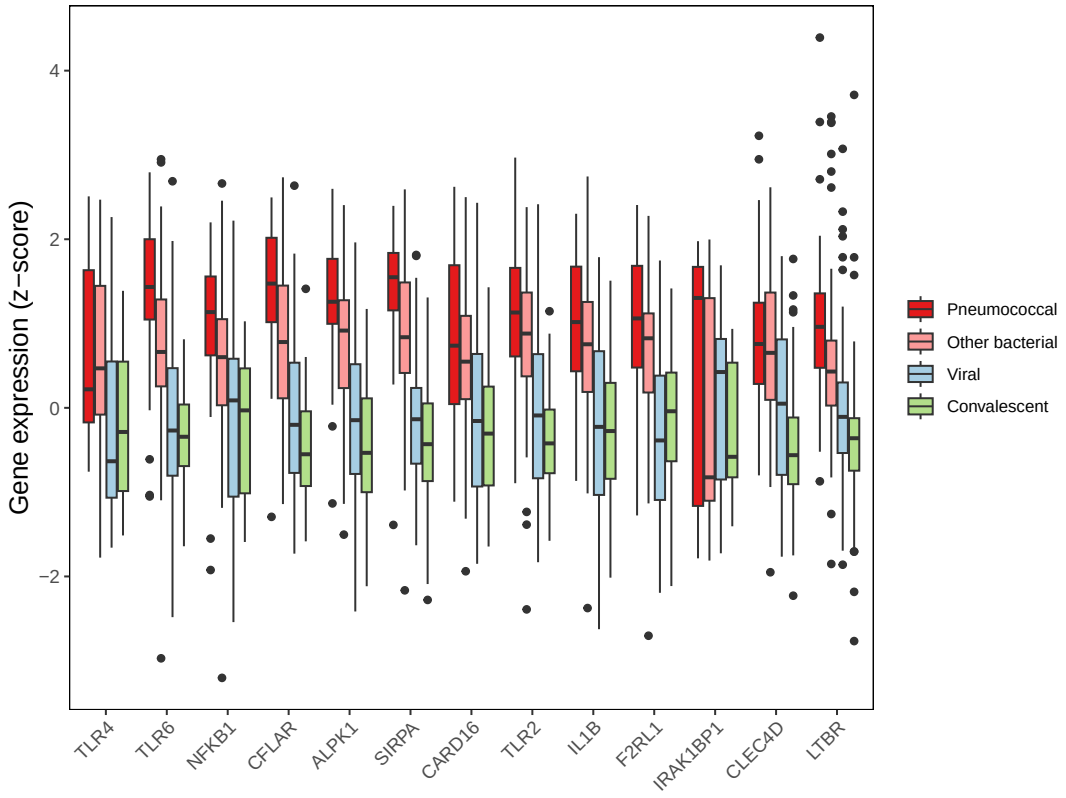


Figure B.12: Boxplots of the common genes between the leading edge of the gene set involving NF-kappa-B signal transduction, GO-BP canonical NF-kappaB signal transduction, and DE genes when acute pneumococcal samples were compared with convalescent samples.

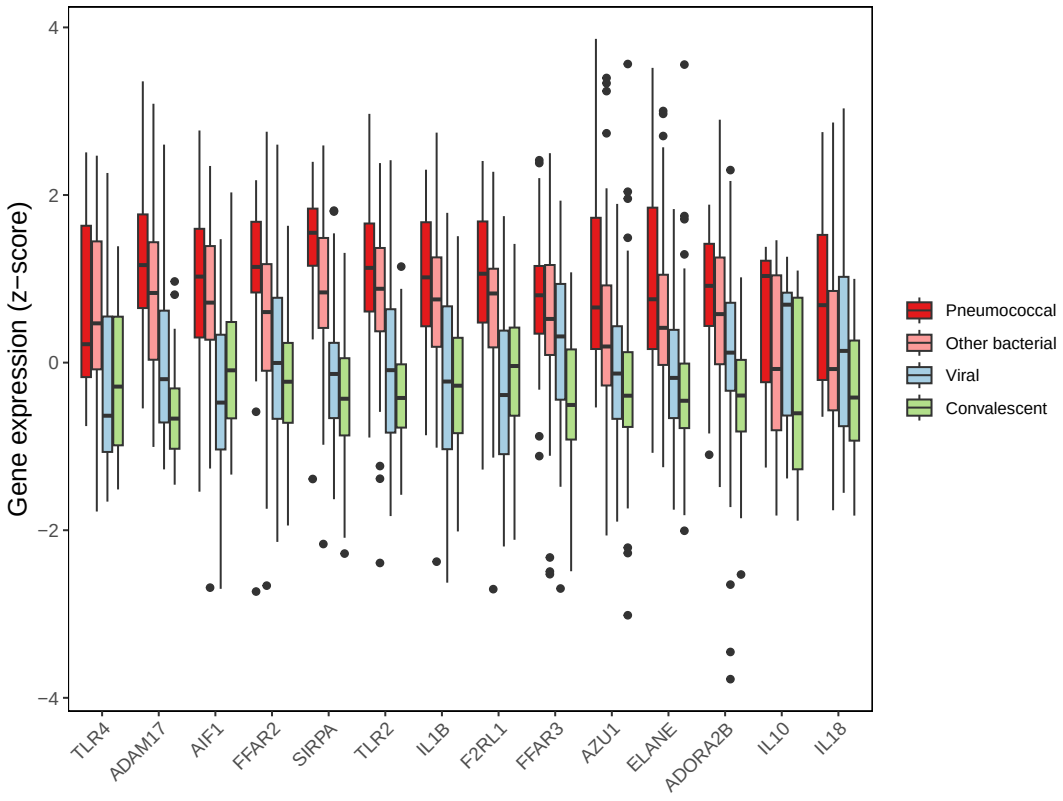


Figure B.13: Boxplots of the common genes between the leading edge of the gene set for chemokine production, GO-BP chemokine production, and DE genes when acute pneumococcal samples were compared with convalescent samples.

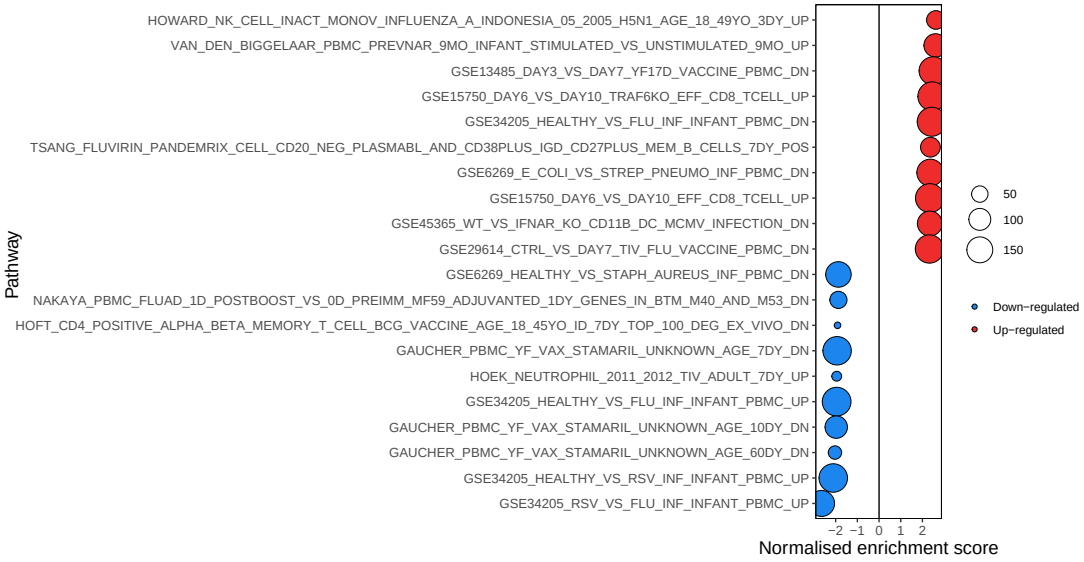


Figure B.14: Gene-set enrichment analysis results showing the ten pathways with the highest and lowest normalised enrichment scores when acute influenza samples were compared with convalescent samples. The size of each circle represents the number of genes in the pathway.

GSEA - Influenza subset genes

When the subset of DE genes between acute and convalescent influenza groups were used for GSEA using the GO-BP there were no significantly enriched pathways. When using the immune gene sets four gene sets were significantly enriched. Two of these gene sets are related to vaccines, measles, mumps and rubella (MMR) and an HIV vaccine. Another pathway related to genes in plasma cells compared with naive B cells. One pathway related to DE genes when acute RSV and acute influenza infections were compared, see Figure B.15.

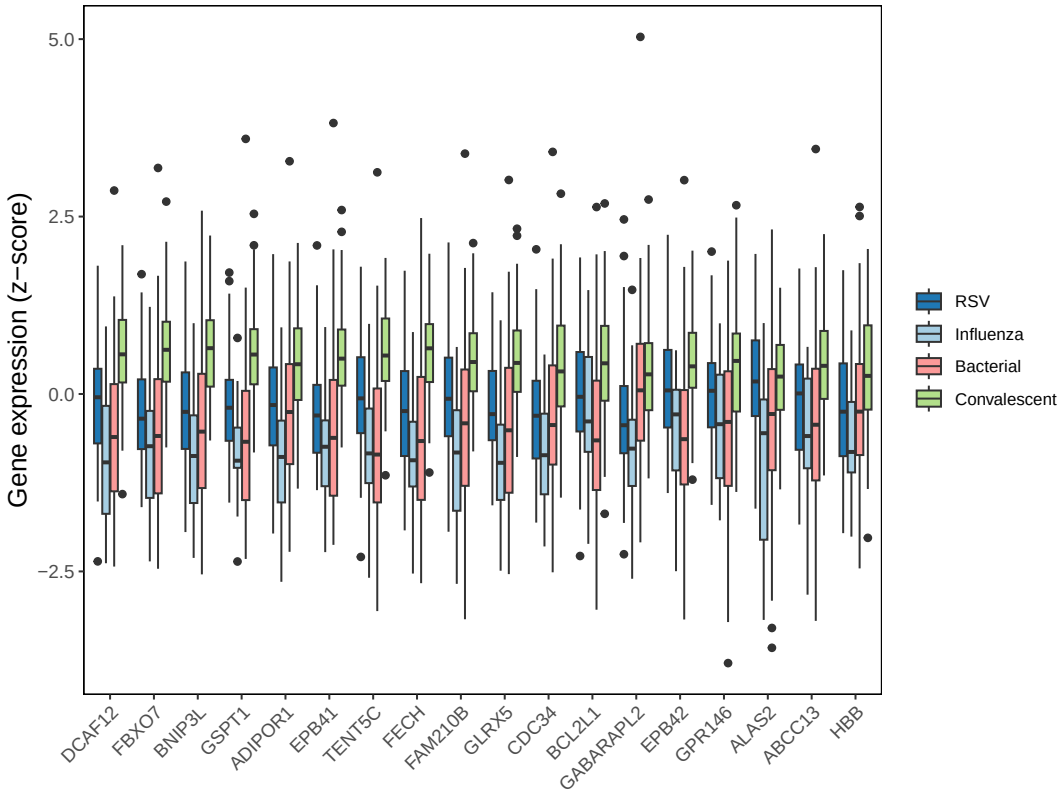


Figure B.15: Boxplots of the common genes between the leading edge of the gene set comparing RSV and influenza infections, (GSE34205 RSV vs flu inf infant PBMC up), and DE genes when acute influenza samples were compared with convalescent samples.

Appendix C

Appendix - Linked to Chapter 3, Proteomics

C.0.1 Supplementary methods

Filtering of low-quality proteins

Twenty-four low-quality samples were removed. These samples were all processed on the same plate, despite discussions with the team who performed the LC-MS no reason for these low-quality samples was identified, see Figure C.1.

Assessment and removal of undesired proteins

Nine proteins were marked as contaminants against the original spectral library search in DIA-NN. One of these was removed in the filtering steps. Of the remaining eight potential contaminants, each was assessed to see if it was

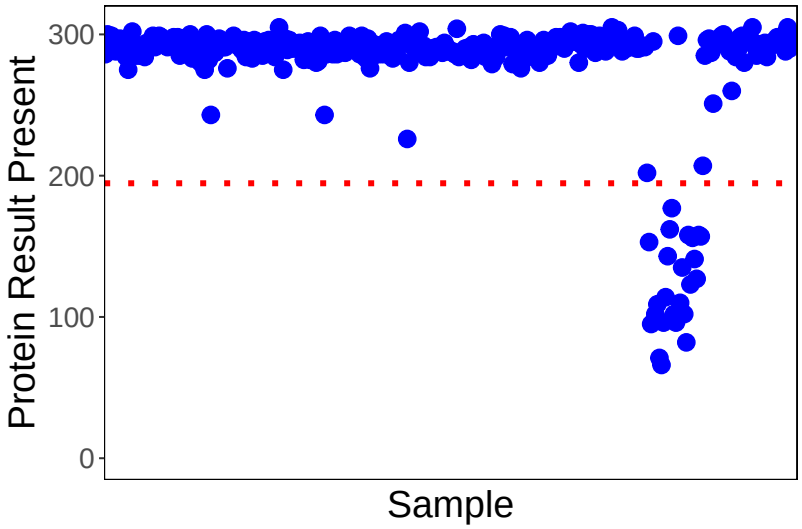


Figure C.1: Number of protein results present by sample, pre-filtering for low-quality samples. Each dot represents a sample, samples are plotted as they were analysed by the mass spectrometer. The dotted line represents 70% of the mean protein results across the entire cohort, n=195.

biologically likely that this protein was present in this cohort, see Table C.1. These contaminant proteins did not have a formal protein ID assigned to them, and it was not possible to confidently assign protein names to all these proteins. No evidence was found to support these proteins being routinely found in blood and so all were excluded.

Quality control plots

Boxplots of the normalised expression values were plotted for all samples, looking for outlying samples, see Figures C.2 and C.3. A scatter plot of the sample medians of the normalised expression values is shown in Figure C.4. Two outlier samples were identified, there was no biological reason for these samples to be outliers so they were removed from further analyses.

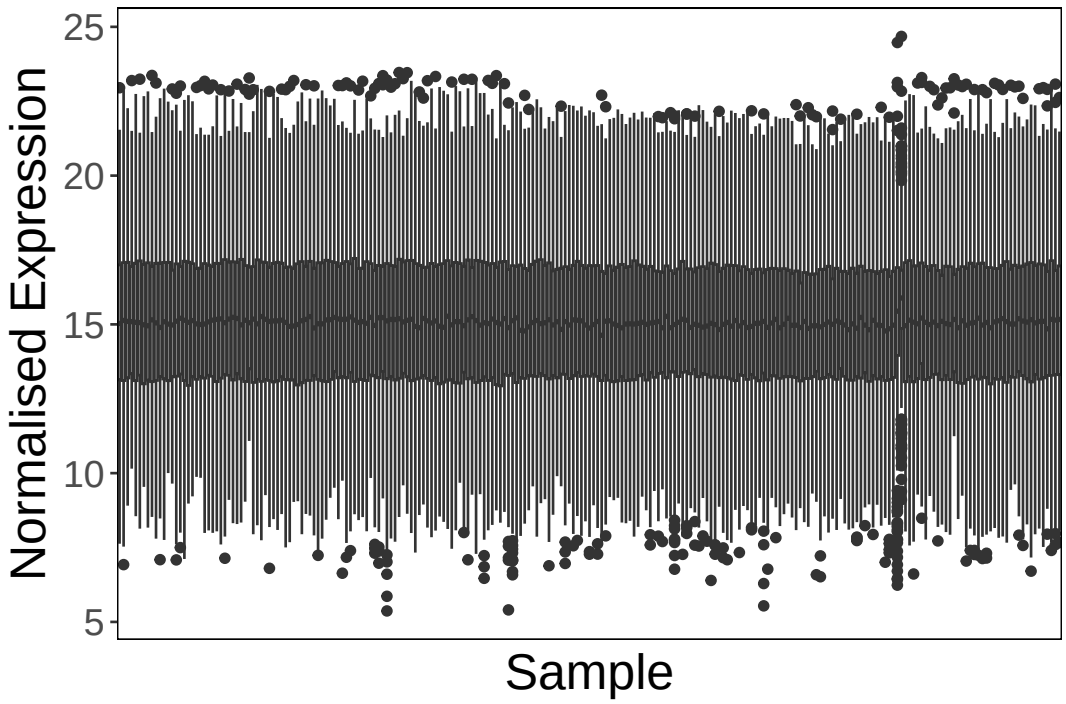


Figure C.2: Normalised median values for each acute sample. The box represents the interquartile range (IQR). The upper whisker extends to the largest value no further than 1.5 times IQR from the hinge. The lower whisker extends to the smallest value at most 1.5 times IQR from the hinge. Data beyond the end of the whiskers are outlying points and are plotted individually.

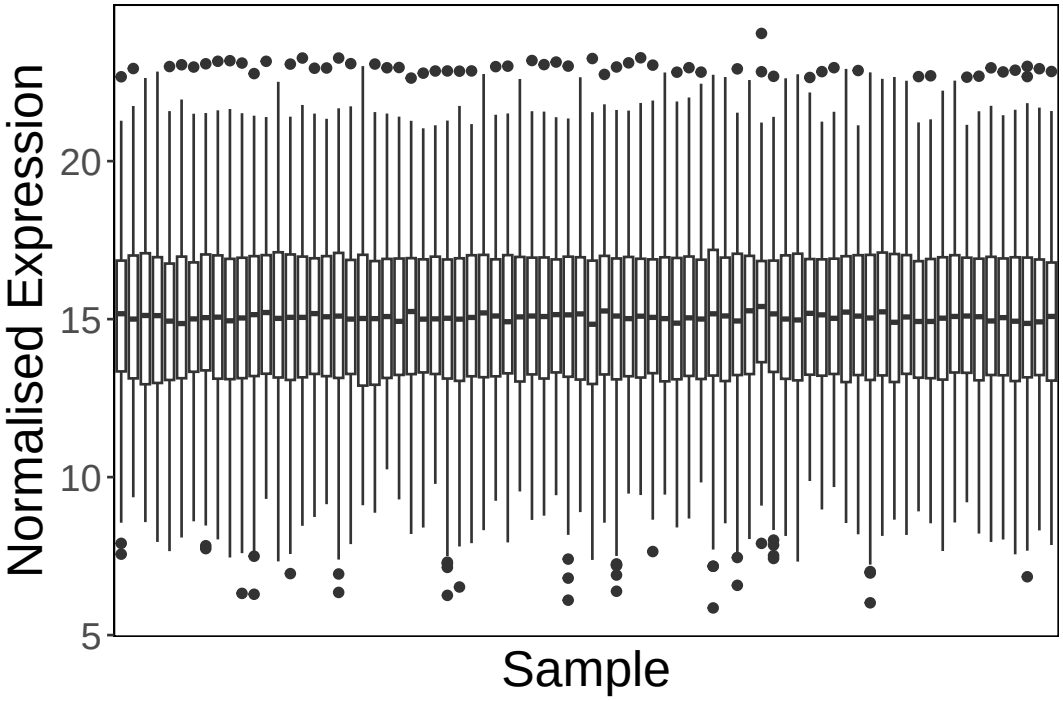


Figure C.3: Normalised median values for each convalescent sample. The box represents the interquartile range (IQR). The upper whisker extends to the largest value no further than 1.5 times IQR from the hinge. The lower whisker extends to the smallest value at most 1.5-time IQR from the hinge. Data beyond the end of the whiskers are outlying points and are plotted individually.

Table C.1: List of proteins labelled as contaminants following results of the spectral library search in DIA-NN, the "Decision" column states whether or not to include these proteins in further analyses.

Protein ID	Protein Name	Protein Description – From UniProt Atlas	Decision
I141_HUMAN	Unable to match to protein in UniProt or Protein Atlas database	Not found	Exclude
KERATIN02	Keratin, type II cytoskeletal 2 epidermal	Commonly found contaminant	Exclude
KERATIN03	Keratin, type II cytoskeletal 3	Involved in keratinisation, likely contaminant	Exclude
KERATIN06	Keratin, type I cytoskeletal 6A	Commonly found contaminant	Exclude
KERATIN13	Keratin, type I cytoskeletal 13	Found in tongue cells, likely contaminant	Exclude
NRL_1MCOH	Neural retina-specific leucine zipper protein	Found in retina, likely contaminant	Exclude
PIGR_HUMAN	Polymeric immunoglobulin receptor	Mediates selective transcytosis of polymeric IgA and IgM across mucosal epithelial cells	Exclude
TRYPSIN	Trypsin	Digestion enzyme used in sample preparation	Exclude

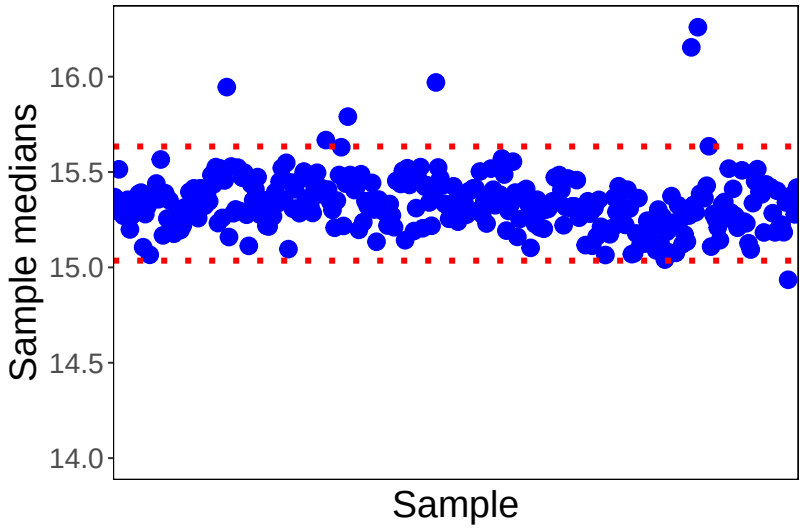


Figure C.4: Scatter plot showing the sample medians of the normalised expression values in the MS proteomics experiment. The red dotted lines represent 5% above and below the median for the cohort.

Density graphs of protein abundance were plotted for each sample, and each density graph was examined looking for abnormal density patterns. No abnormally skewed patterns, suggesting poor sample quality, were identified.

C.0.2 Supplementary results

To compare acute bacterial and acute viral LRTIs, PCA was repeated with acute LRTI samples only to look for differences between different LRTI groups. The biplot in Figure C.5 shows that while there is overlap between bacterial and viral cases, there is some separation between bacterial and viral cases. The weighted proteins along the direction of PC1, towards the bacterial samples, are CRP, lipopolysaccharide-binding protein LBP, alpha-1-antitrypsin (A1AT), complement component C9 (CO9). N-acetylmuramoyl-L-alanine amidase (PGRP2) and immunoglobulin kappa variable 1D-12 (KVD12) have high weightings along PC1, towards the viral grouping.

C.0.3 Differential protein abundance - co-infections

Differential protein abundance analysis was carried out comparing acute viral samples to acute bacterial-viral co-infections. There were no significantly differentially abundant proteins, as shown in Figure C.6. Similarly, there were no significantly differentially abundant proteins when acute bacterial samples were compared with acute co-infections, as shown in Figure C.7.

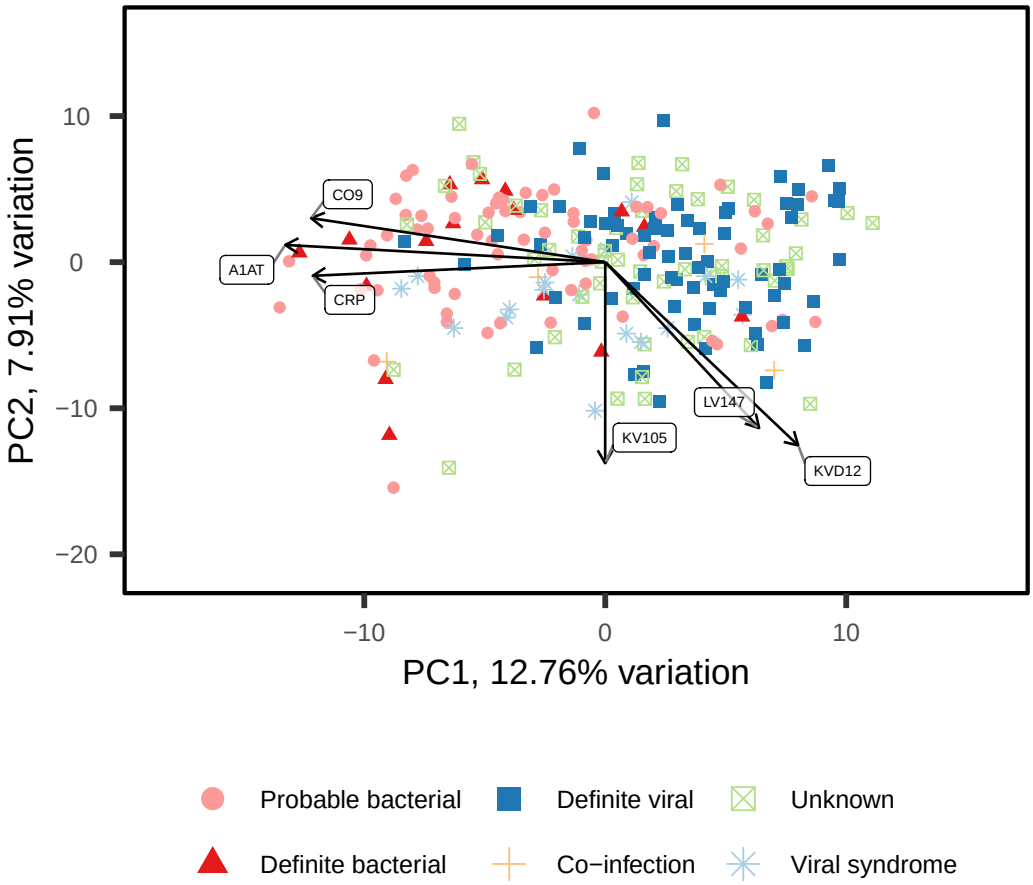


Figure C.5: PCA in the MS proteomics experiment. Principal component biplot comparing PC1 and PC2 coloured by LRTI classification, post-batch correction. The loadings for the top-weighted proteins are shown.

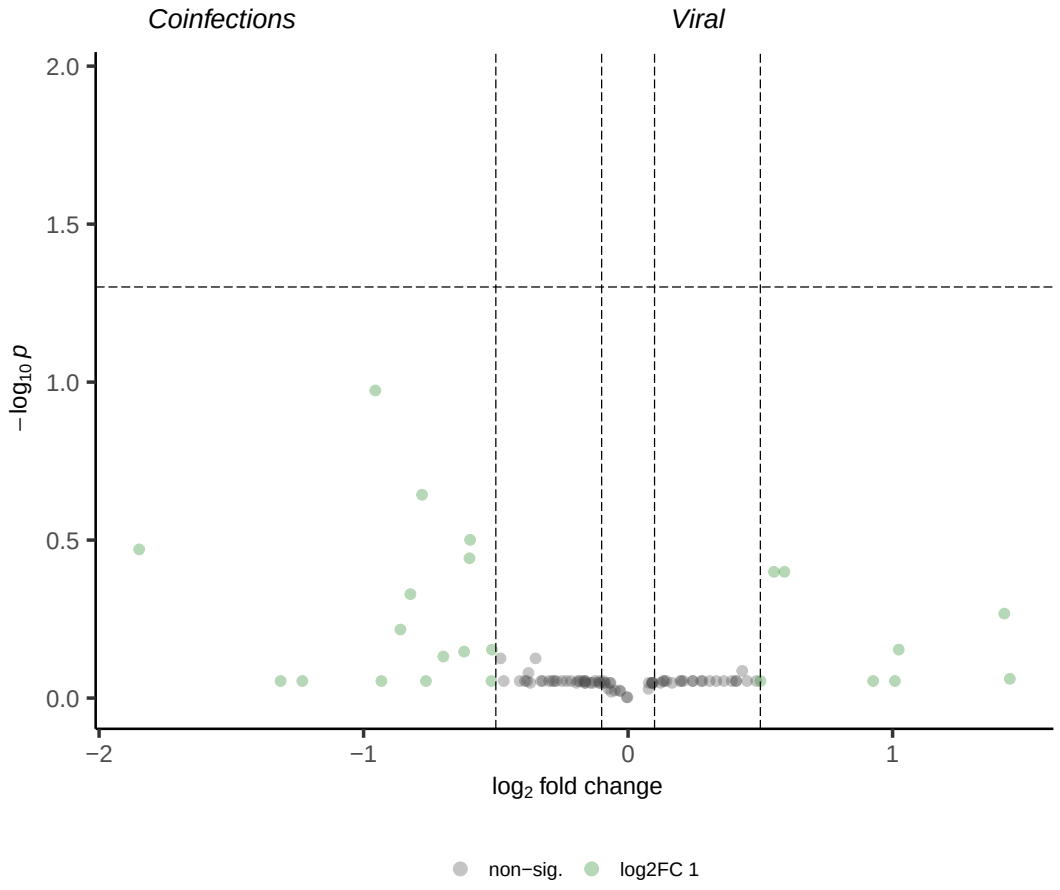


Figure C.6: Volcano plot showing differentially expressed genes between participants in the acute viral group and the acute co-infection LRTI group, in the MS proteomics experiment. Participants in the training dataset only included. The labelled genes are the top ten up- and down-regulated genes, ranked by Benjamini-Hochberg adjusted p-values.

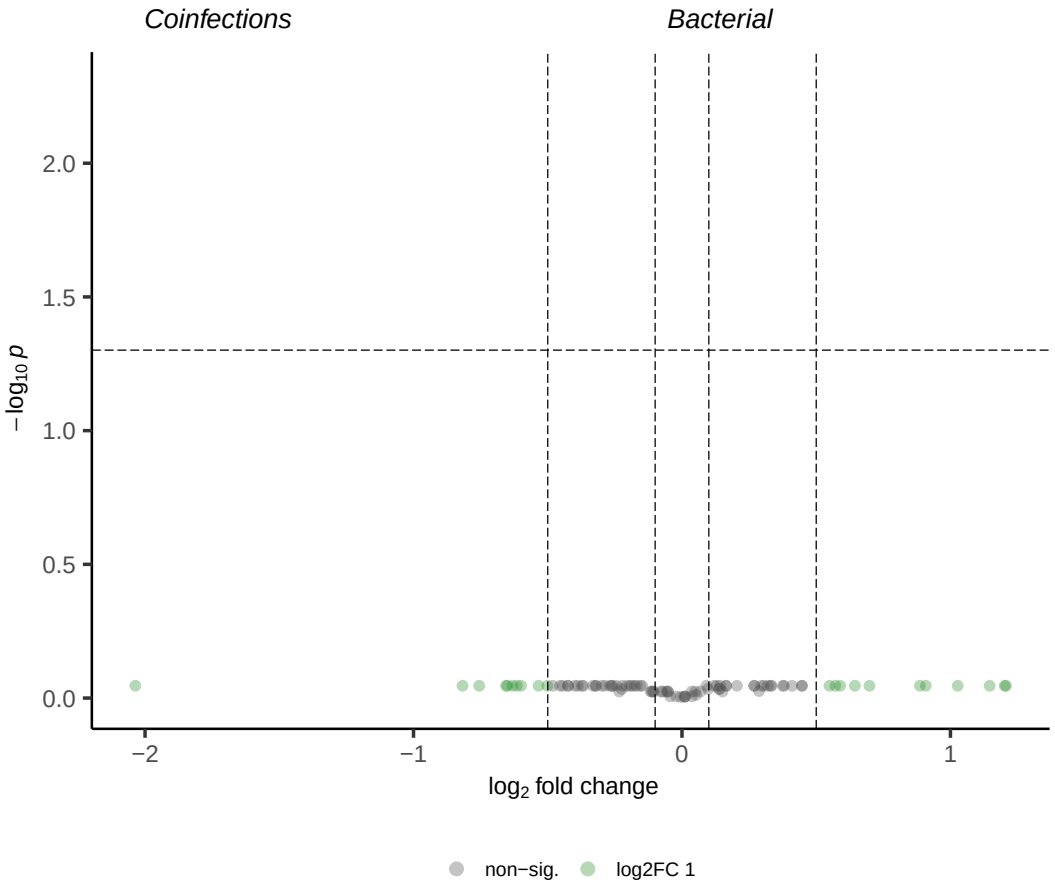


Figure C.7: Volcano plot showing differentially expressed genes between participants in the acute bacterial group and the acute co-infection LRTI group, in the MS proteomics experiment. Participants in the training dataset only included. The labelled genes are the top ten up- and down-regulated genes, ranked by Benjamini-Hochberg adjusted p-values.

C.0.4 Pathway analysis - influenza

Using GSEA, figure C.8 shows box plots of proteins contributing to the high-density lipoprotein particle clearance pathway. This pathway was over-represented when acute influenza samples were compared with convalescent samples from individuals who had previously presented with viral LRTIs.

C.0.5 Pathway analysis - pneumococcal

When comparing the acute pneumococcal group with acute samples from participants with other, non-pneumococcal LRTIs there were no significantly differentially abundant proteins. The results were the same whether only samples that were blood or pleural fluid positive for *S. pneumoniae* were included in the pneumococcal group, see Figure C.9 or if samples that were positive for *S. pneumoniae* Serotype 1 or 5 in their nasopharyngeal sample were added to the acute pneumococcal group, see Figure C.10.

Using ORA, when the acute pneumococcal group are compared with convalescent samples from participants with bacterial LRTIs, plasminogen activation pathways are over-represented, see Figure C.11 for box plots of proteins that contribute to this pathway grouped by viral, pneumococcal, other bacterial and convalescent samples.

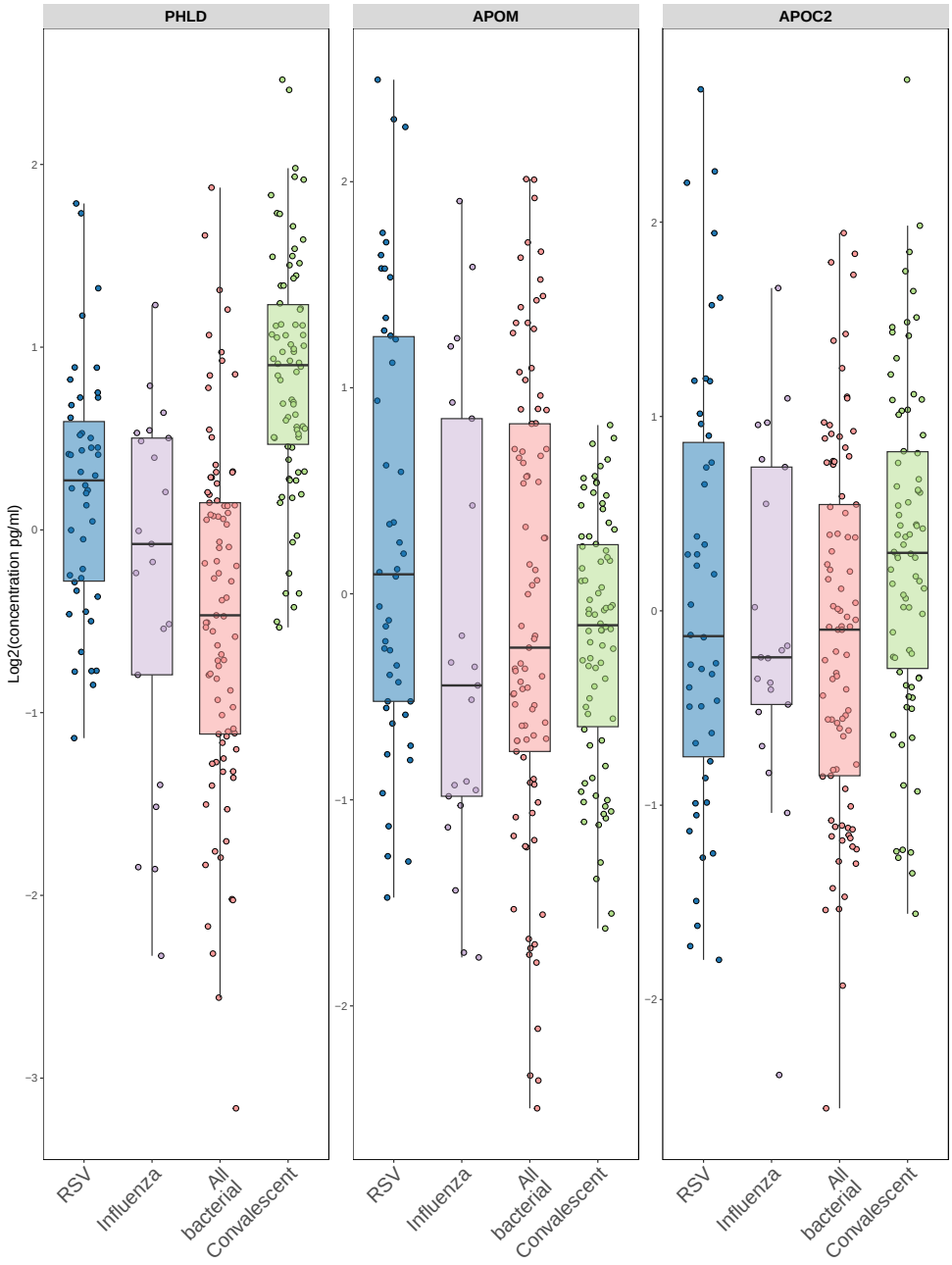


Figure C.8: Proteins which contribute to the high-density lipoprotein particle clearance pathway from the over-representation analysis results when participants in the acute influenza group are compared with convalescent samples from participants with viral LRTIs. All of the bacterial sample results and all of the convalescent sample results are grouped for comparison.

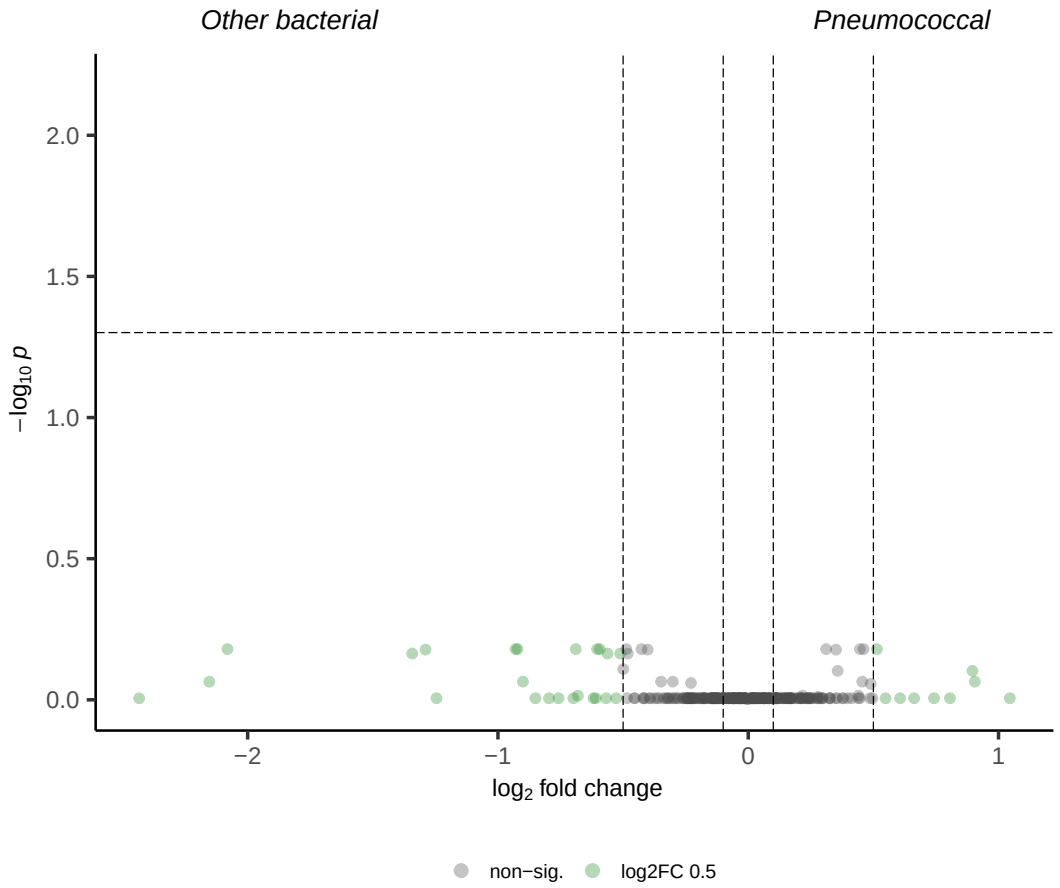


Figure C.9: Volcano plot showing that there were no differentially abundant proteins between participants in the acute pneumococcal group and acute samples from participants with other, non-pneumococcal, bacterial, LRTI. Only samples that were blood or pleural fluid positive for *S. pneumoniae* were included in the pneumococcal group. The labelled proteins are the top ten up- and down-regulated proteins, ranked by Benjamini-Hochberg adjusted p-values.

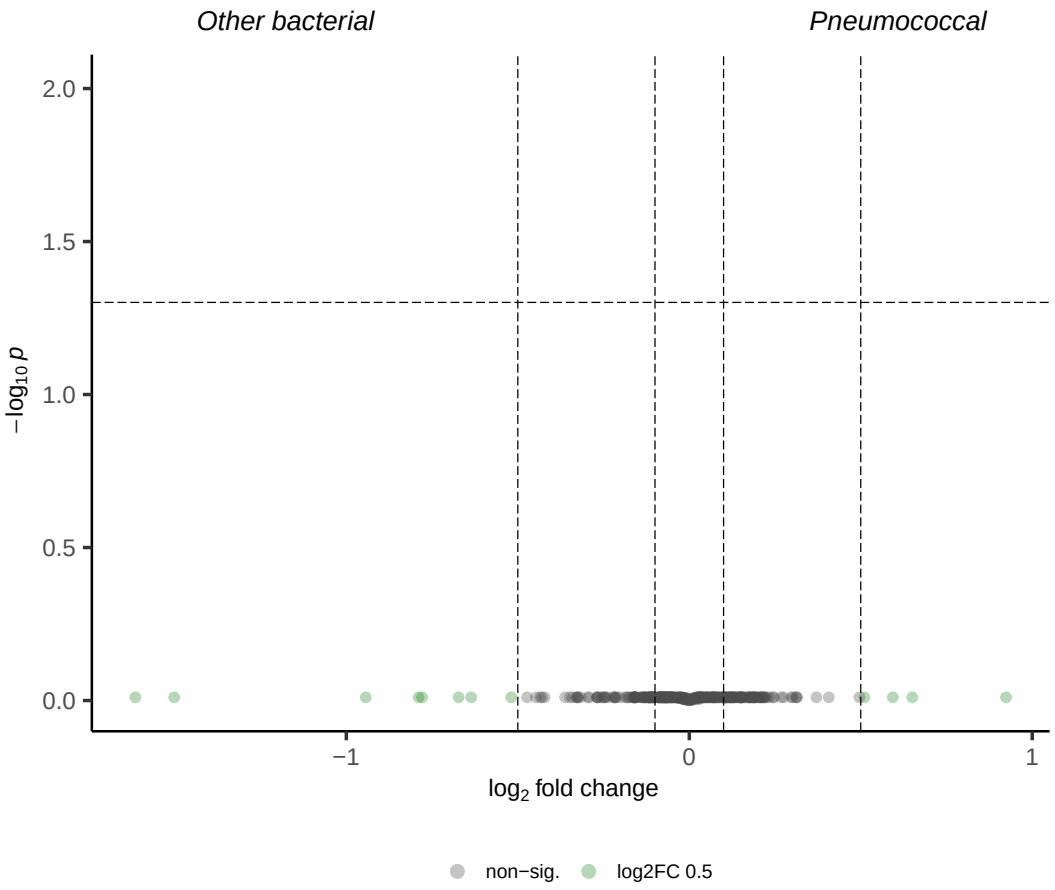


Figure C.10: Volcano plot showing that there were no differentially abundant proteins between participants in the acute pneumococcal group and acute samples from participants with other, non-pneumococcal, bacterial, LRTI. Samples that were blood or pleural fluid positive for *S. pneumoniae* or samples in the probable bacterial group that were positive for *S. pneumoniae* Serotype 1 or 5 in their nasopharyngeal sample. The labelled proteins are the top ten up- and down-regulated proteins, ranked by Benjamini-Hochberg adjusted p-values.

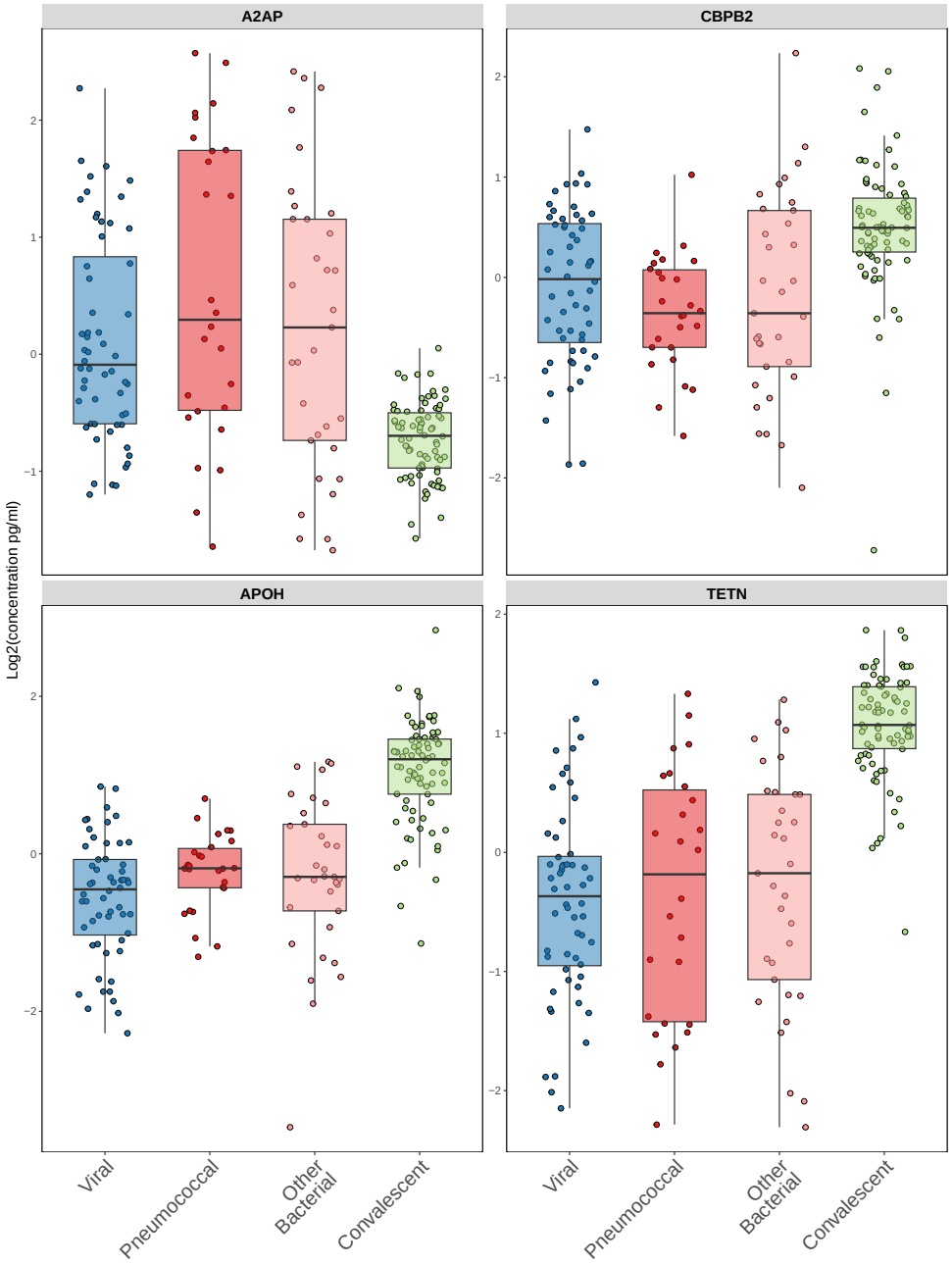


Figure C.11: Proteins which contribute to the plasminogen activation pathways from the over-representation analysis results when participants in the acute pneumococcal group are compared with convalescent samples from participants with bacterial LRTIs. All of the viral sample results and all of the convalescent sample results are grouped for comparison.

Appendix D

Appendix - Linked to Chapter 4, Cytokine panel

D.0.1 Supplementary methods

Step-by-step MSD panel laboratory methods are outlined here.

Day Before MSD Assay

1. Choose 40 samples to include in assay. Total on plate: 16 to create a standard curve and 80 sample positions (40 samples in duplicate needed per plate)

2. Make Excel file with layout of plate in table.

3.

Day 1 of MSD Assay

Taken from MSD U-PLEX Platform – U-PLEX Custom Biomarker Group 1 (Human) Multiplex Assays – 18166-V4-2021Feb

1. When arrive – take MSD kit out of fridge – should be at room temperature before starting (45 mins to 1 hour at room temperature prior to starting)
2. Note timings of taking kit out of fridge and when started protocol
3. Extra consumables to get ready on day 1
    (a) Jar with disinfectant for used tips etc.
    (b) Small tubes (10) with screw top lids for mixture of antibody and linker (location: freezer room)
    (c) Falcon tubes x1 – bigger 10ml+ tubes with blue top.
    (d) Plastic trays for accessing solutions (x1).
    (e) Glass jar if need to make new wash buffer.
    (f) Single and multichannel pipettes.
    (g) Filtered pipette tips to fit the above.
    (h) Wider plastic container to tip plate contents into.
    (i) Adhesive plate seal (and light proof seal for top).

STEP 1 – Create Individual U-PLEX Linker-Coupled Antibody Solutions

1. Assign antibody to Linker (assign the coloured numbers (these represent each linker) on the spot map to the antibody of interest). Couple each biotinylated capture antibody to a unique linker and record the antibody identity next to the linker number on the spot map.

2. Take 10 small tubes with screw top lids and mark the name of the antibody on the top of each lid

3. Start with 300 microL of the assigned linker to the small tubes – vortex linker and put linker in new tube.

   (a) Add 200 microL of each biotinylated antibody to each linker – vortex antibody briefly before adding.

   (b) Mix by vortexing for at least 5 seconds after.

   (c) Incubate at room temperature for 30 minutes WITHOUT shaking – start timer at first linker and check time again when last linker added.

4. Add 200 microL of stop solution. Mix by vortexing (at least 5 seconds). Incubate at room temperature for 30 mins WITHOUT shaking.

STEP 2: Prepare the Multiplex Coating Solution (technique NB, as taking most of fluid from tubes)

1. Take falcon tube. Add 600 microL of each U-PLEX linker-coupled antibody solution (10x concentration) to the falcon tube and VORTEX (at least 5 seconds)

2. This will be 6mls if combining 10 antibodies, and do not need to add

anything else at this point. (If combining less than 10 antibodies then bring solution up to 6mls with stop solution.)

3. VORTEX for at least 5 seconds

STEP 3: Coat the U-PLEX Plate

1. Take the U-PLEX plate
2. Add 50 microL of multiplex coating solution to each well. Seal the plate with an adhesive plate seal. And add light protective seal. Incubate with shaking at room temperature for 1 hour. Shaking is best around 700 rpm, but anywhere from 500 to 1000 is fine.
3. Take MSD wash buffer, if have run out then need to prepare this
   (a) Take a glass jar with a screw on lid.
   (b) Add 10mls of concentrated wash buffer.
   (c) Add 190mls dH20 (deionised water) from tap in lab at the end.
4. Pour wash buffer into plastic discard tray.
5. Wash the plate 3 times.
   (a) Using the multichannel pipette, take up 150 microL of the diluted wash buffer (x1).
   (b) Add 150 microL of wash buffer to each well.
   (c) Invert the plate into the wide plastic container.
   (d) Gently tap dry onto some paper towels

(e) Repeat (a) to (d) twice more.

(f) Seal the plate.

(g) Plate is now coated and read for use. These coated plates may be stored in original pouch, sealed for up to 7 days at 2-8C.

END of Day 1

Day 2 of MSD Assay

When arrive – take diluents out of freezer (diluents 57 and 3). When arrive– take MSD kit out of fridge – should be at room temperature before starting When arrive - take samples out and put on dry ice to allow to defrost slowly. Extra consumables needed:

1. Eight small tubes with flip caps for calibrator standard.
2. Falcon tube x1.
3. Pipettes and tips.

Prepare calibrator standards.

1. Take 8 small tubes with flip caps.
2. Label the tubes 1 to 8.
3. Take each calibrator vial, three provided
   (a) Add 250 microL of Diluent 57 to each calibrator vial.

(b) This results in a 5x concentrated stock of each calibrator.

(c) Invert the reconstituted calibrator 3 times, DO NOT VORTEX.

(d) Let me reconstituted solution equilibrate at room temperature for 15-30 mins, THEN VORTEX briefly.

4. Combine the calibrators in a clean polypropylene tube (small tubes with flip caps), to make the calibrator standard 1, see Table D.1.

5. Serially dilute the calibrator standard 1 to generate the standard curve.

(a) Add 225 microL (note change in vol from previous step) of Diluent 57 to each of the 8 tubes labelled 2 to 8.

(b) Add an extra 75 microL of Diluent 57 to tube 8 (see Table D.2).

(c) Vortex calibrator standard 1.

(d) Transfer 75 microL from standard 1 to 2.

(e) Vortex standard 2.

(f) Transfer 75 microL from standard 2 to 3.

(g) Vortex standard 3.

(h) Repeat the same process until standard 7.

(i) Vortex standard 7.

(j) DO NOT TRANSFER any fluid from standard 7 to 8

STEP 1: Add Samples and Calibrators

1. Take MSD plate and add 25 microL of Diluent 57 (this may change with different antibodies) to each well.

Table D.1: Using the MSD platform: Combining calibrators to generate the calibrator standard 1 (top of the curve) level.

No. of Calibrator Blends	Volume of Reconstituted Calibrator (microL)	Diluent 57 (microL)	Total Volume (microL)
1	50	200	250
2	50 each	150	250
3	50 each	100	250
4	50 each	50	250
5	50 each	0	250

Table D.2: Using the MSD platform: Serial dilution to generate the standard curve. One part calibrator to 3 parts diluent = 4-fold dilution.

Calibrator Standard No.	Tube No.	Source of Calibrator	Volume of reconstituted calibrator (microL)	Assay diluent (microL)	Total volume (microL)
1	1	Calibrator standard 1 (top of curve)	75	225	300
2	2	From tube 1	75	225	300
3	3	From tube 2	75	225	300
4	4	From tube 3	75	225	300
5	5	From tube 4	75	225	300
6	6	From tube 5	75	225	300
7	7	From tube 6	75	225	300
8 (zero calibrator)	8	-	0	300	300

2. Tap the plate gently on all sides.

3. Vortex STANDARD or SAMPLE just PRIOR to putting on plate.

4. Add 25 microL of the prepared calibrator standard or sample to each well.

5. Seal the plate with adhesive seal.

6. Put on shaker, at room temperature for 1 hour. Incubate, while shaking, at room temperature for 1 hour.

7. While plate is shaking prepare detection antibody solution.

8.

Prepare Detection Antibody Solution

1. Take the detection antibody MSD bottle (yellow tubes). These are 100x when they arrive.

2. Take a falcon tube and add 60 microL of each detection antibody (yellow tubes) to the falcon tube.

3. Add Diluent 3 to bring the final volume to 6 mls.

4. If using 10 antibodies then add 5.4 mls of Diluent 3 to the falcon tube.

5. Vortex for a few seconds and use immediately, if not using immediately put in fridge.

STEP 2: Wash and Add Detection Antibody Solution

1. Wash the plate 3 times with at least 150 microL of wash buffer.

    (a) Using the multichannel pipette. Take up 150 microL of the diluted wash buffer (x1).

    (b) Add 150 microL of wash buffer to each well.

    (c) Invert the plate into the wide plastic container.

    (d) Gently tap dry onto some paper towels.

    (e) Repeat (a) to (e) two more times.

2. Add 50 microL of detection antibody solution to each well.

3. Seal the plate.

4. Incubate while shaking at room temperature for 1 hour.

5.

STEP 3: Wash and Read

1. 17. Wash the plate 3 times with at least 150 microL of wash buffer.

    (a) Using the multichannel pipette. Take up 150 microL of the diluted wash buffer (x1).

    (b) Add 150 microL of wash buffer to each well.

    (c) Invert the plate into the wide plastic container.

    (d) Gently tap dry onto some paper towels.

    (e) Repeat “a to e” twice more.

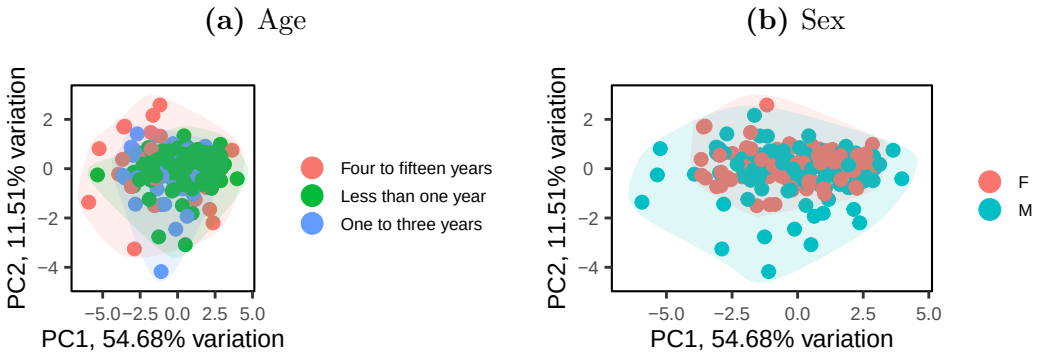


Figure D.1: PCA of results in the cytokine experiment. Principal component biplot post-batch correction coloured by age, panel (a), and sex, panel (b). PC1 and PC2 plotted. Samples from all nine MSD plates included if samples had a complete set of results.

2. Add 150 microL of MSD GOLD Read Buffer B to each well (incubation in GOLD buffer is not required before reading).

3. Analyse the plate on the MSD instrument. Download the .txt file.

4. At end of day – if this is the first thaw for Diluents 3 and 57 then aliquot these into suitable volumes before refreezing.

5. Record details of plate/MSD kit and record if any samples have low volume.

D.0.2 Supplementary results

PCA was perfomred post-batch correction. Figure D.1 shows PC biplots comparing the first two components, coloured by age, Panel (a) and sex, Panel (b).

Acute viral group compared with acute definite bacterial group

When the acute viral LRTI group was compared with the acute definite bacterial group, there were no significant differences between any of the ten cytokine levels, see Figure D.2.

Statistical results for comparisons across groups

Table D.3 shows the raw and adjusted p-values when the acute definite bacterial group was compared with the acute definite viral group.

Table D.4 shows the raw and adjusted p-values when the acute RSV group was compared with acute samples from bacterial groups and other, non-RSV, viral groups.

Table D.5 shows the raw and adjusted p-values when the acute RSV group was compared with acute other, non-RSV, viral groups only.

Model performance without confounding factors - cytokines only

When known confounders, such as age, are not built into the model, a ten-cytokine signature had the highest accuracy. See Figure D.3 for the accuracy measures using recursive feature selection. However, there was a marginal difference in accuracy between a ten-cytokine model and a three-cytokine model including IL-6, G-CSF and IL-15.

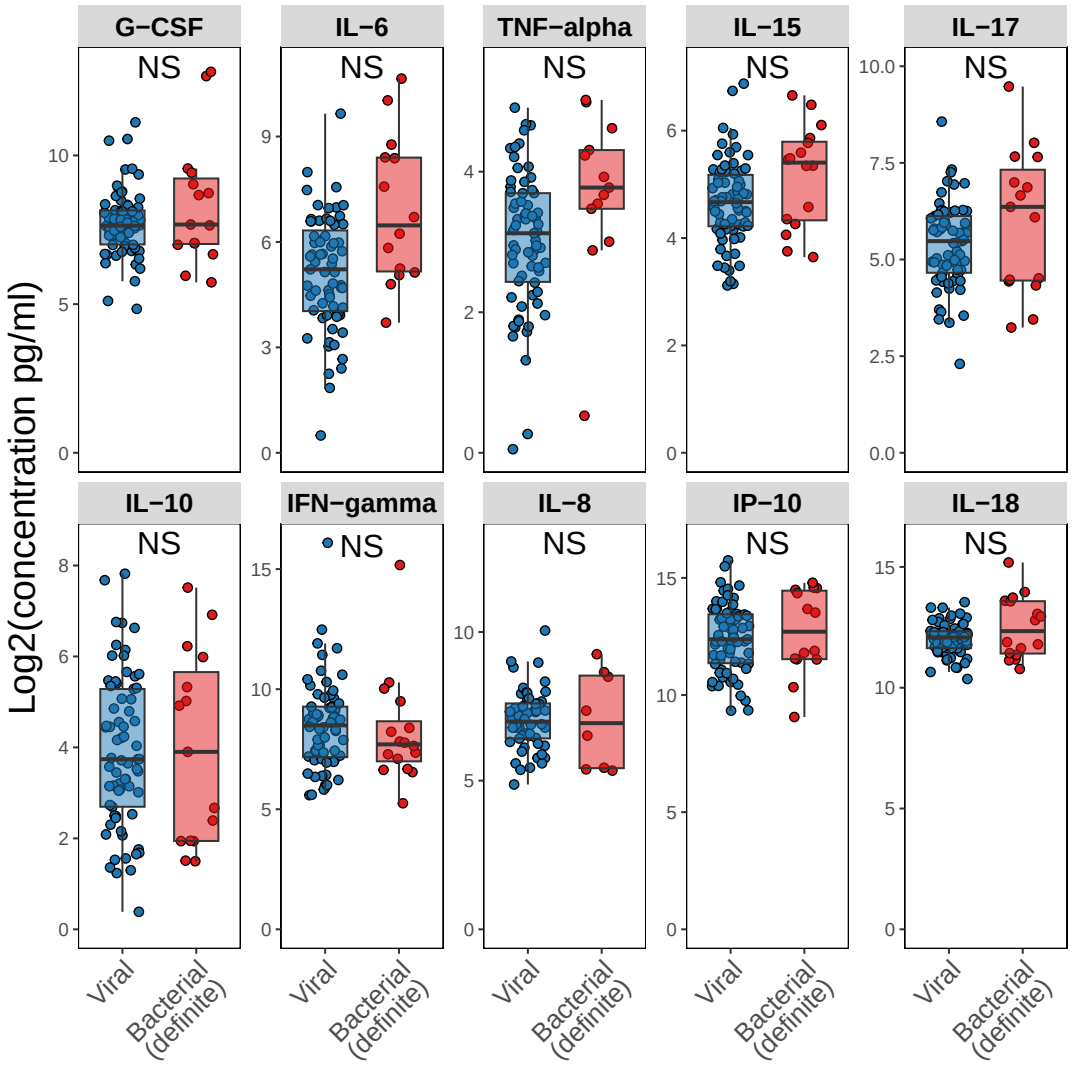


Figure D.2: In the LRTI cohort study, results from the viral LRTI group, n=70, for each of the ten cytokines tested, compared with the definite bacterial LRTI group, n=16.

Table D.3: Raw and adjusted p-values when the acute definite bacterial group was compared with the acute definite viral group. Mann-Whitney U tests were used to look for statistical significance between the groups, results were adjusted for multiple testing using the Benjamini-Hochberg Procedure.

Definite bacterial compared with definite viral	Adjusted p-values	Raw p-values
IFN-gamma	0.456525	0.358725
IL-6	0.000044	0.000004
IL-8	0.383382	0.230029
IP-10	0.383382	0.217797
IL-10	0.373029	0.149212
IL-18	0.966503	0.966503
G-CSF	0.003224	0.000645
TNF-alpha	0.966503	0.937666
IL-17	0.456525	0.36522
IL-15	0.312778	0.093833

Table D.4: Raw and adjusted p-values when the acute RSV group was compared with acute samples from bacterial groups and other, non-RSV, viral groups. Mann-Whitney U tests were used to look for statistical significance between the groups, results were adjusted for multiple testing using the Benjamini-Hochberg Procedure.

RSV acute compared with other acute samples (bacterial and non-RSV viral groups)	Raw p-value	Adjusted p-value
G_CSF	7.52E-03	0.03365
IL_6	4.52E-05	0.000452
TNF_alpha	3.54E-01	0.707616
IL_15	1.01E-02	0.03365
IL_17	6.94E-01	0.841753
IL_10	6.63E-01	0.841753
IFN_gamma	6.42E-01	0.841753
IL_8	3.06E-01	0.707616
IP_10	9.23E-01	0.923373
IL_18	7.58E-01	0.841753

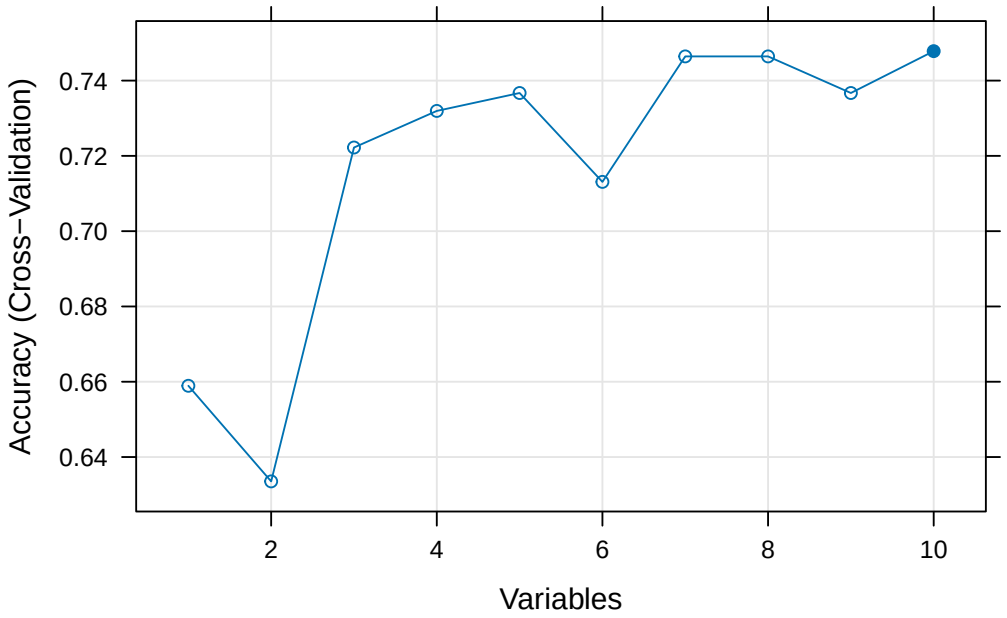


Figure D.3: Using recursive feature elimination to determine the accuracy of different models and the ideal number of variables to include, considering all ten cytokines. A model including ten cytokines had the highest accuracy level.

Table D.5: Raw and adjusted p-values when the acute RSV group was compared with acute other, non-RSV, viral groups only. Mann-Whitney U tests were used to look for statistical significance between the groups, results were adjusted for multiple testing using the Benjamini-Hochberg Procedure.

RSV acute compared with acute other virus groups	Adjusted p-values	Raw p-values
IFN-gamma	0.8518789	0.681503
IL-6	0.854048	0.768643
IL-8	0.8518789	0.40333
IP-10	0.8518789	0.64038
IL-10	0.8518789	0.190352
IL-18	0.8518789	0.680973
G-CSF	0.6800934	0.068009
TNF-alpha	0.8518789	0.593451
IL-17	0.9123469	0.912347
IL-15	0.8518789	0.464905

Model performance - Training dataset

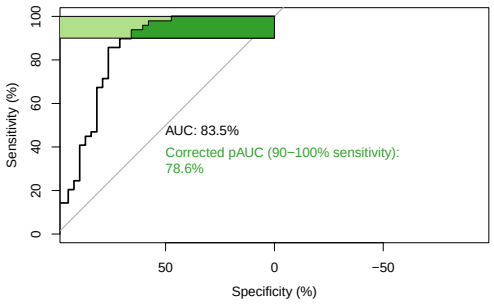
The ten-cytokine and three-cytokine models performed similarly in the training dataset at differentiating between bacterial and viral LRTIs. The specificity was the same for both models, 86%. The three-cytokine model had a slightly lower sensitivity, 68% compared with 74% for the ten-cytokine model.

The AUC for the ROC curves was similar for both models, 83.5% for the ten-cytokine model and 80% for the three-cytokine model. When the AUC was restricted to a high-sensitivity range, 90%-100% the pAUC was 78.6% for the ten-cytokine model and 68.3% for the three-cytokine model.

Table D.6: Cytokine model performance statistics in the training data. Comparison of results when a ten-cytokine and a three-cytokine model are used to differentiate between bacterial and viral LRTIs.

Performance metric	**Ten-cytokine signature**		**Three-cytokine signature**	
Proteins included	IL-6, G-CSF, IL-15, IL-8, IL-10, IL-17, TNF-alpha, IFN-gamma, IL-18, IP-10		IL-6, G-CSF, IL-15	
	Reference - bacterial	**Reference - viral**	**Reference - bacterial**	**Reference - viral**
Predicted bacterial	28	7	26	7
Predicted viral	10	42	12	42
Accuracy (95% CI)	0.80 (0.71, 0.88)		0.78 (0.68, 0.86)	
Sensitivity	0.74		0.68	
Specificity	0.86		0.86	
Positive predictive value	0.80		0.79	
Negative predictive value	0.81		0.78	

(a) Ten-cytokine



(b) Three-cytokine

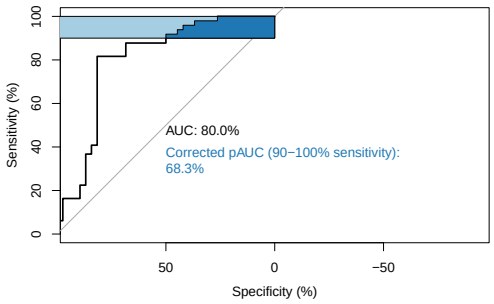


Figure D.4: Receiver operating characteristic curves for two different cytokine models to differentiate between the acute bacterial and acute viral groups, using the training dataset. The area under the curve (AUC) presented for each with a partial AUC at sensitivity range from 90% to 100% also shown.

Model performance - Test dataset

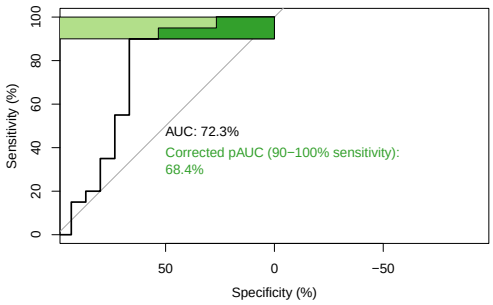
The performances of both models was measured in the test dataset. The numbers of cases correctly predicted as bacterial and viral were the same for both models giving a sensitivity of 67% and a specificity of 85% for both.

The AUC for the ROC curves were AUC of 72.3% for the ten-cytokine model, and 80.7% for the three-cytokine model. Restricting the AUC to the sensitivity range of 90% to 100% resulted in a pAUC of 68.4% for the ten-cytokine model, and a pAUC of 75.4% for the three-cytokine model.

Table D.7: Cytokine model performance statistics in the test data. Comparison of results when a ten-cytokine and a three-cytokine model are used to differentiate between bacterial and viral LRTIs.

Performance metric	**Ten-cytokine signature**		**Three-cytokine signature**	
Proteins included	IL-6, G-CSF, IL-15, IL-8, IL-10, IL-17, TNF-alpha, IFN-gamma, IL-18, IP-10		IL-6, G-CSF, IL-15	
	Reference - bacterial	**Reference - viral**	**Reference - bacterial**	**Reference - viral**
Predicted bacterial	10	3	10	3
Predicted viral	5	17	5	17
Accuracy (95% CI)	0.77 (0.6, 0.9)		0.77 (0.6, 0.9)	
Sensitivity	0.67		0.67	
Specificity	0.85		0.85	
Positive predictive value	0.77		0.77	
Negative predictive value	0.77		0.77	

(a) Ten-cytokine



(b) Three-cytokine

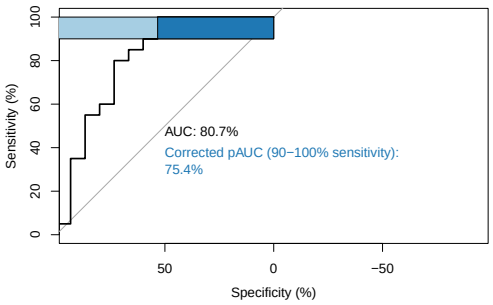


Figure D.5: Receiver operating characteristic curves for two different cytokine models to differentiate between the acute bacterial and acute viral groups, using the test data set. The area under the curve (AUC) presented for each with a partial AUC at sensitivity range from 90% to 100% also shown.

Appendix E

Appendix - Linked to Chapter 5, Multiomics

E.0.1 Supplementary methods

The following steps were following in performing mixOmics analyses in R.

Selection of data

With mixOmics it is advised to have less than 10,000 features per omic platform included. This is not an issue for the proteomic results and cytokine panels but after filtering the transcriptomic results still contained 22,132 genes. Each dataset will be scaled using z-scores to allow for comparison across platforms.

1. Cytokine panel
    (a) Include all 10 cytokines ideally but IL-8 has many missing results as IL-8 failed on two plates and as full data are required, exclude IL-8

and include the other nine cytokines.

(b) Include data after adjusting for batch effect and age.

2. MS Proteomics: Include all 310 proteins that passed the QC steps and well included for the single proteomic level analysis.

3. RNA-seq: As the primary interest in this chapter is to look at comparisons between bacterial and viral datasets, include only the 500 DE genes when bacterial and viral LRTIs were compared.

4. Only samples from the acute bacterial and viral groups, as defined using the semi-supervised approach in Section 2.3.3 were included in these analysis

General steps in R

Packages used include: mixOmics, PCAtools, gplots, ggplot2, RColorBrewer, NMF and dplyr.

Dimension reduction

1. Load the RNA-seq, MS proteomics and cytokine panel datasets into R

2. Parse the data:

   (a) Subset the metadata to only include samples with results across all three platforms.

   (b) Remove CRP results from the MS proteomics data.

   (c) Remove IL-8 results from cytokine panel.

   (d) Subset RNA-seq data to only include 500 genes of interest

3. Split the data into test and train datasets with a 70:30 split

4. Perform PCA analysis, generate PC biplots for each of the three platforms.

5. Perform PCA-DA analysis for each of the three platforms

6. Calculate the error rate for each platform across the first ten components.

7. Generate PCA-DA biplots, examine each of the components that would improve the error rate.

N-integration of datasets

1. Create a design matrix. Perform regression analysis with PLS to calculate the cross-correlation between datasets, using Pearson's correlation coefficient. Use these correlation calculations to add weights to the design matrix.

2. Use block PLS-DA approach to calculate the effect on the error rate of adding components to the model.

   (a) The number of components included in the model which results in the lowest error rate is the optimal number of components.

   (b) Choose the performance measure with the lowest error rate.

3. Variable selection. Decide on the maximum number of features of include for each platform.

   (a) Maximum = 9 cytokines, 20 MS proteomic proteins and 20 RNA transcripts.

(b) A total of 495 models were fitted for each component and each repeated 10 times.

4. Create the final model using the weighted design matrix, the optimal number of components and the optimal number of variables. Calculate how well the model integrates different platforms by calculating Pearson's correlation coefficient across platforms.

5. Create a loadings plot and a heatmap to show how each feature selected contributes to the final model.

6. Test the model performance in the training data, calculate the:

   (a) Sensitivity

   (b) Specificity

   (c) Positive and negative predictive values

   (d) AUC-ROC curve

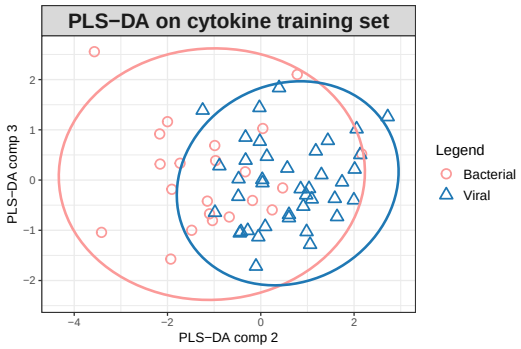
   (e) Partial AUC at a restricted sensitivity range – 90%-100%

7. Test the model performance in the test data, calculate the same performance metrics.

E.0.2 Supplementary results

PLS-DA biplots were plotted for each of the first five components in the cytokine panel results. This was to see if there was any improved differentiation of bacterial and viral groups using different components, see Figures E.1 and E.2.

(a) Components 1 and 3



(b) Components 2 and 3

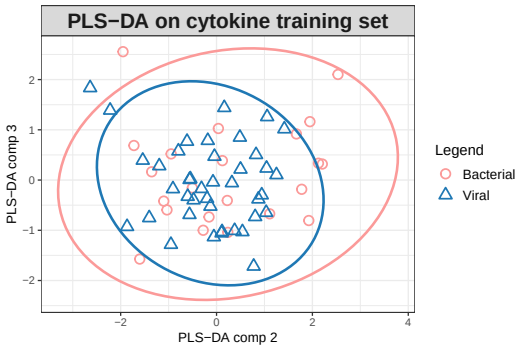
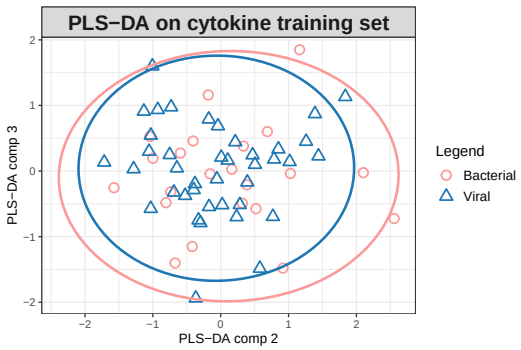


Figure E.1: PLS-DA analysis of MSD cytokine panel results, biplots comparing components 1 and 3, panel (a) and comparing components 2 and 3, panel (b)

(a) Components 3 and 4



(b) Components 4 and 5

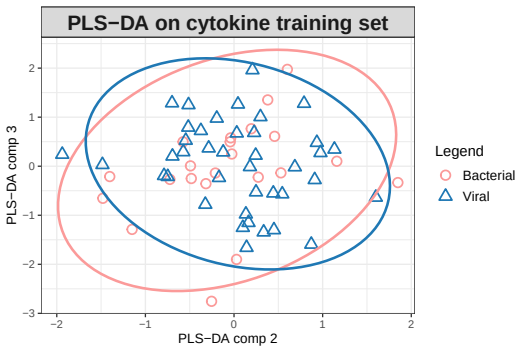


Figure E.2: PLS-DA analysis of MSD cytokine panel results, biplots comparing components 3 and 4, panel (a) and comparing components 4 and 5, panel (b)

Appendix F

Appendix - Linked to Chapter 6, Fever Cohort

F.1 Supplementary methods

Target genes of the FTD Tropical fever CORE RUO kit, information received from Siemens when asked to confirm gene targets used:

- Chikungunya virus - Glycoprotein E1
- Dengue virus - 3´UTR
- *Leptospira species* - LipL32 gene, coding for the major outer membrane lipoprotein
- *Plasmodium species* - 18S rRNA gene
- *Rickettsia species* - Citrate synthase gene (gltA)

- *Salmonella species* - Tetrathionate subunit B (ttrB) gene
- West Nile virus - NS2A region

Target genes of the FTD Dengue differentiation RUO kit:

- Dengue-1 - Polyprotein gene
- Dengue-2 - 3´UTR
- Dengue-3 - 3´UTR
- Dengue-4 - 3´UTR

F.1.1 Nucleic acid extraction and Tropical Fever Core Panel

The laboratory protocol followed for nucleic acid extraction and use of the Tropical Fever Core Panel are outlied here.

Extraction Kit – CWBio Viral DNA/RNA kit (currently in use in at Patan Hospital). Applicable samples: swab, serum, plasma, bronchoalveolar lavage fluid and other cell-free body fluids.

Extraction Kit Components:

1. Lysis buffer
2. Washing buffer 1 (concentrate)
3. Washing buffer 2 (concentrate)

4. RNase-free water
5. Adsorption columns
6. Collection tubes

Consumables needed:

1. Chemicals: Isopropanol, 100% ethanol, RNAse free water
2. Plasticware: 1.5 ml centrifugal tube, pipettes and tips, clear top 0.2ml PCR tubes for detection, two sample racks.
3. Instruments required: Room temperature mixer/shaker, vortex, centrifuge
4. Samples to be extracted with each extraction: Negative control from the PCR kit, samples (selected)

Protocol Steps, Day before extraction of nucleic acid.

1. Select samples and put these samples together in new box and return to -80C freezer.
2. Ensure all of the materials for extraction and PCR are available.
3. Arrange time to do extractions and run PCR with molecular laboratory team.

Day of extractions, preparation for extraction

1. Add isopropanol to Washing Buffer 1 (concentrate) and 100% ethanol to Washing Buffer 2 (concentrate) according to the label of the reagent bottle, then mark them.

2. Mix all the reagents and gently invert 3-5 times before use.

3. Remove the NC from the freezer and leave thaw at room temperature until ready for use.

4. Take the samples to be tested and the internal control (IC) out of the freezer and place in freezer until ready for use.

5. Ensure all of the listed consumables including appropriate pipettes and tips are available.

Note: Please check whether there is crystallization or precipitation in the lysis buffer before use. If this happens, please dissolve it in a 56℃ water bath.

General purification procedure

1. Take 1.5 ml centrifugal tube, add 100 micrloL lysis buffer, 150 microL isopropanol and 100 microL sample (the sample needs to be at room temperature) to tube, (sample: lysis buffer: Isopropanol = 1:1:1.5). Vortex for 5 seconds, then shake by hand for 10 mins.

2. Transfer the solution from Step 1 to the adsorption column of a collection tube. Centrifuge the collection tube at 12,000 rpm (approx. 13,400 g force) for 1 minute and discard the liquid. Then place the adsorption column back into the collection tube.

3. Mix (5 x n+1) microL of isopropanol and (1.5 x n+1) microL IC. Add 6.5 microL of the isopropanol/IC mixture to each tube.

4. Add 500 microL Washing Buffer 1 (confirming that isopropanol has been added) to the adsorption column, centrifuge at 12,000 rpm for 1 minute, discard the liquid and put the adsorption column back into the collection tube.

5. Add 500 microL Washing Buffer 2 (confirming that 100% ethanol has been added) to the adsorption column, centrifuge at 12,000 rpm for 1 minute, discard the liquid and put the adsorption column back into the collection tube.

6. Centrifuge at 12,000 rpm for 2 minutes, and discard waste liquid in the collection tube. Put the adsorption column into a 1.5ml tube and open the column cap (to allow to dry). Place the column at room temperature for 2 minutes and let it dry.

7. Add 40 microL RNase-Free water to the middle part of the adsorption column membrane (Do not touch the membrane), leave at room temperature for 2 minutes and then centrifuge for 2 minutes at 9,000 rpm. If not being used immediately (or within one hour, store in 2-8°C fridge during this time), the nucleic acid solution should be stored at -80°C to prevent degradation. Change tips on each addition of elution buffer/RNase-free water.

Note: Change in centrifuge speed and timing in Step 7 is because the larger centrifuge is needed for this step due to the two lids for the samples not fitting in the smaller centrifuge, speed of centrifuge reduced to ensure caps on tubes don't break (and time of centrifuge increased).

PCR Kit Protocol – FTD Tropical Fever Core (64 reaction kit)

Storage: The components of this PCR kit should be stored in the original packaging at -20°C (-10°C to -30°C).

Notes:

1. Nucleic acid needs to be extracted from the PCR kit negative control, using the same process as above.
2. Positive control does NOT need to be extracted before use.
3. Use extraction methods that yield high-quality and inhibitor-free purified nucleic acids.
4. Do the calculation for the volume of master mix required prior to entering the master mix room.

Topical Fever Core kit components

1. TF1 PP - Primer/probe mix for dengue virus, *Rickettsia spp.*, *Salmonella spp.* West Nile Virus (WNV)
2. TF2 PP - Primer/probe mix for *Plasmodium spp.*, chikungunya virus, *Leptospira spp.* IC (*Streptococcus equi*)
3. TF PC - Positive control (plasmid pool) for the identification of dengue virus, *Rickettsia spp.*, *Salmonella spp.*, WNV, *Plasmodium spp.*, chikungunya virus *Leptospira spp.* To be used with PPmixes 1-2
4. Negative control
5. Internal control

6. 25x RT-PCR Enzyme mix (Fast-track master mix)

7. 2 x RT-PCR Buffer (Fast-track master mix)

PCR Kit Steps

Note: The volume of extracted nucleic acid must have a minimum volume of at least 20 microL. A day before PCR:

1. For each day run, assign a number to participant ID. Have this list made a day before.

2. Prepare and save the PCR programme for each day run

3. Prepare the PCR tubes by labelling with the above assigned number: make sure to not cover the area where the detection is done by PCR machine

4. Next day:

Thaw all reagents and keep in ice. Prepare a PCR master mix. For one reaction, mix: 1.5 microL PP + 1 microL Enzyme + 12.5 microL Buffer.

1. Add one PC and one extracted NC for each day PCR run.

2. Recommended: the 'n' calculated below needs to include extra sample for volume loss in the process of preparation of master mix. If samples number is above 5, n+2; above 10, n+4

3. Preparation of master mix as follows (for both TF1 and TF2): Primer probe(PP) 1.5 microL per reaction, enzyme 1 microL per reaction, buffer 12.5 microL per reaction.

Table F.1: Recommended settings for real-time PCR instruments when using the Siemens Tropical Core Panel. *Excitation wavelengths listed are from Applied Biosystems 7500 Real-Time PCR System - wavelengths may vary for other real-time PCR systems.

PP Mix and Thermocycler Detection Settings				
	Master Mix		Dye	Detection wavelength (nm)*
	TF1	TF2		
Pathogen	DENV	*Plasmodium spp.*	Green	520
	Rickettsia spp.	CHIKV	Yellow	550
	WNV	Leptospira spp.	Orange	610
	Salmonella spp.	IC (*S. equi*)	Red	670

4. Label the side of the 0.2ml PCR tubes with number that has been assigned to the participant ID.

5. Aliquot 15 microL of PCR master mix in each well.

6. Add 10 microL of PC or 10 microL of the extracted NC or 10 microL of extracted samples to the aliquoted PCR master mix. For the recommended settings for real-time PCR instruments refer to Table F.1.

7. Run the PCR programme, see Table F.2 for settings.

8. After completing PCR process, store any remaining extracted RNA/DNA at -80°C.

9. After completing the PCR process, freeze the PCR kit components again at -20°C.

Data acquisition:

1. Export the file in Microsoft Excel or csv format. Save in the relevant folder with today's date in the drive.

Table F.2: Recommended PCR programme when using the Siemens Tropical Core Panel. This programme was used for each of the Tropical Core Panel runs during the experiment presented in Chapter 6

Stage	Cycles	Acquisition	Temperature	Time
Hold	/	/	50°C	15 minutes
Hold	/	/	94°C	1 minute
Cycling	40	/	94°C	8 seconds
		Yes	60°C	1 minute

2. Keep a paper inventory or an Excel inventory for the data so obtained. Information should include, assigned number, participant ID, extraction date, PCR-TF1 date, PCR-TF1-saved file name, PCR-TF2 date, PCR-TF2-saved file name, concentration data.

3. Also save the image file showing the PCR run.

F.2 Severe lower respiratory tract infections

For LRTIs, different severity measures can be used depending on the age and presentation. The WHO criteria for severe pneumonia were described in previous chapters.[18] The BTS definition of severe pneumonia consists of eight features, as shown in Figure F.1.[2] Children presenting to hospital with some of these features can be classified as severe LRTIs and require hospital admission.

For children with bronchiolitis different severity measures are available including the Wood-Downes, Respiratory Severity Score, Bronchiolitis risk of admission score and the ReSVinet score.[332, 333] All of these scores have their advantages and disadvantages, and different scores are used in different settings. I have

Features of severe community acquired pneumonia
Oxygen saturation <92% OR cyanosis
Raised respiratory rate*
Significant tachycardia for level of fever†
Prolonged central capillary refill time (>2 seconds)
Difficulty in breathing
Grunting or intermittent apnoea
Decreased feeding or dehydration
Chronic conditions or immunodeficiency‡

* For children less than 12 months of age, respiratory rate >70 breaths/min. For children one year of age or more, respiratory rate >50 breaths/min
‡ Chronic conditions include: congenital heart disease, chronic lung disease of prematurity, chronic respiratory conditions leading to infection such as cystic fibrosis, bronchiectasis, immune deficiency

†Temperature ≥ 38°C with a heart rate above these age thresholds:

< 1 year of age	≥ 170 beats/min
1-2 years of age	≥ 160 beats/min
2-5 years of age	≥ 150 beats/min
≥ 5 years of age	≥ 130 beats/min

Figure F.1: British Thoracic Society criteria for severe paediatric pneumonia.[2]

selected a score that was suitable for use in this cohort setting.

The ReSVinet score was validated in a study of 170 children under two years of age, admitted to hospital with acute respiratory infections. The ReSVinet score relies on universally available parameters, and was designed so that it could be used by parents as well as physicians. The ReSVinet score has previously been shown to correlate well with length of stay, PICU admission and another commonly used bronchiolitis scale.[333, 334] See Figure F.2 for ReSVinet score parameters.

F.2.1 Methods - Severity

Different severity measures were used based on previous publications. Accurately assigning severity to cases is important when looking for biomarkers associated with severe disease. I looked at the clinical criteria recorded at admission to classify the severity of cases.

For respiratory cases, in participants less than two years of age, the previously validated ReSVinet score was used to assess severity, see Figure F.2. The features required to calculate this score included feeding intolerance, the need for medical intervention, work of breathing, respiratory rate, apnoea, general condition of participant and fever.[333]

Each LRTI case will be assessed to see if it meets the WHO definition of severe pneumonia, as has been done in the previous LRTI cohort study.[18] The BTS criteria for severe pneumonia will also be used to classify severe LRTIs. The outcomes of cases classed as severe using these different definitions will be

Feature	Zero points	One point	Two points	Three points
Feeding intolerance	No	**Mild** Decreased appetite and/or isolated vomits with cough	**Partial** Frequent vomits with cough, able to tolerate fluids sufficiently to ensure hydration	**Total** Oral intolerance or rejection of feed, not able to guarantee adequate hydration orally. Required nasogastric and/or intravenous fluids
Medical intervention	No	**Basic** Nasal secretions aspiration, antipyretics	**Intermediate** Oxygen therapy required. Nebulised therapy with bronchodilators	**High** Required respiratory support with positive pressure(either non-invasive with CPAP, BiPAP or high-flow O2; or invasive through endotracheal tube)
Respiratory difficulty	No	**Mild** No difficulty at rest. Wheezing only audible with stethoscope, good air entrance	**Moderate** Makes some extra respiratory effort (intercostal and/or tracheo-sternal recession). Added sounds audible with stethoscope. Expiratory wheezing audible even without stethoscope.	**Severe** Respiratory effort is obvious. Inspiratory and expiratory wheezing and/or clearly decreased air entry
Respiratory frequency	Normal*	**Occasional or intermittent tachypnoea** Well tolerated, limited in time by self-resolution or response to secretion, aspiration or nebulisation	**Prolonged or recurrent tachypnea** Tachypnea persisted or recurred despite secretion aspiration and/or nebulisation with bronchodilators	**Severe and sustained tachypnoea†** Very superficial and quick breathing rate. Normal/low breath rate with obvious increased respiratory effort and/or mental status affected
Apnoea	No			**Yes** At least one episode of respiratory pause medically documented or strongly suggested through history
General condition	Normal	**Mild** Child is mildly uncomfortable but does not appear to be in a severe distress. Parents are not alarmed. Could wait in the waiting room or be discharged home	**Moderate** Patient looks ill, and will need prompt medical review. Parents are concerned. Cannot wait in the waiting room	**Severe** Agitated, apathetic, lethargic. No need of medical training to realise severity. Parents are very concerned. Immediate medical evaluation and/or intervention are required
Fever	No	**Mild** Central temperature 38°C to 38.4°C	**Yes** Central temperature ≥38.5°C	

Figure F.2: ReSVinet score features for assessing the severity of acute respiratory infections in infants. * Normal respiratory rates: <2 months: 40–50 breaths per minute (bpm), 2–6 months: 35–45 bpm, 6-12 months: 30–40 bpm, 12-24 months: 25–35 bpm, 24-36 months: 20–30 bpm. † Severe tachypnoea: <2 months: >70bpm, 2–6 months: >60 bpm, 6-12 months: >55 bpm, 12-24 months: >50 bpm, 24-36m: >40 bpm.[333]

compared.

Finally, I will look at cases which had poor outcomes, including death, PICU admission, change in disability level at discharge, or prolonged hospital stay. I will look for data points collected at admission which are associated with poor outcomes.

F.2.2 Results - Severity

Identifying cases at admission that are at risk of poor outcomes is important in providing the appropriate level of care. Various tools are available to identify severe infections which might progress to a poor outcome. In the following sections, I describe the participants that went on to have a poor outcome, and different severity measures.

Correctly identifying severe cases is important when looking for biomarkers of severity in future experiments.

Data were collected to define severe disease in the LRTI group. The WHO classification of severe pneumonia classified 18.2% (42/231) of LRTI cases as severe. Using the BTS definition of severe pneumonia, 61.5% (142/231) of cases had one or more features of severe pneumonia, and 13.4% (31/231) of cases had three or more BTS features. The ReSVinet scale was only calculated in children under two years of age. Of the 81 cases with complete data to calculate a ReSVinet score, 24.7% (20/81) had a score of ten or more.

Of the LRTI cases, 13.4% (31/231) had a poor outcome. Using the WHO classification of severe pneumonia, 32.3% (10/31) of the poor outcome cases were

classified as severe at admission. All of the LRTI cases with a poor outcome had at least one of the BTS features of severe pneumonia, with 51.6% (16/31) of poor outcome cases having three or more BTS features.

Of the LRTI cases in children under 2 years of age, 16.9% (20/118) had a poor outcome. Of these participants, 50% (10/20) had a ReSVinet score of ten or more. The ReSVinet score in participants with a poor outcome ranged from 2 to 15.

F.2.3 Discussion - Severity

Three different tools were used to measure LRTI severity at admission. The WHO classification of severe pneumonia performed worst at predicting participants who would go on to have a poor outcome, with a third of poor outcome cases correctly predicted as severe. A ReSVinet score over ten, or the presence of three or more BTS features of severe pneumonia performed better and could identify half of the cases that went on to have a poor outcome.

These results show that there is room for improvement in severity scores, and biomarkers for severity could be useful. Studies of severe infections are needed to address this problem, including studies to identify biomarkers of severity which can be used in conjunction with available severity scoring systems.

Table F.3: Cases of lower respiratory tract infection in the cohort, divided by their diagnostic classification. Demographics, laboratory measures at admission and the proportion of severe cases, using different measures are shown. One case was classified as an uncertain diagnosis and was excluded from this table. * The ReSVinet scale was calculated in children under 2 years of age only, 81 participants had complete data entered to calculate a ReSVinet score. † Severe cases were defined as having three of more features of severe pneumonia using the British Thoracic Society (BTS) definition. CRP, C-reactive protein; WCC, white cell count; WHO, World Health Organisation.

	Definite viral	**Probable viral**	**Viral syndrome - high or no inflammatory markers**	**Unknown**	**Bacterial syndrome - low or no inflammatory markers**	**Probable bacterial**	**Definite bacterial**	**Total cases**
Number of cases	64	22	11	38	53	31	11	230
Age, years (median)	1.214	1.7355	1.859	2.2805	1.859	3.433	3.647	1.955
Female (%)	39.1%	54.5%	36.4%	36.8%	37.7%	48.4%	27.3%	34.3%
Duration of hospital stay, days (median)	6	5	6	6	4	5.5	8	5
CRP (mg/L) - mean	10.0	14.2	44.2	13.0	17.0	85.1	95.0	18.4
WCC x10^9/L	9.2	11.7	17.3	12.9	11.6	14.1	14.9	11.6
Neutrophils x 10^9/L	5.1	6.2	13.1	8.2	5.9	9.2	12.1	6.7
Severe cases - WHO	15.6%	9.1%	36.4%	21.1%	22.6%	19.4%	0.0%	18.2%
Severe cases - BTS*	20.3%	13.6%	9.1%	7.9%	5.7%	19.4%	18.2%	13.4%
Severe cases - ReSVinet scale†	25.6%	7.1%	0.0%	27.8%	6.9%	20.0%	0.0%	24.7%

F.3 Description of samples sent for RNA sequencing

A subset of participants from the cohort were selected for the RNA-seq experiment. Section 6.2.9 describes the criteria for selecting samples for RNA-seq. Table F.4 shows the classification of cases selected for RNA-seq with demographics and blood laboratory results at admission. Most of the cases in the dengue cohort were included in the experiment, 95% (96/101), along with 40.7% (94/231) of the LRTI cases.

At the time of writing these samples have been sent for RNA-seq analysis.

Table F.4: Cases sent for RNA-sequencing, grouped by diagnostic classification. The number of cases with a lower respiratory tract infection (LRTI) diagnosis and a dengue diagnosis are shown. Demographics and blood inflammatory markers at admission are also shown. Two cases were classified as uncertain diagnosis and were excluded from this table. CRP, C-reactive protein; WCC, white cell count.

	Definite viral	Probable viral	Viral syndrome - high or no inflammatory markers	Unknown	Bacterial syndrome - low or no inflammatory markers	Probable bacterial	Definite bacterial	Total cases	Healthy controls
Number of cases	124	11	17	8	23	20	36	241	42
Number of LRTI cases	32	7	6	7	19	14	8	94	NA
Number of dengue cases	82		10				4	96	NA
Age, years (median)	6.33	2.842	3.814	2.816	1.415	3.73	2.5355	3.7	4.1
Female (%)	0.266129	0.454546	0.2352941	0.375	0.304348	0.4	0.555556	33.6	41.5
Duration of hospital stay, days (median)	4	5	5	4	5	6	8	4	NA
CRP (mg/L)	5	16	64.3	4	19	88.7	41.5	22.9	NA
WCC x10^9/L	5.2	10.9	15.1	14.65	9.7	14.2	11.6	8.3	NA
Neutrophils x 10^9/L	2.73	7.93	10.05	9.9445	5.4325	9.16	7.149	4.4	NA

Bibliography

[1] World Health Organisation. *WHO Guidelines Approved by the Guidelines Review Committee. Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control*. World Health Organization Copyright © 2009, World Health Organization., 2009.

[2] M Harris, J Clark, N Coote, P Fletcher, A Harnden, M McKean, and A Thomson. British thoracic society guidelines for the management of community acquired pneumonia in children: update 2011. *Thorax*, 66 Suppl 2:ii1–23, 2011.

[3] CB van Houten, A Cohen, D Engelhard, JP Hays, R Karlsson, E Moore, D Fernández, R Kreisberg, LV Collins, W de Waal, KM de Winter-de Groot, TFW Wolfs, P Meijers, B Luijk, JJ Oosterheert, R Heijligenberg, SUC Sankatsing, AWJ Bossink, A Stubbs, M Stein, S Reisfeld, A Klein, R Rachmilevitch, J Ashkar, I Braverman, V Kartun, I Chistyakov, E Bamberger, I Srugo, M Odeh, E Schiff, Y Dotan, O Boico, R Navon, T Friedman, L Etshtein, M Paz, TM Gottlieb, E Pri-Or, G Kronenfeld, E Simon, K Oved, E Eden, and LJ Bont. Antibiotic misuse in respiratory tract infections in children and adults-a prospective, multicentre study (tailored treatment). *Eur J Clin Microbiol Infect Dis*, 38(3):505–514, 2019.

[4] A Mann, K Nehra, JS Rana, and T Dahiya. Antibiotic resistance in agriculture: Perspectives on upcoming strategies to overcome upsurge in resistance. *Curr Res Microb Sci*, 2:100030, 2021.

[5] A Manesh and GM Varghese. Rising antimicrobial resistance: an evolving epidemic in a pandemic. *The Lancet Microbe*, 2(9):e419–e420, 2021.

[6] CJL Murray, KS Ikuta, F Sharara, L Swetschinski, G Robles Aguilar, A Gray, C Han, C Bisignano, P Rao, E Wool, SC Johnson, AJ Browne, MG Chipeta, F Fell, S Hackett, G Haines-Woodhouse, BH Kashef Hamadani, EAP Kumaran, B McManigal, S Achalapong,

R Agarwal, S Akech, S Albertson, J Amuasi, J Andrews, A Aravkin, E Ashley, F-X Babin, F Bailey, S Baker, B Basnyat, A Bekker, R Bender, JA Berkley, A Bethou, J Bielicki, S Boonkasidecha, J Bukosia, C Carvalheiro, C Castañeda-Orjuela, V Chansamouth, S Chaurasia, S Chiurchiù, F Chowdhury, R Clotaire Donatien, AJ Cook, B Cooper, TR Cressey, E Criollo-Mora, M Cunningham, S Darboe, NPJ Day, M De Luca, K Dokova, A Dramowski, SJ Dunachie, T Duong Bich, T Eckmanns, D Eibach, A Emami, N Feasey, N Fisher-Pearson, K Forrest, C Garcia, D Garrett, P Gastmeier, AZ Giref, RC Greer, V Gupta, S Haller, A Haselbeck, SI Hay, M Holm, S Hopkins, Y Hsia, KC Iregbu, J Jacobs, D Jarovsky, F Javanmardi, AWJ Jenney, M Khorana, S Khusuwan, N Kissoon, E Kobeissi, T Kostyanev, F Krapp, R Krumkamp, A Kumar, HH Kyu, C Lim, K Lim, D Limmathurotsakul, MJ Loftus, M Lunn, J Ma, A Manoharan, F Marks, J May, M Mayxay, N Mturi, T Munera-Huertas, P Musicha, LA Musila, MM Mussi-Pinhata, RN Naidu, T Nakamura, R Nanavati, S Nangia, P Newton, C Ngoun, A Novotney, D Nwakanma, CW Obiero, TJ Ochoa, A Olivas-Martinez, P Olliaro, E Ooko, E Ortiz-Brizuela, P Ounchanum, GD Pak, JL Paredes, AY Peleg, C Perrone, T Phe, K Phommasone, N Plakkal, A Ponce-de Leon, M Raad, T Ramdin, S Rattanavong, A Riddell, T Roberts, JV Robotham, A Roca, VD Rosenthal, KE Rudd, N Russell, HS Sader, W Saengchan, J Schnall, JAG Scott, S Seekaew, M Sharland, M Shivamallappa, J Sifuentes-Osornio, AJ Simpson, N Steenkeste, AJ Stewardson, T Stoeva, N Tasak, A Thaiprakong, G Thwaites, C Tigoi, C Turner, P Turner, HR van Doorn, S Velaphi, A Vongpradith, M Vongsouvath, H Vu, T Walsh, JL Walson, S Waner, T Wangrangsimakul, P Wannapinij, T Wozniak, TEMW Young Sharma, KC Yu, P Zheng, B Sartorius, AD Lopez, A Stergachis, C Moore, C Dolecek, and M Naghavi. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655, 2022.

[7] World Health Organisation. Ten threats to global health in 2019, 2019.

[8] UNICEF, WHO, IBRD, UN Population Division, and UN Inter-Agency Group for Child Mortality Estimation. Levels trends in child mortality, report 2020, estimates. developed by the un inter-agency group for child mortality estimation. Report, 2020.

[9] World Health Organisation. Child mortality and causes of death, 2024.

[10] F Villavicencio, J Perin, H Eilerts-Spinelli, D Yeung, D Prieto-Merino, L Hug, D Sharrow, D You, KL Strong, RE Black, and L Liu. Global, regional, and national causes of death in children and adolescents younger

than 20 years: an open data portal with estimates for 2000x2013;21. *The Lancet Global Health*, 12(1):e16–e17, 2024.

[11] RM Cunningham, MA Walton, and PM Carter. The major causes of death in children and adolescents in the united states. *New England Journal of Medicine*, 379(25):2468–2475, 2018.

[12] C Grüber, T Keil, M Kulig, S Roll, U Wahn, and V Wahn. History of respiratory infections in the first 12 yr among children from a birth cohort. *Pediatr Allergy Immunol*, 19(6):505–12, 2008.

[13] Thomas M and Bomar PA. *Upper Respiratory Tract Infection*. StatPearls, Treasure Island, 2024.

[14] D Tristram. *Laryngitis, Tracheitis, Epiglottitis, and Bronchiolitis*. Springer International Publishing, 2019.

[15] KL Hon, AKC Leung, AHC Wong, A Dudi, and KKY Leung. Respiratory syncytial virus is the most common causative agent of viral bronchiolitis in young children: An updated review. *Curr Pediatr Rev*, 19(2):139–149, 2023.

[16] HMG Wiegers, L van Nijen, JBM van Woensel, RA Bem, MD de Jong, and JCJ Calis. Bacterial co-infection of the respiratory tract in ventilated children with bronchiolitis; a retrospective cohort study. *BMC Infect Dis*, 19(1):938, 2019.

[17] KL O'Brien, HC Baggett, WA Brooks, DR Feikin, LL Hammitt, MM Higdon, SRC Howie, M Deloria Knoll, KL Kotloff, OS Levine, SA Madhi, DR Murdoch, C Prosperi, JAG Scott, Q Shi, DM Thea, Z Wu, SL Zeger, PV Adrian, P Akarasewi, TP Anderson, M Antonio, JO Awori, VL Baillie, C Bunthi, J Chipeta, MJ Chisti, J Crawley, AN DeLuca, AJ Driscoll, BE Ebruke, HP Endtz, N Fancourt, W Fu, D Goswami, MJ Groome, M Haddix, L Hossain, Y Jahan, EW Kagucia, A Kamau, RA Karron, S Kazungu, N Kourouma, L Kuwanda, G Kwenda, M Li, EM Machuka, G Mackenzie, N Mahomed, SA Maloney, JL McLellan, JL Mitchell, DP Moore, SC Morpeth, A Mudau, L Mwananyanda, J Mwansa, M Silaba Ominde, U Onwuchekwa, DE Park, J Rhodes, P Sawatwong, P Seidenberg, A Shamsul, EAF Simões, S Sissoko, S Wa Somwe, SO Sow, M Sylla, B Tamboura, MD Tapia, S Thamthitiwat, A Toure, NL Watson, K Zaman, and SMA Zaman. Causes of severe pneumonia requiring hospital admission in children without hiv infection from africa and asia: the perch multi-country case-control study. *The Lancet*, 394(10200):757–779, 2019.

[18] World Health Organisation. *WHO Guidelines Approved by the Guidelines Review Committee*. World Health Organization Copyright © World Health Organization 2014., 2014.

[19] ML Everard. Paediatric respiratory infections. *European Respiratory Review*, 25(139):36, 2016.

[20] GL Collaborators. Age-sex differences in the global burden of lower respiratory infections and risk factors, 1990-2019: results from the global burden of disease study 2019. *Lancet Infect Dis*, 22(11):1626–1647, 2022.

[21] S Safiri, A Mahmoodpoor, AA Kolahi, SA Nejadghaderi, MJM Sullman, MA Mansournia, K Ansarin, GS Collins, JS Kaufman, and M Abdollahi. Global burden of lower respiratory infections during the last three decades. *Front Public Health*, 10:1028525, 2022.

[22] S Berman. Epidemiology of acute respiratory infections in children of developing countries. *Reviews of Infectious Diseases*, 13(Supplement$_6$) : $S454 - -S462, 1991$.

[23] CQ Fritz, KM Edwards, WH Self, CG Grijalva, Y Zhu, SR Arnold, JA McCullers, K Ampofo, AT Pavia, RG Wunderink, EJ Anderson, AM Bramley, S Jain, and DJ Williams. Prevalence, risk factors, and outcomes of bacteremic pneumonia in children. *Pediatrics*, 144(1), 2019.

[24] TJ Evans and PA Riley. Principles of microscopy, culture and serology-based diagnostics. *Medicine*, 49(10):648–653, 2021.

[25] I Rudan, C Boschi-Pinto, Z Biloglav, K Mulholland, and H Campbell. Epidemiology and etiology of childhood pneumonia. *Bull World Health Organ*, 86(5):408–16, 2008.

[26] JN Oliwa and BJ Marais. Vaccines to prevent pneumonia in children - a developing country perspective. *Paediatr Respir Rev*, 22:23–30, 2017.

[27] ZA Bhutta, JK Das, N Walker, A Rizvi, H Campbell, I Rudan, and RE Black. Interventions to address deaths from childhood pneumonia and diarrhoea equitably: what works and at what cost? *Lancet*, 381(9875):1417–1429, 2013.

[28] O Ruuskanen, E Lahti, LC Jennings, and DR Murdoch. Viral pneumonia. *Lancet*, 377(9773):1264–75, 2011.

[29] T Shi, K McLean, H Campbell, and H Nair. Aetiological role of common respiratory viruses in acute lower respiratory infections in children under five years: A systematic review and meta-analysis. *J Glob Health*, 5(1):010408, 2015.

[30] B Kampmann, SA Madhi, I Munjal, EAF Simões, BA Pahud, C Llapur, J Baker, G Pérez Marc, D Radley, E Shittu, J Glanternik, H Snaggs, J Baber, P Zachariah, SL Barnabas, M Fausett, T Adam, N Perreras, MA Van Houten, A Kantele, LM Huang, LJ Bont, T Otsuki, SL Vargas, J Gullam, B Tapiero, RT Stein, FP Polack, HJ Zar, NB Staerke, M Duron Padilla, PC Richmond, K Koury, K Schneider, EV Kalinina, D Cooper, KU Jansen, AS Anderson, KA Swanson, WC Gruber, and A Gurtman. Bivalent prefusion f vaccine in pregnancy to prevent rsv illness in infants. *N Engl J Med*, 388(16):1451–1464, 2023.

[31] EE Walsh, G Pérez Marc, AM Zareba, AR Falsey, Q Jiang, M Patton, FP Polack, C Llapur, PA Doreski, K Ilangovan, M Rämet, Y Fukushima, N Hussen, LJ Bont, J Cardona, E DeHaan, G Castillo Villa, M Ingilizova, D Eiras, T Mikati, RN Shah, K Schneider, D Cooper, K Koury, MM Lino, AS Anderson, KU Jansen, KA Swanson, A Gurtman, WC Gruber, and B Schmoele-Thoma. Efficacy and safety of a bivalent rsv prefusion f vaccine in older adults. *N Engl J Med*, 388(16):1465–1477, 2023.

[32] GL Collaborators. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990-2016: a systematic analysis for the global burden of disease study 2016. *Lancet Infect Dis*, 18(11):1191–1210, 2018.

[33] DR Murdoch, CW Woods, MD Zimmerman, PM Dull, RH Belbase, AJ Keenan, RM Scott, B Basnyat, LK Archibald, and LB Reller. The etiology of febrile illness in adults presenting to patan hospital in kathmandu, nepal. *Am J Trop Med Hyg*, 70(6):670–5, 2004.

[34] JE Meiring, M Shakya, F Khanam, M Voysey, MT Phillips, S Tonks, D Thindwa, TC Darton, S Dongol, A Karkey, K Zaman, S Baker, C Dolecek, SJ Dunstan, G Dougan, KE Holt, RS Heyderman, F Qadri, VE Pitzer, B Basnyat, MA Gordon, J Clemens, and AJ Pollard. Burden of enteric fever at three urban sites in africa and asia: a multicentre population-based study. *Lancet Glob Health*, 9(12):e1688–e1696, 2021.

[35] KR Rijal, B Adhikari, B Ghimire, B Dhungel, UR Pyakurel, P Shah, A Bastola, B Lekhak, MR Banjara, BD Pandey, DM Parker, and P Ghimire. Epidemiology of dengue virus infections in nepal, 2006-2019. *Infect Dis Poverty*, 10(1):52, 2021.

[36] N Kandel, GD Thakur, and A Andjaparidze. Leptospirosis in nepal. *JNMA J Nepal Med Assoc*, 52(187):151–3, 2012.

[37] E Bottieau, L Van Duffel, S El Safi, KD Koirala, B Khanal, S Rijal, NR Bhattarai, T Phe, K Lim, D Mukendi, JL Kalo, P Lutumba, B Barbé, J Jacobs,

M Van Esbroeck, N Foqué, A Tsoumanis, P Parola, CP Yansouni, M Boelaert, K Verdonck, and F Chappuis. Etiological spectrum of persistent fever in the tropics and predictors of ubiquitous infections: a prospective four-country study with pooled analysis. *BMC Med*, 20(1):144, 2022.

[38] KD Koirala, F Chappuis, K Verdonck, S Rijal, and M Boelaert. Persistent febrile illnesses in nepal: A systematic review. *Indian J Med Res*, 148(4):385–395, 2018.

[39] P Lamichhane, KM Pokhrel, B Alghalyini, ARZ Zaidi, MZ Alshehery, K Khanal, M Bhattarai, and A Yadav. Epidemiology, clinical characteristics, diagnosis, and complications of scrub typhus infection in nepal: a systematic review. *Ann Med Surg (Lond)*, 85(10):5022–5030, 2023.

[40] CE Smith. The history of dengue in tropical asia and its probable relationship to the mosquito aedes aegypti. *J Trop Med Hyg*, 59(10):243–51, 1956.

[41] X Yang, MBM Quam, T Zhang, and S Sang. Global burden for dengue and the evolving pattern in the past 30 years. *J Travel Med*, 28(8), 2021.

[42] S Bhatt, PW Gething, OJ Brady, JP Messina, AW Farlow, CL Moyes, JM Drake, JS Brownstein, AG Hoen, O Sankoh, MF Myers, DB George, T Jaenisch, GR Wint, CP Simmons, TW Scott, JJ Farrar, and SI Hay. The global distribution and burden of dengue. *Nature*, 496(7446):504–7, 2013.

[43] AT Mairuhu, J Wagenaar, DP Brandjes, and EC van Gorp. Dengue: an arthropod-borne disease of global importance. *Eur J Clin Microbiol Infect Dis*, 23(6):425–33, 2004.

[44] MG Guzman, SB Halstead, H Artsob, P Buchy, J Farrar, DJ Gubler, E Hunsperger, A Kroeger, HS Margolis, E Martínez, MB Nathan, JL Pelegrino, C Simmons, S Yoksan, and RW Peeling. Dengue: a continuing global threat. *Nat Rev Microbiol*, 8(12 Suppl):S7–16, 2010.

[45] SB Halstead. Etiologies of the experimental dengues of siler and simmons. *Am J Trop Med Hyg*, 23(5):974–82, 1974.

[46] DA Muller, AC Depelsenaire, and PR Young. Clinical and laboratory diagnosis of dengue virus infection. *J Infect Dis*, $215(\text{suppl}_2) : S89 - -s95, 2017$.

[47] EG Kallás, MA Cintra, JA Moreira, EG Patiño, PE Braga, JC Tenório, V Infante, R Palacios, MVG de Lacerda, and D Batista Pereira. Live, attenuated, tetravalent butantan–dengue vaccine in children and adults. *New England Journal of Medicine*, 390(5):397–408, 2024.

[48] S Sridhar, A Luedtke, E Langevin, M Zhu, M Bonaparte, T Machabert, S Savarino, B Zambrano, A Moureau, A Khromava, Z Moodie, T Westling, C Mascareñas, C Frago, M Cortés, D Chansinghakul, F Noriega, A Bouckenooghe, J Chen, SP Ng, PB Gilbert, S Gurunathan, and CA DiazGranados. Effect of dengue serostatus on dengue vaccine safety and efficacy. *N Engl J Med*, 379(4):327–340, 2018.

[49] S Biswal, H Reynales, X Saez-Llorens, P Lopez, C Borja-Tabora, P Kosalaraksa, C Sirivichayakul, V Watanaveeradej, L Rivera, F Espinoza, L Fernando, R Dietze, K Luz, R Venâncio da Cunha, J Jimeno, E López-Medina, A Borkowski, M Brose, M Rauscher, I LeFevre, S Bizjajeva, L Bravo, and D Wallace. Efficacy of a tetravalent dengue vaccine in healthy children and adolescents. *N Engl J Med*, 381(21):2009–2019, 2019.

[50] SB Halstead. Three dengue vaccines — what now? *New England Journal of Medicine*, 390(5):464–465, 2024.

[51] B Amatya, E Schwartz, A Biber, O Erster, Y Lustig, R Pradhan, B Khadka, and P Pandey. Dengue serotype characterization during the 2022 dengue epidemic in kathmandu, nepal. *Journal of Travel Medicine*, 30(7):taad034, 2023.

[52] S Malla, GD Thakur, SK Shrestha, MK Banjeree, LB Thapa, G Gongal, P Ghimire, BP Upadhyay, P Gautam, S Khanal, A Nisaluk, RG Jarman, and RV Gibbons. Identification of all dengue serotypes in nepal. *Emerg Infect Dis*, 14(10):1669–70, 2008.

[53] P Bhatt, SP Sabeena, M Varma, and G Arunkumar. Current understanding of the pathogenesis of dengue virus infection. *Curr Microbiol*, 78(1):17–32, 2021.

[54] S Rhedin, A Lindstrand, A Hjelmgren, M Ryd-Rinder, L Öhrmalm, T Tolfvenstam, Å Örtqvist, M Rotzén-Östlund, B Zweygberg-Wirgart, B Henriques-Normark, K Broliden, and P Naucler. Respiratory viruses associated with community-acquired pneumonia in children: matched case-control study. *Thorax*, 70(9):847–53, 2015.

[55] N Salez, A Vabret, M Leruez-Ville, L Andreoletti, F Carrat, F Renois, and X de Lamballerie. Evaluation of four commercial multiplex molecular tests for the diagnosis of acute respiratory infections. *PLOS ONE*, 10(6):e0130378, 2015.

[56] MJ Carter, M Gurung, B Pokhrel, SM Bijukchhe, S Karmacharya, B Khadka, A Maharjan, S Bhattarai, S Shrestha, B Khadka, A Khulal, S Gurung, B Dhital, KG Prajapati, I Ansari, GP Shah, B Wahl, R Kandasamy, R Pradhan, S Kelly, M Voysey, DR Murdoch, N Adhikari, S Thorson, D Kelly, S Shrestha, and

AJ Pollard. Childhood invasive bacterial disease in kathmandu, nepal (2005-2013). *Pediatr Infect Dis J*, 41(3):192–198, 2022.

[57] AS Shah, MD Knoll, PR Sharma, JC Moisi, P Kulkarni, MK Lalitha, M Steinhoff, and K Thomas. Invasive pneumococcal disease in kanti children's hospital, nepal, as observed by the south asian pneumococcal alliance network. *Clin Infect Dis*, 48 Suppl 2:S123–8, 2009.

[58] JA Scott, AJ Hall, R Dagan, JM Dixon, SJ Eykyn, A Fenoll, M Hortal, LP Jetté, JH Jorgensen, F Lamothe, C Latorre, JT Macfarlane, DM Shlaes, LE Smart, and A Taunay. Serogroup-specific epidemiology of streptococcus pneumoniae: associations with age, sex, and geography in 7,000 episodes of invasive disease. *Clin Infect Dis*, 22(6):973–81, 1996.

[59] JS Bradley, CL Byington, SS Shah, B Alverson, ER Carter, C Harrison, SL Kaplan, SE Mace, Jr. McCracken, GH, MR Moore, SD St Peter, JA Stockwell, and JT Swanson. The management of community-acquired pneumonia in infants and children older than 3 months of age: clinical practice guidelines by the pediatric infectious diseases society and the infectious diseases society of america. *Clin Infect Dis*, 53(7):e25–76, 2011.

[60] JW Paisley, BA Lauer, K McIntosh, MP Glode, J Schachter, and C Rumack. Pathogens associated with acute lower respiratory tract infection in young children. *Pediatr Infect Dis*, 3(1):14–9, 1984.

[61] E Vuori-Holopainen, E Salo, H Saxén, K Hedman, T Hyypiä, R Lahdenperä, M Leinonen, E Tarkka, M Vaara, and H Peltola. Etiological diagnosis of childhood pneumonia by use of transthoracic needle aspiration and modern microbiological methods. *Clin Infect Dis*, 34(5):583–90, 2002.

[62] PH Patel, MH Antoine, and S Ullah. *Bronchoalveolar Lavage*. StatPearls Publishing Copyright © 2024, StatPearls Publishing LLC., 2024.

[63] AG Falade, EK Mulholland, RA Adegbola, and BM Greenwood. Bacterial isolates from blood and lung aspirate cultures in gambian children with lobar pneumonia. *Ann Trop Paediatr*, 17(4):315–9, 1997.

[64] NICE. *National Institute for Health and Care Excellence: Guidelines*. National Institute for Health and Care Excellence (NICE) Copyright © NICE 2021., 2021.

[65] P Toikka, K Irjala, T Juvén, R Virkki, J Mertsola, M Leinonen, and O Ruuskanen. Serum procalcitonin, c-reactive protein and interleukin-6 for distinguishing bacterial and viral pneumonia in children. *The Pediatric infectious disease journal*, 19(7):598–602, 2000.

[66] R Virkki, T Juven, H Rikalainen, E Svedström, J Mertsola, and O Ruuskanen. Differentiation of bacterial and viral pneumonia in children. *Thorax*, 57(5):438–41, 2002.

[67] FS Dawood, SS Chaves, A Pérez, A Reingold, J Meek, MM Farley, P Ryan, R Lynfield, C Morin, J Baumbach, NM Bennett, S Zansky, A Thomas, ML Lindegren, W Schaffner, and L Finelli. Complications and associated bacterial coinfections among children hospitalized with seasonal or pandemic influenza, united states, 2003-2010. *J Infect Dis*, 209(5):686–94, 2014.

[68] K Ampofo, J Bender, X Sheng, K Korgenski, J Daly, AT Pavia, and CL Byington. Seasonal invasive pneumococcal disease in children: role of preceding respiratory viral infection. *Pediatrics*, 122(2):229–37, 2008.

[69] DM Weinberger, KP Klugman, CA Steiner, L Simonsen, and C Viboud. Association between respiratory syncytial virus activity and pneumococcal disease in infants: A time series analysis of us hospitalization data. *PLOS Medicine*, 12(1):e1001776, 2015.

[70] GA Capraro. Replacement of culture with molecular testing for diagnosis infectious diseases. *Advances in Molecular Pathology*, 1(1):91–96, 2018.

[71] A Szlachta-McGinn, KM Douglass, UYR Chung, NJ Jackson, JC Nickel, and AL Ackerman. Molecular diagnostic methods versus conventional urine culture for diagnosis and treatment of urinary tract infection: A systematic review and meta-analysis. *Eur Urol Open Sci*, 44:113–124, 2022.

[72] KA Fleming, S Horton, ML Wilson, R Atun, K DeStigter, J Flanigan, S Sayed, P Adam, B Aguilar, S Andronikou, C Boehme, W Cherniak, AN Cheung, B Dahn, L Donoso-Bach, T Douglas, P Garcia, S Hussain, HS Iyer, M Kohli, AB Labrique, LM Looi, JG Meara, J Nkengasong, M Pai, KL Pool, K Ramaiya, L Schroeder, D Shah, R Sullivan, BS Tan, and K Walia. The lancet commission on diagnostics: transforming access to diagnostics. *Lancet*, 398(10315):1997–2050, 2021.

[73] World Health Organisation. Building laboratory capacity for diagnostic testing and sequencing of covid-19 in nepal, 2024.

[74] P Ramanan, AL Bryson, MJ Binnicker, BS Pritt, and R Patel. Syndromic panel-based testing in clinical microbiology. *Clin Microbiol Rev*, 31(1), 2018.

[75] A Calderaro, M Buttrini, B Farina, S Montecchini, F De Conto, and C Chezzi. Respiratory tract infections and laboratory diagnostic methods: A review with a focus on syndromic panel-based assays. *Microorganisms*, 10(9), 2022.

[76] P Shah, M Voice, L Calvo-Bado, I Rivero-Calle, S Morris, R Nijman, C Broderick, T De, I Eleftheriou, R Galassini, A Khanijau, L Kolberg, M Kolnik, A Rudzate, MG Sagmeister, NA Schweintzger, F Secka, C Thakker, F van der Velden, C Vermont, K Vincek, PKA Agyeman, AJ Cunnington, R De Groot, M Emonts, K Fidler, TW Kuijpers, M Mommert-Tripon, K Brengel-Pesce, F Mallet, H Moll, S Paulus, M Pokorn, A Pollard, LJ Schlapbach, CF Shen, M Tsolia, E Usuf, M van der Flier, U von Both, S Yeung, D Zavadska, W Zenz, V Wright, ED Carrol, M Kaforou, F Martinon-Torres, C Fink, M Levin, and J Herberg. Relationship between molecular pathogen detection and clinical disease in febrile children across europe: a multicentre, prospective observational study. *Lancet Reg Health Eur*, 32:100682, 2023.

[77] M Largman-Chalamish, A Wasserman, A Silberman, T Levinson, O Ritter, S Berliner, D Zeltser, I Shapira, O Rogowski, and S Shenhar-Tsarfaty. Differentiating between bacterial and viral infections by estimated crp velocity. *PLoS One*, 17(12):e0277401, 2022.

[78] MU Bhuiyan, CC Blyth, R West, J Lang, T Rahman, C Granland, C de Gier, ML Borland, RB Thornton, LS Kirkham, A Martin, PC Richmond, DW Smith, A Jaffe, and TL Snelling. Combination of clinical symptoms and blood biomarkers can improve discrimination between bacterial or viral community-acquired pneumonia in children. *BMC Pulm Med*, 19(1):71, 2019.

[79] B Müller, KL Becker, H Schächinger, PR Rickenbacher, PR Huber, W Zimmerli, and R Ritz. Calcitonin precursors are reliable markers of sepsis in a medical intensive care unit. *Crit Care Med*, 28(4):977–83, 2000.

[80] L Hu, Q Shi, M Shi, R Liu, and C Wang. Diagnostic value of pct and crp for detecting serious bacterial infections in patients with fever of unknown origin: A systematic review and meta-analysis. *Applied Immunohistochemistry Molecular Morphology*, 25(8), 2017.

[81] X Yang, J Zeng, X Yu, Z Wang, D Wang, Q Zhou, T Bai, and Y Xu. Pct, il-6, and il-10 facilitate early diagnosis and pathogen classifications in bloodstream infection. *Annals of Clinical Microbiology and Antimicrobials*, 22(1):103, 2023.

[82] E Eden, I Srugo, T Gottlieb, R Navon, O Boico, A Cohen, E Bamberger, A Klein, and K Oved. Diagnostic accuracy of a trail, ip-10 and crp combination for discriminating bacterial and viral etiologies at the emergency department. *Journal of Infection*, 73(2):177–180, 2016.

[83] C Papan, A Argentiero, M Porwoll, U Hakim, E Farinelli, I Testa, MB Pasticci, D Mezzetti, K Perruccio, L Etshtein, N Mastboim, E Moscoviz, TI Ber, A Cohen,

E Simon, O Boico, L Shani, TM Gottlieb, R Navon, E Barash, K Oved, E Eden, A Simon, JG Liese, M Knuf, M Stein, R Yacobov, E Bamberger, S Schneider, S Esposito, and T Tenenbaum. A host signature based on trail, ip-10, and crp for reducing antibiotic overuse in children by differentiating bacterial from viral infections: a prospective, multicentre cohort study. *Clin Microbiol Infect*, 28(5):723–730, 2022.

[84] SA Tegethoff, G Danziger, D Kühn, C Kimmer, T Adams, L Heintz, C Metz, K Reifenrath, R Angresius, S Mang, T Rixecker, A Becker, J Geisel, C Jentgen, F Seiler, MC Reichert, F Fröhlich, S Meyer, J Rissland, S Ewen, G Wagenpfeil, K Last, S Smola, R Bals, F Lammert, SL Becker, M Krawczyk, PM Lepper, and C Papan. Tnf-related apoptosis-inducing ligand, interferon gamma-induced protein 10, and c-reactive protein in predicting the progression of sars-cov-2 infection: a prospective cohort study. *Int J Infect Dis*, 122:178–187, 2022.

[85] S Dittrich, BT Tadesse, F Moussy, A Chua, A Zorzet, T Tängdén, DL Dolinger, AL Page, JA Crump, V D'Acremont, Q Bassat, Y Lubell, PN Newton, N Heinrich, TJ Rodwell, and IJ González. Target product profile for a diagnostic assay to differentiate between bacterial and non-bacterial infections and reduce antimicrobial overuse in resource-limited settings: An expert consensus. *PLoS One*, 11(8):e0161721, 2016.

[86] Encyclopaedia Britannica. Geography and travel - kathmandu, 2024.

[87] MB Khan, ZS Yang, CY Lin, MC Hsu, AN Urbina, W Assavalapsakul, WH Wang, YH Chen, and SF Wang. Dengue overview: An updated systemic review. *J Infect Public Health*, 16(10):1625–1642, 2023.

[88] CollaboratorsGBD. The global burden of typhoid and paratyphoid fevers: a systematic analysis for the global burden of disease study 2017. *Lancet Infect Dis*, 19(4):369–381, 2019.

[89] DO Garrett, AT Longley, K Aiemjoy, MT Yousafzai, C Hemlock, AT Yu, K Vaidya, D Tamrakar, S Saha, II Bogoch, K Date, S Saha, MS Islam, KMI Sayeed, C Bern, S Shakoor, IF Dehraj, J Mehmood, MSI Sajib, M Islam, RS Thobani, A Hotwani, N Rahman, S Irfan, SR Naga, AM Memon, S Pradhan, K Iqbal, R Shrestha, H Rahman, MM Hasan, SH Qazi, AM Kazi, NS Saddal, R Jamal, MJ Hunzai, T Hossain, F Marks, AS Carter, JC Seidman, FN Qamar, SK Saha, JR Andrews, and SP Luby. Incidence of typhoid and paratyphoid fever in bangladesh, nepal, and pakistan: results of the surveillance for enteric fever in asia project. *The Lancet Global Health*, 10(7):e978–e988, 2022.

[90] CM Parry, L Wijedoru, A Arjyal, and S Baker. The utility of diagnostic tests for enteric fever in endemic locations. *Expert Review of Anti-infective Therapy*, 9(6):711–725, 2011.

[91] DP Neupane, HP Dulal, and J Song. Enteric fever diagnosis: Current challenges and future directions. *Pathogens*, 10(4), 2021.

[92] WJ Terpstra, WH Organization, and IL Society. *Human Leptospirosis: Guidance for Diagnosis, Surveillance and Control*. World Health Organization, 2003.

[93] D Kala, S Gupta, R Nagraik, V Verma, A Thakur, and A Kaushal. Diagnosis of scrub typhus: recent advancements and challenges. *3 Biotech*, 10(9):396, 2020.

[94] AG Stewart and AGA Stewart. An update on the laboratory diagnosis of rickettsia spp. infection. *Pathogens*, 10(10), 2021.

[95] R Chaudhary, S Bhatta, A Singh, M Pradhan, B Moktan, S Duwal, and R Pandit. A comparative study of rapid sars-cov-2 antigen detection assay against rt-pcr assay for diagnosis of covid-19 in a tertiary hospital of kathmandu. *Kathmandu Univ Med J (KUMJ)*, 20(79):337–341, 2022.

[96] C Chaimayo, B Kaewnaphan, N Tanlieng, N Athipanyasilp, R Sirijatuphat, M Chayakulkeeree, N Angkasekwinai, R Sutthent, N Puangpunngam, T Tharmviboonsri, O Pongraweewan, S Chuthapisith, Y Sirivatanauksorn, W Kantakamalakul, and N Horthongkham. Rapid sars-cov-2 antigen detection assay in comparison with real-time rt-pcr assay for laboratory diagnosis of covid-19 in thailand. *Virol J*, 17(1):177, 2020.

[97] ML Levin, AN Snellgrove, and GE Zemtsova. Comparative value of blood and skin samples for diagnosis of spotted fever group rickettsial infection in model animals. *Ticks Tick Borne Dis*, 7(5):1029–1034, 2016.

[98] W Watthanaworawit, P Turner, C Turner, A Tanganuchitcharnchai, AL Richards, KM Bourzac, SD Blacksell, and F Nosten. A prospective evaluation of real-time pcr assays for the detection of orientia tsutsugamushi and rickettsia spp. for early diagnosis of rickettsial infections during the acute phase of undifferentiated febrile illness. *Am J Trop Med Hyg*, 89(2):308–310, 2013.

[99] C de Abreu Fonseca, VL Teixeira de Freitas, E Caló Romero, C Spinosa, MC Arroyo Sanches, MV da Silva, and MA Shikanai-Yasuda. Polymerase chain reaction in comparison with serological tests for early diagnosis of human leptospirosis. *Trop Med Int Health*, 11(11):1699–707, 2006.

[100] P Mahajan, N Kuppermann, A Mejias, N Suarez, D Chaussabel, TC Casper, B Smith, ER Alpern, J Anders, SM Atabaki, JE Bennett, S Blumberg, B Bonsu, D Borgialli, A Brayer, L Browne, DM Cohen, EF Crain, AT Cruz, PS Dayan, R Gattu, R Greenberg, Jr. Hoyle, JD, DM Jaffe, DA Levine, K Lillis, JG Linakis, J Muenzer, LE Nigrovic, EC Powell, AJ Rogers, G Roosevelt, RM Ruddy, M Saunders, MG Tunik, L Tzimenatos, M Vitale, JM Dean, and O Ramilo. Association of rna biosignatures with bacterial infections in febrile infants aged 60 days or younger. *Jama*, 316(8):846–57, 2016.

[101] HR Jackson, J Zandstra, S Menikou, MS Hamilton, AJ McArdle, R Fischer, AM Thorne, H Huang, MW Tanck, MH Jansen, T De, PKA Agyeman, U Von Both, ED Carrol, M Emonts, I Eleftheriou, M Van der Flier, C Fink, J Gloerich, R De Groot, HA Moll, M Pokorn, AJ Pollard, LJ Schlapbach, MN Tsolia, E Usuf, VJ Wright, S Yeung, D Zavadska, W Zenz, LJM Coin, C Casals-Pascual, AJ Cunnington, F Martinon-Torres, JA Herberg, MI de Jonge, M Levin, TW Kuijpers, M Kaforou, HR Jackson, J Zandstra, S Menikou, S Hamilton, AJ McArdle, T De, PKA Agyeman, U Von Both, ED Carrol, M Emonts, I Eleftheriou, M Van der Flier, C Fink, R De Groot, HA Moll, M Pokorn, A Pollard, LJ Schlapbach, M Tsolia, E Usuf, V Wright, S Yeung, D Zavadska, W Zenz, LJM Coin, AJ Cunnington, F Martinon-Torres, J Herberg, MI De Jonge, M Levin, T Kuijpers, M Kaforou, A Abdulla, C Aebi, R Agbeko, L Ali, W Alkema, K Allen, S Anderson, I Ansari, T Arif, T Avramoska, B Baas, N Bahovec, A Balode, A Bārdzdina, AM Barendregt, R Barral-Arca, D Bath, S Bauchinger, L Baumard, H Baumgart, F Baxter, K Bell, A Bell, X Bello, E Bellos, M Benesch, J Bennet, C Berger, S Bernhard-Stirnemann, S Bibi, C Bidlingmaier, A Binder, V Binder, J Blackmore, K Bojang, DM Borensztajn, K Brengel-Pesce, C Broderick, J Buschbeck, L Calvo-Bado, S Carnota, MJ Carter, MB Castro, M Cebey-López, S Ceesay, A Ceolotto, A Chan, E Cocklin, K Collings, S Crulley, MJ Curras-Tuala, U D'Alessandro, G D'Souza, K Danhauser, S Darboe, S Darnell, L De Haan, G De Vries, D Deksne, K Devine, JE Dewez, W Dik, J Dudley, E Eber, D Fabian, CB Farto, SS Fernández, K Fidler, E Fitchett, R Galassini, S Gallisti, MB García, D Gardovska, J Geissler, GPJM Gerrits, E Giannoni, J Gloerich, A Gómez-Carballa, FÁ González, G Gores, D Grāvele, M Griese, I Grope, M Gurung, N Haas, D Habgood-Coote, NN Hagedoorn, H Haidl, R Harrison, A Hauer, J Heidema, U Heininger, S Henriet, M Hibberd, C Hoggart, S Hösele, S Hourmat, C Hude, M Huijnen, PL Iglesias, MV Iglesias, R Jennings, J Johnson, I Jongerius, R Jorgensen, C Kahlert, R Kandasamy, M Kappler, M Keldorfer, DF Kelly, A Khanijau, N Kim, E Kim, S King, L Kolberg, M Kolnik, L Kloosterhuis, DS Kohlfürst, B Kohlmaier, L Krenn, S Leigh, M Leitner, B Leurent, E Lim, N Lin, C-C Liu, S Löffler, E Lurz, C Mackerness, I Maconochie, F Mallet, A Marmarinos, A Martin, M Martin, JM Martinón Sánchez, N Martinón-Torres,

P McAlinden, S McDonald, A McDonell, A Meiere, A Meierford, CJ Miedema, A Miners, R Mistry, M Mommert, S Morris, G Muench, DR Murdoch, S Mustafa, G Natalucci, C Neeleman, K Newall, S Nichols, A Niederer-Loher, T Niedrist, R Nijman, I Nokalna, G Nordberg, D O'Connor, CC Obihara, Z Oliver, W Oosthoek, MS Ora, V Osterman, A Pachot, D Pajkrt, J Pardo-Seco, J Pavãre, IP Paz, S Paulus, BM Pérez, S Persand, A Pfleger, K Pfurtscheller, R Philipsen, A Pickering, B Pierce, H Pilch, S Pischedda, L Pölz, KM Posfay-Barbe, O Powell, P Prunk, Z Pučuka, G Rajic, A Rashid, L Redondo-Collazo, K Reiter, C Relly, M Rhodes, JG Rial, V Richmond, T Riedel, I Rivero Calle, A Roca, S Rödl, LP Rodríguez, C Rodríguez-Tenreiro, S Romaine, E Rowlands, A Rudzate, M Sagmeister, M Saidykhan, A Sallas, I Sarr, C Schoen, D Schonenberg, N Schweintzger, F Secka, K Selecka, P Shah, C-F Shen, S Shrestha, A Skrabl-Baumgartner, J Soon, M Sperl, E Sprenkeler, N Spyridis, TP Srovin, L Stampfer, M Stevens, M Stocker, V Strenger, CD Suárez, D Svile, K Syggelou, C Tal, M Tambouratzi, E Tavliavini, C Thakker, E Thomson, S Throson, H Till, GA Tramper-Stranders, CS Trasorras, A Trobisch, UN Urbãne, M Usman, L Valentine, K Van Aerde, JM Van den Berg, B Van den Broek, I Van der Giessen, M Van der Kuip, F Van der Velden, AM Van Furth, AJ Van Gool, M Van Leur, G van Mierlo, SR Vázquez, C Vermont, LG Vicente, K Vincek, O Vito, M Voice, D Wallia, B Walsh, S-M Wang, C Wedderburn, E Willems, C Wilson, A Wood, P Woodsford, V Wyss, M Xagorari, J Zachariasse, SMA Zaman, C Zurl, and M Zwerenz. A multi-platform approach to identify a blood-based host protein signature for distinguishing between bacterial and viral infections in febrile children (perform): a multi-cohort machine learning study. *The Lancet Digital Health*, 5(11):e774–e785, 2023.

[102] Z Dong and Y Chen. Transcriptomics: advances and approaches. *Sci China Life Sci*, 56(10):960–7, 2013.

[103] JA Herberg, M Kaforou, VJ Wright, H Shailes, H Eleftherohorinou, CJ Hoggart, M Cebey-López, MJ Carter, VA Janes, S Gormley, C Shimizu, AH Tremoulet, AM Barendregt, A Salas, J Kanegaye, AJ Pollard, SN Faust, S Patel, T Kuijpers, F Martinón-Torres, JC Burns, LJM Coin, M Levin, and IC for the. Diagnostic test accuracy of a 2-transcript host rna signature for discriminating bacterial vs viral infection in febrile children. *JAMA*, 316(8):835–845, 2016.

[104] ST Anderson, M Kaforou, AJ Brent, VJ Wright, CM Banwell, G Chagaluka, AC Crampin, HM Dockrell, N French, MS Hamilton, ML Hibberd, F Kern, PR Langford, L Ling, R Mlotha, THM Ottenhoff, S Pienaar, V Pillay, JAG Scott, H Twahir, RJ Wilkinson, LJ Coin, RS Heyderman, M Levin, and B Eley. Diagnosis of childhood tuberculosis and host rna expression in africa. *N Engl J Med*, 370(18):1712–1723, 2014.

[105] VJ Wright, JA Herberg, M Kaforou, C Shimizu, H Eleftherohorinou, H Shailes, AM Barendregt, S Menikou, S Gormley, M Berk, LT Hoang, AH Tremoulet, JT Kanegaye, LJM Coin, MP Glodé, M Hibberd, TW Kuijpers, CJ Hoggart, JC Burns, and M Levin. Diagnosis of kawasaki disease using a minimal whole-blood gene expression signature. *JAMA Pediatr*, 172(10):e182293, 2018.

[106] SA Byron, KR Van Keuren-Jensen, DM Engelthaler, JD Carpten, and DW Craig. Translating rna sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics*, 17(5):257–271, 2016.

[107] Z Cao and LR Yu. Mass spectrometry-based proteomics for biomarker discovery. *Methods Mol Biol*, 2486:3–17, 2022.

[108] RJ Rotello and TD Veenstra. Mass spectrometry techniques: Principles and practices for quantitative proteomics. *Curr Protein Pept Sci*, 22(2):121–133, 2021.

[109] P Mallick and B Kuster. Proteomics: a pragmatic perspective. *Nature Biotechnology*, 28(7):695–709, 2010.

[110] Y Mi, K Burnham, P Charles, R Heilig, I Vendrell, J Whalley, H Torrance, D Antcliffe, S May, M Neville, G Berridge, P Hutton, C Goh, J Radhakrishnan, A Nesvizhskii, F Yu, E Davenport, S McKechnie, R Davies, DJP O'Callaghan, P Patel, F Karpe, A Gordon, G Ackland, C Hinds, R Fischer, J Knight, and G Investigators. High-throughput mass spectrometry maps the sepsis plasma proteome and differences in response, 2022.

[111] A Stukalov, V Girault, V Grass, O Karayel, V Bergant, C Urban, DA Haas, Y Huang, L Oubraham, A Wang, MS Hamad, A Piras, FM Hansen, MC Tanzer, I Paron, L Zinzula, T Engleitner, M Reinecke, TM Lavacca, R Ehmann, R Wölfel, J Jores, B Kuster, U Protzer, R Rad, J Ziebuhr, V Thiel, P Scaturro, M Mann, and A Pichlmair. Multilevel proteomics reveals host perturbations by sars-cov-2 and sars-cov. *Nature*, 594(7862):246–252, 2021.

[112] GQ Yin, HX Zeng, ZL Li, C Chen, JY Zhong, MS Xiao, Q Zeng, WH Jiang, PQ Wu, JM Zeng, XY Hu, HH Chen, H Ruo, HJ Zhao, L Gao, C Liu, and SX Cai. Differential proteomic analysis of children infected with respiratory syncytial virus. *Braz J Med Biol Res*, 54(4):e9850, 2021.

[113] LM Palma Medina, H Babačić, M Dzidic, Å Parke, M Garcia, KT Maleki, C Unge, M Lourda, E Kvedaraite, P Chen, JR Muvva, M Cornillet, J Emgård, K Moll, J Michaëlsson, M Flodström-Tullberg, S Brighenti, M Buggert, J Mjösberg, KJ Malmberg, JK Sandberg, S Gredmark-Russ, O Rooyackers, M Svensson,

BJ Chambers, LI Eriksson, M Pernemalm, NK Björkström, S Aleman, HG Ljunggren, J Klingström, K Strålin, and A Norrby-Teglund. Targeted plasma proteomics reveals signatures discriminating covid-19 from sepsis with pneumonia. *Respir Res*, 24(1):62, 2023.

[114] H Kupcova Skalnikova, J Cizkova, J Cervenka, and P Vodicka. Advances in proteomic techniques for cytokine analysis: Focus on melanoma research. *Int J Mol Sci*, 18(12), 2017.

[115] A Mobed, SK Shakouri, and S Dolati. Biosensors: A novel approach to and recent discovery in detection of cytokines. *Cytokine*, 136:155272, 2020.

[116] MJ Cameron and DJ Kelvin. Cytokines and chemokines–their receptors and their genes: an overview. *Adv Exp Med Biol*, 520:8–32, 2003.

[117] AA Justiz Vaillant and A Qurie. *Interleukin*. StatPearls Publishing Copyright © 2024, StatPearls Publishing LLC., 2024.

[118] C Liu, D Chu, K Kalantar-Zadeh, J George, HA Young, and G Liu. Cytokines: From clinical significance to quantification. *Advanced Science*, 8(15):2004433, 2021.

[119] J Scheller, A Chalaris, D Schmidt-Arras, and S Rose-John. The pro- and anti-inflammatory properties of the cytokine interleukin-6. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1813(5):878–888, 2011.

[120] JA Hamilton. Colony-stimulating factors in inflammation and autoimmunity. *Nature Reviews Immunology*, 8(7):533–544, 2008.

[121] D Metcalf. The colony-stimulating factors and cancer. *Cancer Immunol Res*, 1(6):351–6, 2013.

[122] M Croft and RM Siegel. Beyond tnf: Tnf superfamily cytokines as targets for the treatment of rheumatic diseases. *Nat Rev Rheumatol*, 13(4):217–233, 2017.

[123] SP Commins, L Borish, and JW Steinke. Immunologic messenger molecules: Cytokines, interferons, and chemokines. *Journal of Allergy and Clinical Immunology*, 125(2):S53–S72, 2010.

[124] S Manivasagam and RS Klein. Type iii interferons: Emerging roles in autoimmunity. *Front Immunol*, 12:764062, 2021.

[125] PC Ng, SH Cheng, KM Chui, TF Fok, MY Wong, W Wong, RP Wong, and KL Cheung. Diagnosis of late onset neonatal sepsis with cytokines, adhesion molecule, and c-reactive protein in preterm very low birthweight infants. *Arch Dis Child Fetal Neonatal Ed*, 77(3):F221–7, 1997.

[126] L Bergantini, M d'Alessandro, P Cameli, A Otranto, S Luzzi, F Bianchi, and E Bargagli. Cytokine profiles in the detection of severe lung involvement in hospitalized patients with covid-19: The il-8/il-32 axis. *Cytokine*, 151:155804, 2022.

[127] G Zhu, J Zhu, L Song, W Cai, and J Wang. Combined use of biomarkers for distinguishing between bacterial and viral etiologies in pediatric lower respiratory tract infections. *Infectious Diseases*, 47(5):289–293, 2015.

[128] J ten Oever, M Tromp, CP Bleeker-Rovers, LA Joosten, MG Netea, P Pickkers, and FL van de Veerdonk. Combination of biomarkers for the discrimination between bacterial and viral lower respiratory tract infections. *J Infect*, 65(6):490–5, 2012.

[129] EL Chiswick, E Duffy, B Japp, and D Remick. Detection and quantification of cytokines and other biomarkers. *Methods Mol Biol*, 844:15–30, 2012.

[130] WP Stefura, C Graham, L Lotoski, and KT HayGlass. *Improved Methods for Quantifying Human Chemokine and Cytokine Biomarker Responses: Ultrasensitive ELISA and Meso Scale Electrochemiluminescence Assays*. Springer New York, 2019.

[131] A HajYasien. *Introduction to Multiomics Technology*. Springer International Publishing, 2024.

[132] Y Hasin, M Seldin, and A Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, 2017.

[133] COMBAT Consortium. A blood atlas of covid-19 defines hallmarks of disease severity and specificity. *Cell*, 185(5):916–938.e58, 2022.

[134] E Stephenson, G Reynolds, RA Botting, FJ Calero-Nieto, MD Morgan, ZK Tuong, K Bach, W Sungnak, KB Worlock, M Yoshida, N Kumasaka, K Kania, J Engelbert, B Olabi, JS Spegarova, NK Wilson, N Mende, L Jardine, LCS Gardner, I Goh, D Horsfall, J McGrath, S Webb, MW Mather, RGH Lindeboom, E Dann, N Huang, K Polanski, E Prigmore, F Gothe, J Scott, RP Payne, KF Baker, AT Hanrath, ICD Schim van der Loeff, AS Barr, A Sanchez-Gonzalez, L Bergamaschi, F Mescia, JL Barnes, E Kilich, A de Wilton, A Saigal, A Saleh, SM Janes, CM Smith, N Gopee, C Wilson, P Coupland, JM Coxhead, VY Kiselev, S van Dongen, J Bacardit, HW King, AJ Rostron, AJ Simpson, S Hambleton, E Laurenti, PA Lyons, KB Meyer, MZ Nikolić, CJA Duncan, KGC Smith, SA Teichmann, MR Clatworthy, JC Marioni, B Göttgens, and M Haniffa. Single-cell multi-omics analysis of the immune response in covid-19. *Nat Med*, 27(5):904–916, 2021.

[135] JM Wozniak, RH Mills, J Olson, JR Caldera, GD Sepich-Poore, M Carrillo-Terrazas, CM Tsai, F Vargas, R Knight, PC Dorrestein, GY Liu, V Nizet, G Sakoulas, W Rose, and DJ Gonzalez. Mortality risk profiling of staphylococcus aureus bacteremia by multi-omic serum analysis reveals early predictive and pathogenic signatures. *Cell*, 182(5):1311–1327.e14, 2020.

[136] Y Xiao, M Bi, H Guo, and M Li. Multi-omics approaches for biomarker discovery in early ovarian cancer diagnosis. *EBioMedicine*, 79:104001, 2022.

[137] J Montaner, L Ramiro, A Simats, S Tiedt, K Makris, GC Jickling, S Debette, JC Sanchez, and A Bustamante. Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. *Nat Rev Neurol*, 16(5):247–264, 2020.

[138] Y Tao, S Xing, S Zuo, P Bao, Y Jin, Y Li, M Li, Y Wu, S Chen, X Wang, Y Zhu, Y Feng, X Zhang, X Wang, Q Xi, Q Lu, P Wang, and ZJ Lu. Cell-free multi-omics analysis reveals potential biomarkers in gastrointestinal cancer patients' blood. *Cell Rep Med*, 4(11):101281, 2023.

[139] S Tarazona, A Arzalluz-Luque, and A Conesa. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science*, 1(6):395–402, 2021.

[140] R Argelaguet, B Velten, D Arnol, S Dietrich, T Zenz, JC Marioni, F Buettner, W Huber, and O Stegle. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, 2018.

[141] M Habib, BD Porter, and C Satzke. Capsular serotyping of streptococcus pneumoniae using the quellung reaction. *J Vis Exp*, (84):e51208, 2014.

[142] National Planning Commission. Nepal earthquake 2015-post disaster assessment, vol.a: Key findings nepal: Government of nepal. Report, 2015.

[143] S Shrestha, M Gurung, P Amatya, S Bijukchhe, AS Bose, MJ Carter, MC Gautam, S Gurung, J Hinds, R Kandasamy, S Kelly, B Khadka, P Maskey, YF Mujadidi, PJ O'Reilly, B Pokhrel, R Pradhan, GP Shah, S Shrestha, B Wahl, KL O'Brien, MD Knoll, DR Murdoch, DF Kelly, S Thorson, M Voysey, AJ Pollard, K Acharya, B Acharya, I Ansari, R Basi, S Bista, S Bista, AK Budha, S Budhathoki, R Deshar, S Dhungel, S Felle, K Gautam, K Gorham, TY Gurung, P Gurung, R Jha, M K.C, SR Karnikar, A Kattel, L Lama, TKP Magar, M Maharjan, A Mallik, A Michel, D Nepal, J Nepal, KM Park, KG Prajapati, R Pudasaini, S Shrestha, M Smedley, R Weeks, JK Yadav, and SK Yadav. Effect of the of 10-valent pneumococcal conjugate vaccine in nepal 4 years

after introduction: an observational cohort study. *The Lancet Global Health*, 10(10):e1494–e1504, 2022.

[144] TK Karki. Political blackmailing: A case study of india's unofficial blockade on nepal. *SSRN Electronic Journal*, 2022.

[145] A Bastola, R Sah, AJ Rodriguez-Morales, BK Lal, R Jha, HC Ojha, B Shrestha, DKW Chu, LLM Poon, A Costello, K Morita, and BD Pandey. The first 2019 novel coronavirus case in nepal. *Lancet Infect Dis*, 20(3):279–280, 2020.

[146] World Health Organisation. National covid-19 epi dashboard, 2024.

[147] I Koutlakis-Barron and TA Hayden. Essentials of infection prevention in the pediatric population. *Int J Pediatr Adolesc Med*, 3(4):143–152, 2016.

[148] M Suneja, SE Beekmann, G Dhaliwal, AC Miller, and PM Polgreen. Diagnostic delays in infectious diseases. *Diagnosis (Berl)*, 9(3):332–339, 2022.

[149] R Stark, M Grzelak, and J Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

[150] ME Ritchie, B Phipson, D Wu, Y Hu, CW Law, W Shi, and GK Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res*, 43(7):e47, 2015.

[151] F Finotello and B Di Camillo. Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, 14(2):130–142, 2015.

[152] S Liu, Z Wang, R Zhu, F Wang, Y Cheng, and Y Liu. Three differential expression analysis methods for rna sequencing: limma, edger, deseq2. *J Vis Exp*, (175), 2021.

[153] TM Nguyen, A Shafi, T Nguyen, and S Draghici. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol*, 20(1):203, 2019.

[154] A Subramanian, P Tamayo, VK Mootha, S Mukherjee, BL Ebert, MA Gillette, A Paulovich, SL Pomeroy, TR Golub, ES Lander, and JP Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50, 2005.

[155] HD Gliddon, M Kaforou, M Alikian, D Habgood-Coote, C Zhou, T Oni, ST Anderson, AJ Brent, AC Crampin, B Eley, R Heyderman, F Kern, PR Langford,

THM Ottenhoff, ML Hibberd, N French, VJ Wright, HM Dockrell, LJ Coin, RJ Wilkinson, and M Levin. Identification of reduced host transcriptomic signatures for tuberculosis disease and digital pcr-based validation and quantification. *Frontiers in Immunology*, 12, 2021.

[156] HR Jackson, L Miglietta, D Habgood-Coote, G D'Souza, P Shah, S Nichols, O Vito, O Powell, MS Davidson, C Shimizu, PKA Agyeman, CR Beudeker, K Brengel-Pesce, ED Carrol, MJ Carter, T De, I Eleftheriou, M Emonts, C Epalza, P Georgiou, R De Groot, K Fidler, C Fink, D van Keulen, T Kuijpers, H Moll, I Papatheodorou, S Paulus, M Pokorn, AJ Pollard, I Rivero-Calle, P Rojo, F Secka, LJ Schlapbach, AH Tremoulet, M Tsolia, E Usuf, M Van Der Flier, U Von Both, C Vermont, S Yeung, D Zavadska, W Zenz, LJM Coin, A Cunnington, JC Burns, V Wright, F Martinon-Torres, JA Herberg, J Rodriguez-Manzano, M Kaforou, and M Levin. Diagnosis of multisystem inflammatory syndrome in children by a whole-blood transcriptional signature. *J Pediatric Infect Dis Soc*, 12(6):322–331, 2023.

[157] NM Suarez, E Bunsow, AR Falsey, EE Walsh, A Mejias, and O Ramilo. Superiority of transcriptional profiling over procalcitonin for distinguishing bacterial from viral lower respiratory tract infections in hospitalized adults. *J Infect Dis*, 212(2):213–22, 2015.

[158] R Patro, G Duggal, MI Love, RA Irizarry, and C Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*, 14(4):417–419, 2017.

[159] The SAM BAM Format Specification Working Group. Sequence alignment map format specification, 2023.

[160] S Anders, PT Pyl, and W Huber. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–9, 2015.

[161] M Carter. *Childhood pneumococcal pneumonia in Nepal*. Thesis, 2017.

[162] CL Schoch, S Ciufo, M Domrachev, CL Hotton, S Kannan, R Khovanskaya, D Leipe, R McVeigh, K O'Neill, B Robbertse, S Sharma, V Soussov, JP Sullivan, L Sun, S Turner, and I Karsch-Mizrachi. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)*, 2020, 2020.

[163] Marc Carlson. *org.Hs.eg.db: Genome wide annotation for Human*, 2023. R package version 3.17.0.

[164] MD Robinson, DJ McCarthy, and GK Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, 2010.

[165] CW Law, Y Chen, W Shi, and GK Smyth. voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15(2):R29, 2014.

[166] Belinda Phipson, Anna Trigos, Matt Ritchie, Shian Su, Maria Doyle, Harriet Dashnow, and Charity Law. Rna-seq analysis in r, 2020.

[167] M Charrad, N Ghazzali, V Boiteau, and A Niknafs. Nbclust: An r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1 – 36, 2014.

[168] K Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med*, 4(2):627–35, 2013.

[169] J Reimand, R Isserlin, V Voisin, M Kucera, C Tannus-Lopes, A Rostamianfar, L Wadi, M Meyer, J Wong, C Xu, D Merico, and GD Bader. Pathway enrichment analysis and visualization of omics data using g:profiler, gsea, cytoscape and enrichmentmap. *Nature Protocols*, 14(2):482–517, 2019.

[170] A Liberzon, A Subramanian, R Pinchback, H Thorvaldsdóttir, P Tamayo, and JP Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–40, 2011.

[171] JB Vincent, MW Crowder, and BA Averill. Hydrolysis of phosphate monoesters: a biological problem with multiple chemical solutions. *Trends in Biochemical Sciences*, 17(3):105–110, 1992.

[172] E Smiley-Moreno, D Smith, JJ Yu, P Cao, BP Arulanandam, and JP Chambers. Biochemical characterization of a recombinant acid phosphatase from acinetobacter baumannii. *PLoS One*, 16(6):e0252377, 2021.

[173] E du Plessis, J Theron, E Berger, and M Louw. Evaluation of the staphylococcus aureus class c nonspecific acid phosphatase (saps) as a reporter for gene expression and protein secretion in gram-negative and gram-positive bacteria. *Appl Environ Microbiol*, 73(22):7232–9, 2007.

[174] NG Nossal and LA Heppel. The release of enzymes by osmotic shock from escherichia coli in exponential phase. *J Biol Chem*, 241(13):3055–62, 1966.

[175] FS Sharief and SS Li. Structure of human prostatic acid phosphatase gene. *Biochem Biophys Res Commun*, 184(3):1468–76, 1992.

[176] WL Smith and FA Fitzpatrick. *Chapter 11 - The eicosanoids: cyclooxygenase, lipoxygenase, and epoxygenase pathways*. Elsevier, 1996.

[177] NC Ayola-Serrano, N Roy, Z Fathah, MM Anwar, B Singh, N Ammar, R Sah, A Elba, RS Utt, S Pecho-Silva, AJ Rodriguez-Morales, K Dhama, and S Quraishi. The role of 5-lipoxygenase in the pathophysiology of covid-19 and its therapeutic implications. *Inflamm Res*, 70(8):877–889, 2021.

[178] B Dufrusine, S Valentinuzzi, S Bibbò, V Damiani, P Lanuti, D Pieragostino, P Del Boccio, E D'Alessandro, A Rabottini, A Berghella, N Allocati, K Falasca, C Ucciferri, F Mucedola, M Di Perna, L Martino, J Vecchiet, V De Laurenzi, and E Dainese. Iron dyshomeostasis in covid-19: Biomarkers reveal a functional link to 5-lipoxygenase activation. *Int J Mol Sci*, 24(1), 2022.

[179] RH Pan, X Zhang, ZP Chen, and YJ Liu. Arachidonate lipoxygenases 5 is a novel prognostic biomarker and correlates with high tumor immune infiltration in low-grade glioma. *Front Genet*, 14:1027690, 2023.

[180] A Togayachi, T Akashima, R Ookubo, T Kudo, S Nishihara, H Iwasaki, A Natsume, H Mio, J Inokuchi, T Irimura, K Sasaki, and H Narimatsu. Molecular cloning and characterization of udp-glcnac:lactosylceramide beta 1,3-n-acetylglucosaminyltransferase (beta 3gn-t5), an essential enzyme for the expression of hnk-1 and lewis x epitopes on glycolipids. *J Biol Chem*, 276(25):22032–40, 2001.

[181] A Magalhães, R Marcos-Pinto, AV Nairn, M Dela Rosa, RM Ferreira, S Junqueira-Neto, D Freitas, J Gomes, P Oliveira, MR Santos, NT Marcos, W Xiaogang, C Figueiredo, C Oliveira, M Dinis-Ribeiro, F Carneiro, KW Moremen, L David, and CA Reis. Helicobacter pylori chronic infection and mucosal inflammation switches the human gastric glycosylation pathways. *Biochim Biophys Acta*, 1852(9):1928–39, 2015.

[182] STR Moolamalla, R Balasubramanian, R Chauhan, UD Priyakumar, and PK Vinod. Host metabolic reprogramming in response to sars-cov-2 infection: A systems biology approach. *Microb Pathog*, 158:105114, 2021.

[183] JD Storey, J Madeoy, JL Strout, M Wurfel, J Ronald, and JM Akey. Gene-expression variation within and among human populations. *The American Journal of Human Genetics*, 80(3):502–509, 2007.

[184] RS Spielman, LA Bastone, JT Burdick, M Morley, WJ Ewens, and VG Cheung. Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet*, 39(2):226–31, 2007.

[185] K Taniguchi and M Karin. Nf-b, inflammation, immunity and cancer: coming of age. *Nature Reviews Immunology*, 18(5):309–324, 2018.

[186] SJ Cho, M Plataki, D Mitzel, G Lowry, K Rooney, and H Stout-Delgado. Decreased nlrp3 inflammasome expression in aged lung may contribute to increased susceptibility to secondary streptococcus pneumoniae infection. *Exp Gerontol*, 105:40–46, 2018.

[187] B Albiger, S Dahlberg, A Sandgren, F Wartha, K Beiter, H Katsuragi, S Akira, S Normark, and B Henriques-Normark. Toll-like receptor 9 acts at an early stage in host defence against pneumococcal infection. *Cell Microbiol*, 9(3):633–44, 2007.

[188] F Abdallah, S Coindre, M Gardet, F Meurisse, A Naji, N Suganuma, L Abi-Rached, O Lambotte, and B Favier. Leukocyte immunoglobulin-like receptors in regulating the immune response in infectious diseases: A window of opportunity to pathogen persistence and a sound target in therapeutics. *Frontiers in Immunology*, 12, 2021.

[189] K Hirayasu, F Saito, T Suenaga, K Shida, N Arase, K Oikawa, T Yamaoka, H Murota, H Chibana, I Nakagawa, T Kubori, H Nagai, Y Nakamaru, I Katayama, M Colonna, and H Arase. Microbially cleaved immunoglobulins are sensed by the innate immune receptor lilra2. *Nat Microbiol*, 1(6):16054, 2016.

[190] X Wang, J Zhu, D Zhang, and G Liu. Ribosomal control in rna virus-infected cells. *Front Microbiol*, 13:1026887, 2022.

[191] E Champagne. Special issue "gamma delta t cells in immune response against viruses". *Viruses*, 14(4), 2022.

[192] D Chen, Y Guo, J Jiang, P Wu, T Zhang, Q Wei, J Huang, and D Wu. t cell exhaustion: Opportunities for intervention. *Journal of Leukocyte Biology*, 112(6):1669–1676, 2022.

[193] E Cimini, G Grassi, A Beccacece, R Casetti, C Castilletti, MR Capobianchi, E Nicastri, and C Agrati. In acute dengue infection, high tim-3 expression may contribute to the impairment of ifn production by circulating v2 t cells. *Viruses*, 14(1), 2022.

[194] JP Cerapio, M Perrier, F Pont, M Tosolini, C Laurent, S Bertani, and JJ Fournie. Single-cell rnaseq profiling of human t lymphocytes in virus-related cancers and covid-19 disease. *Viruses*, 13(11), 2021.

[195] W De Weerd, WN Twilhaar, and JL Kimpen. T cell subset analysis in peripheral blood of children with rsv bronchiolitis. *Scand J Infect Dis*, 30(1):77–80, 1998.

[196] HH Park. Structure of traf family: Current understanding of receptor recognition. *Frontiers in Immunology*, 9, 2018.

[197] H Konno, T Yamamoto, K Yamazaki, J Gohda, T Akiyama, K Semba, H Goto, A Kato, T Yujiri, T Imai, Y Kawaguchi, B Su, O Takeuchi, S Akira, Y Tsunetsugu-Yokota, and J Inoue. Traf6 establishes innate immune responses by activating nf-kappab and irf7 upon sensing cytosolic viral rna and dna. *PLoS One*, 4(5):e5674, 2009.

[198] SR Shuken. An introduction to mass spectrometry-based proteomics. *J Proteome Res*, 22(7):2151–2171, 2023.

[199] SN Thomas, D French, PJ Jannetto, BA Rappold, and WA Clarke. Liquid chromatography–tandem mass spectrometry for clinical diagnostics. *Nature Reviews Methods Primers*, 2(1):96, 2022.

[200] J Betancourt and S Gottlieb. *Liquid Chromatography - Libretexts*. 2023.

[201] V Demichev, CB Messner, SI Vernardis, KS Lilley, and M Ralser. Dia-nn: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17(1):41–44, 2020.

[202] R Zaru and S Orchard. Uniprot tools: Blast, align, peptide search, and id mapping. *Curr Protoc*, 3(3):e697, 2023.

[203] J Willforss, A Chawade, and F Levander. Normalyzerde: Online tool for improved normalization of omics expression data and high-sensitivity differential expression analysis. *J Proteome Res*, 18(2):732–740, 2019.

[204] ES Nakayasu, M Gritsenko, PD Piehowski, Y Gao, DJ Orton, AA Schepmoes, TL Fillmore, BI Frohnert, M Rewers, JP Krischer, C Ansong, AM Suchy-Dicey, C Evans-Molina, W-J Qian, B-JM Webb-Robertson, and TO Metz. Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation. *Nature Protocols*, 16(8):3737–3760, 2021.

[205] YV Karpievitch, AR Dabney, and RD Smith. Normalization and missing value imputation for label-free lc-ms analysis. *BMC Bioinformatics*, 13(16):S5, 2012.

[206] Y Yu, N Zhang, Y Mai, L Ren, Q Chen, Z Cao, Q Chen, Y Liu, W Hou, J Yang, H Hong, J Xu, W Tong, L Dong, L Shi, X Fang, and Y Zheng. Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. *Genome Biology*, 24(1):201, 2023.

[207] J Čuklina, CH Lee, EG Williams, T Sajic, BC Collins, M Rodríguez Martínez, VS Sharma, F Wendt, S Goetze, GR Keele, B Wollscheid, R Aebersold, and PGA Pedrioli. Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. *Molecular Systems Biology*, 17(8):e10240, 2021.

[208] CW Law, K Zeglinski, X Dong, M Alhamdoosh, GK Smyth, and ME Ritchie. A guide to creating design matrices for gene expression experiments. *F1000Res*, 9:1444, 2020.

[209] GeneOntology. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res*, 47(D1):D330–d338, 2019.

[210] Kee Siong Ng. A simple explanation of partial least squares. *Australian National University Library*, 2013.

[211] Max Kuhn. The caret package, 2019.

[212] Y Zou, Y Xu, X Chen, Y Wu, L Fu, and Y Lv. Research progress on leucine-rich alpha-2 glycoprotein 1: A review. *Front Pharmacol*, 12:809225, 2021.

[213] H Honda, M Fujimoto, S Serada, H Urushima, T Mishima, H Lee, T Ohkawara, N Kohno, N Hattori, A Yokoyama, and T Naka. Leucine-rich -2 glycoprotein promotes lung fibrosis by modulating tgf- signaling in fibroblasts. *Physiol Rep*, 5(24), 2017.

[214] X Wang, S Abraham, JAG McKenzie, N Jeffs, M Swire, VB Tripathi, UFO Luhmann, CAK Lange, Z Zhai, HM Arthur, J Bainbridge, SE Moss, and J Greenwood. Lrg1 promotes angiogenesis by modulating endothelial tgf- signalling. *Nature*, 499(7458):306–11, 2013.

[215] T Shimoyama, T Yamamoto, S Yoshiyama, R Nishikawa, and S Umegae. Leucine-rich alpha-2 glycoprotein is a reliable serum biomarker for evaluating clinical and endoscopic disease activity in inflammatory bowel disease. *Inflamm Bowel Dis*, 29(9):1399–1408, 2023.

[216] J Arredondo Montero, BP Pérez Riveros, OE Bueso Asfura, M Rico Jiménez, N López-Andrés, and N Martín-Calvo. Leucine-rich alpha-2-glycoprotein as a non-invasive biomarker for pediatric acute appendicitis: a systematic review and meta-analysis. *Eur J Pediatr*, 182(7):3033–3044, 2023.

[217] Y Zhang, J Zhang, H Sheng, H Li, and R Wang. *Chapter Two - Acute phase reactant serum amyloid A in inflammation and other diseases*. Elsevier, 2019.

[218] W Su, L Ju, Q Hua, J Hu, and W Qian. Values of combined c-reactive protein, procalcitonin and serum amyloid a in differential diagnosis of bacterial and non-bacterial community acquired pneumonia in children. *Diagn Microbiol Infect Dis*, 105(2):115865, 2023.

[219] CD Russell, SA Unger, M Walton, and J Schwarze. The human immune response to respiratory syncytial virus infection. *Clin Microbiol Rev*, 30(2):481–502, 2017.

[220] A Bonchuk, K Balagurov, and P Georgiev. Btb domains: A structural view of evolution, multimerization, and protein–protein interactions. *BioEssays*, 45(2):2200179, 2023.

[221] F Noro, F Gianfagna, A Gialluisi, A De Curtis, A Di Castelnuovo, E Napoleone, C Cerletti, MB Donati, G de Gaetano, MF Hoylaerts, L Iacoviello, and B Izzi. Zbtb12 dna methylation is associated with coagulation- and inflammation-related blood cell parameters: findings from the moli-family cohort. *Clin Epigenetics*, 11(1):74, 2019.

[222] M Rahmouni, L De Marco, JL Spadoni, M Tison, R Medina-Santos, T Labib, J Noirel, R Tamouza, S Limou, O Delaneau, J Fellay, A Bensussan, S Le Clerc, PJ McLaren, and JF Zagury. The hla-b*57:01 allele corresponds to a very large mhc haploblock likely explaining its massive effect for hiv-1 elite control. *Front Immunol*, 14:1305856, 2023.

[223] E Kostareli, A Hadzidimitriou, N Stavroyianni, N Darzentas, A Athanasiadou, M Gounari, V Bikos, A Agathagelidis, T Touloumenidou, I Zorbas, A Kouvatsi, N Laoutaris, A Fassas, A Anagnostopoulos, C Belessi, and K Stamatopoulos. Molecular evidence for ebv and cmv persistence in a subset of patients with chronic lymphocytic leukemia expressing stereotyped ighv4-34 b-cell receptors. *Leukemia*, 23(5):919–924, 2009.

[224] LM Loomes, K Uemura, RA Childs, JC Paulson, GN Rogers, PR Scudder, JC Michalski, EF Hounsell, D Taylor-Robinson, and T Feizi. Erythrocyte receptors for mycoplasma pneumoniae are sialylated oligosaccharides of ii antigen type. *Nature*, 307(5951):560–3, 1984.

[225] CI Mockridge, A Rahman, S Buchan, T Hamblin, DA Isenberg, FK Stevenson, and KN Potter. Common patterns of b cell perturbation and expanded v4-34 immunoglobulin gene usage in autoimmunity and infection. *Autoimmunity*, 37(1):9–15, 2004.

[226] J Salazar, MS Martínez, M Chávez-Castillo, V Núñez, R Añez, Y Torres, A Toledo, M Chacín, C Silva, E Pacheco, J Rojas, and V Bermúdez. C-reactive protein:

An in-depth look into structure, function, and regulation. *Int Sch Res Notices*, 2014:653045, 2014.

[227] Alhajj M, Zubair M, and Farhana A. Enzyme linked immunosorbent assay. [updated 2023 apr 23]., 2024.

[228] Y Liu, W Guo, and B Su. Recent advances in electrochemiluminescence imaging analysis based on nanomaterials and micro-/nanostructures. *Chinese Chemical Letters*, 30(9):1593–1599, 2019.

[229] H Wang. Advances in electrochemiluminescence co-reaction accelerator and its analytical applications. *Analytical and Bioanalytical Chemistry*, 413(16):4119–4135, 2021.

[230] C Wang, S Liu, and H Ju. Electrochemiluminescence nanoemitters for immunoassay of protein biomarkers. *Bioelectrochemistry*, 149:108281, 2023.

[231] Meso Scale Discovery. U-plex biomarker group 1 (human) multiplex assays, 2023.

[232] P Gallo, K McNerney, C Foley, L Kagami, K Wagner, W Petrosa, S Canna, C Diorio, K Sullivan, H Bassiri, DT Teachey, EM Behrens, M Paessler, and MP Lambert. Analysis of clinical cytokine panel profiles as diagnostic and prognostic biomarkers in pediatric hyperinflammatory conditions. *Blood*, 140(Supplement 1):8341–8342, 2022.

[233] I Andrijevic, J Matijasevic, L Andrijevic, T Kovacevic, and B Zaric. Interleukin-6 and procalcitonin as biomarkers in mortality prediction of hospitalized patients with community acquired pneumonia. *Ann Thorac Med*, 9(3):162–7, 2014.

[234] Gaggero-A. A. Pinto R. A. Mamani R. Uasapud P. A. Bono M. R. Diaz, PV. Levels of inflammatory cytokines and plasma cortisol in respiratory syncytial virus bronchiolitis. *Revista Medica de Chile*, 141, 2013.

[235] N Mukaida and K Matsushima. *Interleukin 8 and its Receptor*. Elsevier, 1998.

[236] C Mella, MC Suarez-Arrabal, S Lopez, J Stephens, S Fernandez, MW Hall, O Ramilo, and A Mejias. Innate immune dysfunction is associated with enhanced disease severity in infants with severe respiratory syncytial virus bronchiolitis. *J Infect Dis*, 207(4):564–73, 2013.

[237] Martin-Matthew S. Anklam Kelly A. Shen Kunling Roberg Kathy A. Carlson-Dakes Kirstin T. Adler Kiva Gilbertson-White Stephanie Hamilton Rebekah Shult Peter A. Kirk Carol J. Da Silva Douglas F. Sund Sarah A. Kosorok Michael R. Lemanske Robert F. Jr. Gern, JE. Relationships among specific viral

pathogens, virus-induced interleukin-8, and respiratory symptoms in infancy. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology*, 13, 2002.

[238] N Bishara. *Chapter 18 - The Use of Biomarkers for Detection of Early- and Late-Onset Neonatal Sepsis*. W.B. Saunders, 2012.

[239] TJ Standiford and JC Deng. *INTERLEUKINS — IL-10*. Academic Press, 2006.

[240] L Bont, CJ Heijnen, A Kavelaars, WM van Aalderen, F Brus, JT Draaisma, SM Geelen, and JL Kimpen. Monocyte il-10 production during respiratory syncytial virus bronchiolitis is associated with recurrent wheezing in a one-year follow-up study. *Am J Respir Crit Care Med*, 161(5):1518–23, 2000.

[241] G Kaplanski. Interleukin-18: Biological properties and role in disease pathogenesis. *Immunol Rev*, 281(1):138–153, 2018.

[242] JE Sims and DE Smith. The il-1 family: regulators of immunity. *Nature Reviews Immunology*, 10(2):89–102, 2010.

[243] JL Wynn, CS Wilson, J Hawiger, PO Scumpia, AF Marshall, JH Liu, I Zharkikh, HR Wong, P Lahni, JT Benjamin, EJ Plosa, JH Weitkamp, ER Sherwood, LL Moldawer, R Ungaro, HV Baker, MC Lopez, SJ McElroy, N Colliou, M Mohamadzadeh, and DJ Moore. Targeting il-17a attenuates neonatal sepsis mortality induced by il-18. *Proc Natl Acad Sci U S A*, 113(19):E2627–35, 2016.

[244] B Liu, I Mori, MJ Hossain, L Dong, K Takeda, and Y Kimura. Interleukin-18 improves the early defence system against influenza virus infection by augmenting natural killer cell-mediated cytotoxicity. *J Gen Virol*, 85(Pt 2):423–428, 2004.

[245] AE Denton, PC Doherty, SJ Turner, and NL La Gruta. Il-18, but not il-12, is required for optimal cytokine production by influenza virus-specific cd8+ t cells. *Eur J Immunol*, 37(2):368–75, 2007.

[246] X Li, R Bechara, J Zhao, MJ McGeachy, and SL Gaffen. Il-17 receptor–based signaling and implications for disease. *Nature Immunology*, 20(12):1594–1602, 2019.

[247] II Ivanov, K Atarashi, N Manel, EL Brodie, T Shima, U Karaoz, D Wei, KC Goldfarb, CA Santee, SV Lynch, T Tanoue, A Imaoka, K Itoh, K Takeda, Y Umesaki, K Honda, and DR Littman. Induction of intestinal th17 cells by segmented filamentous bacteria. *Cell*, 139(3):485–98, 2009.

[248] RA Pinto, SM Arredondo, MR Bono, AA Gaggero, and PV Díaz. T helper 1/t helper 2 cytokine imbalance in respiratory syncytial virus infection is associated with increased endogenous plasma cortisol. *Pediatrics*, 117(5):e878–86, 2006.

[249] PY Perera, JH Lichy, TA Waldmann, and LP Perera. The role of interleukin-15 in inflammation and immune responses to infection: implications for its therapeutic use. *Microbes Infect*, 14(3):247–61, 2012.

[250] M Altfeld, L Fadda, D Frleta, and N Bhardwaj. Dcs and nk cells: critical effectors in the immune response to hiv-1. *Nat Rev Immunol*, 11(3):176–86, 2011.

[251] MJ Maeurer, P Trinder, G Hommel, W Walter, K Freitag, D Atkins, and S Störkel. Interleukin-7 or interleukin-15 enhances survival of mycobacterium tuberculosis-infected mice. *Infect Immun*, 68(5):2962–70, 2000.

[252] TR Leahy, M Ross, GD Derek, G Robert, C Tanya, S Paul, B Gordon, S Orla, JC Michael, P Nikhil, G Michael, MH Hennie, J Riny, B Louis, S Dubhfeasa, and R Thomas. Interleukin-15 is associated with disease severity in viral bronchiolitis. *European Respiratory Journal*, 47(1):212, 2016.

[253] A Theyab, M Algahtani, KF Alsharif, YM Hawsawi, A Alghamdi, A Alghamdi, and J Akinwale. New insight into the mechanism of granulocyte colony-stimulating factor (g-csf) that induces the mobilization of neutrophils. *Hematology*, 26(1):628–636, 2021.

[254] AD Panopoulos and SS Watowich. Granulocyte colony-stimulating factor: molecular mechanisms of action during steady state and 'emergency' hematopoiesis. *Cytokine*, 42(3):277–88, 2008.

[255] H Li, Z Wang, and X Li. G-csf as a potential early biomarker for diagnosis of bloodstream infection. *J Clin Lab Anal*, 35(12):e23592, 2021.

[256] M Tsujimoto, M Sawaki, M Sakamoto, K Mikasa, K Hamada, K Maeda, S Teramoto, K Mori, K Ueda, N Narita, and E Kita. [the evaluation of interleukin-6 (il-6) and tumor necrosis factor alpha (tnf-alpha) level in peripheral blood of patients with chronic lower respiratory tract infection]. *Kansenshogaku Zasshi*, 71(5):430–6, 1997.

[257] Aberle-S. W. Rebhandl W. Pracher E. Kundi M. Popow-Kraupp T. Aberle, JH. Decreased interferon-gamma response in respiratory syncytial virus compared to other respiratory viral infections in infants. *Clinical and Experimental Immunology*, 137, 2004.

[258] M Liu, S Guo, JM Hibbert, V Jain, N Singh, NO Wilson, and JK Stiles. Cxcl10/ip-10 in infectious diseases pathogenesis and potential therapeutic implications. *Cytokine Growth Factor Rev*, 22(3):121–30, 2011.

[259] D Machado, J Hoffmann, M Moroso, M Rosa-Calatrava, H Endtz, O Terrier, and G Paranhos-Baccalà. Rsv infection in human macrophages promotes cxcl10/ip-10 expression during bacterial co-infection. *Int J Mol Sci*, 18(12), 2017.

[260] Meso Scale Discovery. Discovery workbench 4.0 user guide, 2013.

[261] JA Espinoza, S Jabeen, R Batra, E Papaleo, V Haakensen, V Timmermans Wielenga, ML Møller Talman, N Brunner, AL Børresen-Dale, P Gromov, Å Helland, VN Kristensen, and I Gromova. Cytokine profiling of tumor interstitial fluid of the breast and its relationship with lymphocyte infiltration and clinicopathological characteristics. *Oncoimmunology*, 5(12):e1248015, 2016.

[262] S Coleman, K Nicholls, K Nicholls, XC Dopico, G Karlsson Hedestam, PDW Kirk, and C Wallace. A semi-supervised bayesian mixture modelling approach for joint batch correction and classification, 2022.

[263] WE Johnson, C Li, and A Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[264] K Korthauer, PK Kimes, C Duvallet, A Reyes, A Subramanian, M Teng, C Shukla, EJ Alm, and SC Hicks. A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20(1):118, 2019.

[265] S Iwase, TA Nakada, N Hattori, W Takahashi, N Takahashi, T Aizimu, M Yoshida, T Morizane, and S Oda. Interleukin-6 as a diagnostic marker for infection in critically ill patients: A systematic review and meta-analysis. *Am J Emerg Med*, 37(2):260–265, 2019.

[266] EM Rincón-López, ML Navarro Gómez, T Hernández-Sampelayo Matos, D Aguilera-Alonso, E Dueñas Moreno, J Saavedra-Lozano, B Santiago García, MDM Santos Sebastián, M García Morín, C Beléndez Bieler, J Lorente Romero, and E Cela de Julián. Interleukin 6 as a marker of severe bacterial infection in children with sickle cell disease and fever: a case-control study. *BMC Infect Dis*, 21(1):741, 2021.

[267] P Xu, X Liu, Y Li, X Yin, and X Liu. Gene expression levels of peripheral blood mononuclear cells il-6 and tnf-alpha; in children with bacterial and viral infectious diarrhea. *J Coll Physicians Surg Pak*, 28(12):934–936, 2018.

[268] J Karhu, TI Ala-Kokko, T Vuorinen, P Ohtonen, I Julkunen, and HT Syrjälä. Interleukin-5, interleukin-6, interferon induced protein-10, procalcitonin and c-reactive protein among mechanically ventilated severe community-acquired viral and bacterial pneumonia patients. *Cytokine*, 113:272–276, 2019.

[269] C Cheers, AM Haigh, A Kelso, D Metcalf, ER Stanley, and AM Young. Production of colony-stimulating factors (csfs) during infection: separate determinations of macrophage-, granulocyte-, granulocyte-macrophage-, and multi-csfs. *Infect Immun*, 56(1):247–51, 1988.

[270] M Kawakami, H Tsutsumi, T Kumakawa, H Abe, M Hirai, S Kurosawa, M Mori, and M Fukushima. Levels of serum granulocyte colony-stimulating factor in patients with infections. *Blood*, 76(10):1962–4, 1990.

[271] E Kolaczkowska and P Kubes. Neutrophil recruitment and function in health and inflammation. *Nat Rev Immunol*, 13(3):159–75, 2013.

[272] A Ashkar Ali, S Reid, F Verdu Elena, K Zhang, and K Coombes Brian. Interleukin-15 and nk1.1+ cells provide innate protection against acute salmonella enterica serovar typhimurium infection in the gut and in systemic tissues. *Infection and Immunity*, 77(1):214–222, 2009.

[273] M Frahm, ND Goswami, K Owzar, E Hecker, A Mosher, E Cadogan, P Nahid, G Ferrari, and JE Stout. Discriminating between latent and active tuberculosis with multiple biomarker responses. *Tuberculosis*, 91(3):250–256, 2011.

[274] RJ Bishnoi, RF Palmer, and DR Royall. Serum interleukin (il)-15 as a biomarker of alzheimer's disease. *PLoS One*, 10(2):e0117282, 2015.

[275] T Shimizu, T Koga, K Furukawa, Y Horai, K Fujikawa, A Okada, M Okamoto, Y Endo, S Tsuji, A Takatani, M Umeda, S Fukui, R Sumiyoshi, SY Kawashiri, N Iwamoto, T Igawa, K Ichinose, M Tamai, N Sakamoto, H Nakamura, T Origuchi, H Mukae, M Kuwana, and A Kawakami. Il-15 is a biomarker involved in the development of rapidly progressive interstitial lung disease complicated with polymyositis/dermatomyositis. *J Intern Med*, 289(2):206–220, 2021.

[276] X Wang, J Jiang, Z Cao, B Yang, J Zhang, and X Cheng. Diagnostic performance of multiplex cytokine and chemokine assay for tuberculosis. *Tuberculosis (Edinb)*, 92(6):513–20, 2012.

[277] G Chen, TG Gharib, CC Huang, JM Taylor, DE Misek, SL Kardia, TJ Giordano, MD Iannettoni, MB Orringer, SM Hanash, and DG Beer. Discordant protein

and mrna expression in lung adenocarcinomas. *Mol Cell Proteomics*, 1(4):304–13, 2002.

[278] A Ghazalpour, B Bennett, VA Petyuk, L Orozco, R Hagopian, IN Mungrue, CR Farber, J Sinsheimer, HM Kang, N Furlotte, CC Park, PZ Wen, H Brewer, K Weitz, 2nd Camp, DG, C Pan, R Yordanova, I Neuhaus, C Tilford, N Siemers, P Gargalovic, E Eskin, T Kirchgessner, DJ Smith, RD Smith, and AJ Lusis. Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet*, 7(6):e1001393, 2011.

[279] I Subramanian, S Verma, S Kumar, A Jere, and K Anamika. Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights*, 14:11779322198990 51, 2020.

[280] F Rohart, B Gautier, A Singh, and K-A Lê Cao. mixomics: An r package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11):e1005752, 2017.

[281] S Wold, M Sjöström, and L Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[282] DV Nguyen and DM Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, 2002.

[283] M Pérez-Enciso and M Tenenhaus. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (pls-da) approach. *Human Genetics*, 112(5):581–592, 2003.

[284] A Singh, CP Shannon, B Gautier, F Rohart, M Vacher, SJ Tebbutt, and LC KA. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, 2019.

[285] PG Gavin, JA Mullaney, D Loo, KL Cao, PA Gottlieb, MM Hill, D Zipris, and EE Hamilton-Williams. Intestinal metaproteomics reveals host-microbiota interactions in subjects at risk for type 1 diabetes. *Diabetes Care*, 41(10):2178–2186, 2018.

[286] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[287] C The UniProt. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.

[288] SK Byeon, AK Madugundu, K Garapati, MG Ramarajan, M Saraswat, MP Kumar, T Hughes, R Shah, MM Patnaik, N Chia, S Ashrafzadeh-Kian, JD Yao, BS Pritt, R Cattaneo, ME Salama, RM Zenka, BR Kipp, SKG Grebe, RJ Singh, AA Sadighi Akha, A Algeciras-Schimnich, S Dasari, JE Olson, JR Walsh, AJ Venkatakrishnan, G Jenkinson, JC O'Horo, AD Badley, and A Pandey. Development of a multiomics model for identification of predictive biomarkers for covid-19 severity: a retrospective cohort study. *Lancet Digit Health*, 4(9):e632–e645, 2022.

[289] L Li, RM Sun, and GQ Jiang. Atf3 demethylation promotes the transcription of arl4c, which acts as a tumor suppressor in human breast cancer. *Onco Targets Ther*, 13:3467–3476, 2020.

[290] P Zhang, Y Xu, S Chen, Z Wang, L Zhao, C Chen, W Kang, R Han, J Qiu, Q Wang, H Gao, G Wu, and Q Xia. Arl4c regulates the progression of clear cell renal cell carcinoma by affecting the wnt/-catenin signaling pathway. *J Oncol*, 2022:2724515, 2022.

[291] Q Lin, R Zeng, J Yang, Z Xu, S Jin, and G Wei. Prognostic stratification of sepsis through dna damage response based riskscore system: insights from single-cell rna-sequencing and transcriptomic profiling. *Front Immunol*, 15:1345321, 2024.

[292] B Bender, M Baranyi, A Kerekes, L Bodrogi, R Brands, P Uhrin, and Z Bösze. Recombinant human tissue non-specific alkaline phosphatase successfully counteracts lipopolysaccharide induced sepsis in mice. *Physiol Res*, 64(5):731–8, 2015.

[293] JF Augusto, C Beauvillain, C Poli, L Paolini, I Tournier, P Pignon, S Blanchard, L Preisser, R Soleti, C Delépine, M Monnier, I Douchet, P Asfar, F Beloncle, O Guisset, R Prével, A Mercat, E Vinatier, J Goret, JF Subra, D Couez, MR Wilson, P Blanco, P Jeannin, and Y Delneste. Clusterin neutralizes the inflammatory and cytotoxic properties of extracellular histones in sepsis. *Am J Respir Crit Care Med*, 208(2):176–187, 2023.

[294] G Santilli, BJ Aronow, and A Sala. Essential requirement of apolipoprotein j (clusterin) signaling for ikappab expression and regulation of nf-kappab activity. *J Biol Chem*, 278(40):38214–9, 2003.

[295] E Piktel, I Levental, B Durnaś, PA Janmey, and R Bucki. Plasma gelsolin: Indicator of inflammation and its potential as a diagnostic tool and therapeutic target. *Int J Mol Sci*, 19(9), 2018.

[296] H Wang, B Cheng, Q Chen, S Wu, C Lv, G Xie, Y Jin, and X Fang. Time course of plasma gelsolin concentrations during severe sepsis in critically ill surgical patients. *Crit Care*, 12(4):R106, 2008.

[297] E Suhler, W Lin, HL Yin, and WM Lee. Decreased plasma gelsolin concentrations in acute liver failure, myocardial infarction, septic shock, and myonecrosis. *Crit Care Med*, 25(4):594–8, 1997.

[298] M Asare-Werehene, M McGuinty, A Vranjkovic, Y Galipeau, J Cowan, B Cameron, CL Cooper, MA Langlois, AM Crawley, and BK Tsang. Longitudinal profiles of plasma gelsolin, cytokines and antibody expression predict covid-19 severity and hospitalization outcomes. *Front Immunol*, 13:1011084, 2022.

[299] G Carvalho Mda, ML Tondella, K McCaustland, L Weidlich, L McGee, LW Mayer, A Steigerwalt, M Whaley, RR Facklam, B Fields, G Carlone, EW Ades, R Dagan, and JS Sampson. Evaluation and improvement of real-time pcr assays targeting lyta, ply, and psaa genes for detection of pneumococcal dna. *J Clin Microbiol*, 45(8):2460–6, 2007.

[300] J Rello, T Lisboa, M Lujan, M Gallego, C Kee, I Kay, D Lopez, and GW Waterer. Severity of pneumococcal pneumonia associated with genomic bacterial load. *Chest*, 136(3):832–840, 2009.

[301] PY Iroh Tam, N Hernandez-Alvarado, MR Schleiss, AJ Yi, F Hassan-Hanga, C Onuchukwu, D Umoru, and SK Obaro. Detection of streptococcus pneumoniae from culture-negative dried blood spots by real-time pcr in nigerian children with acute febrile illness. *BMC Res Notes*, 11(1):657, 2018.

[302] M Troendle and A Pettigrew. A systematic review of cases of meningitis in the absence of cerebrospinal fluid pleocytosis on lumbar puncture. *BMC Infectious Diseases*, 19(1):692, 2019.

[303] J Pardo, KP Klinker, SJ Borgert, BM Butler, KH Rand, and NM Iovine. Detection of neisseria meningitidis from negative blood cultures and cerebrospinal fluid with the filmarray blood culture identification panel. *J Clin Microbiol*, 52(6):2262–4, 2014.

[304] S Guiducci, M Moriondo, F Nieddu, S Ricci, E De Vitis, A Casini, GM Poggi, G Indolfi, M Resti, and C Azzari. Culture and real-time polymerase chain reaction sensitivity in the diagnosis of invasive meningococcal disease: Does culture miss less severe cases? *PLOS ONE*, 14(3):e0212922, 2019.

[305] DL Stevens. Invasive group a streptococcus infections. *Clinical Infectious Diseases*, 14(1):2–13, 1992.

[306] CW Obiero, W Gumbi, S Mwakio, H Mwangudzah, AC Seale, M Taniuchi, J Liu, E Houpt, and JA Berkley. Detection of pathogens associated with early-onset neonatal sepsis in cord blood at birth using quantitative pcr. *Wellcome Open Res*, 7:3, 2022.

[307] S Makvana and LR Krilov. Escherichia coli infections. *Pediatr Rev*, 36(4):167–70; quiz 171, 2015.

[308] B Lucignano, S Ranno, O Liesenfeld, B Pizzorno, L Putignani, P Bernaschi, and D Menichella. Multiplex pcr allows rapid and accurate diagnosis of bloodstream infections in newborns and children with suspected sepsis. *J Clin Microbiol*, 49(6):2252–8, 2011.

[309] EL Tsalik, D Jones, B Nicholson, L Waring, O Liesenfeld, LP Park, SW Glickman, LB Caram, RJ Langley, JC van Velkinburgh, CB Cairns, EP Rivers, RM Otero, SF Kingsmore, T Lalani, VG Fowler, and CW Woods. Multiplex pcr to diagnose bloodstream infections in patients admitted from the emergency department with sepsis. *J Clin Microbiol*, 48(1):26–33, 2010.

[310] A Heininger, M Binder, S Schmidt, K Unertl, K Botzenhart, and G Döring. Pcr and blood culture for detection of escherichia coli bacteremia in rats. *J Clin Microbiol*, 37(8):2479–82, 1999.

[311] M Beld, R Minnaar, J Weel, C Sol, M Damen, H van der Avoort, P Wertheim-van Dillen, A van Breda, and R Boom. Highly sensitive assay for detection of enterovirus in clinical specimens by reverse transcription-pcr with an armored rna internal control. *Journal of Clinical Microbiology*, 42(7):3059–3064, 2004.

[312] H Harvala, E Broberg, K Benschop, N Berginc, S Ladhani, P Susi, C Christiansen, J McKenna, D Allen, P Makiello, G McAllister, M Carmen, K Zakikhany, R Dyrdak, X Nielsen, T Madsen, J Paul, C Moore, K von Eije, A Piralla, M Carlier, L Vanoverschelde, R Poelman, A Anton, FX López-Labrador, L Pellegrinelli, K Keeren, M Maier, H Cassidy, S Derdas, C Savolainen-Kopra, S Diedrich, S Nordbø, J Buesa, J-L Bailly, F Baldanti, A MacAdam, A Mirand, S Dudman, I Schuffenecker, S Kadambari, J Neyts, MJ Griffiths, J Richter, C Margaretto, S Govind, U Morley, O Adams, S Krokstad, J Dean, M Pons-Salort, B Prochazka, M Cabrerizo, M Majumdar, G Nebbia, M Wiewel, S Cottrell, P Coyle, J Martin, C Moore, S Midgley, P Horby, K Wolthers, P Simmonds, H Niesters, and TK Fischer. Recommendations for enterovirus diagnostics and characterisation within and beyond europe. *Journal of Clinical Virology*, 101:11–17, 2018.

[313] G Iqbal, H Javed, FA Raza, UF Gohar, W Fatima, and M Khurshid. Diagnosis of acute dengue virus infection using enzyme-linked immunosorbent assay and real-time pcr. *Can J Infect Dis Med Microbiol*, 2023:3995366, 2023.

[314] AM Anand, S Sistla, R Dhodapkar, A Hamide, N Biswal, and B Srinivasan. Evaluation of ns1 antigen detection for early diagnosis of dengue in a tertiary hospital in southern india. *J Clin Diagn Res*, 10(4):Dc01–4, 2016.

[315] BL Innis, A Nisalak, S Nimmannitya, S Kusalerdchariya, V Chongswasdi, S Suntayakorn, P Puttisri, and CH Hoke. An enzyme-linked immunosorbent assay to characterize dengue infections where dengue and japanese encephalitis co-circulate. *Am J Trop Med Hyg*, 40(4):418–27, 1989.

[316] AJ Hessels, J Liu, B Cohen, J Shang, and EL Larson. Severity of illness measures for pediatric inpatients. *J Healthc Qual*, 40(5):e77–e89, 2018.

[317] N Peshimam and S Nadel. Sepsis in children: state-of-the-art treatment. *Ther Adv Infect Dis*, 8:20499361211055332, 2021.

[318] L Pruinelli, BL Westra, P Yadav, A Hoff, M Steinbach, V Kumar, CW Delaney, and G Simon. Delay within the 3-hour surviving sepsis campaign guideline on mortality for patients with severe sepsis and septic shock. *Crit Care Med*, 46(4):500–505, 2018.

[319] TP Htun, Z Xiong, and J Pang. Clinical signs and symptoms associated with who severe dengue classification: a systematic review and meta-analysis. *Emerg Microbes Infect*, 10(1):1116–1128, 2021.

[320] SRS Hadinegoro. The revised who dengue case classification: does the system need to be modified? *Paediatrics and International Child Health*, 32(sup1):33–38, 2012.

[321] QIAGEN GmbH for PreAnalytiX. Paxgene - blood rna kit instructions for use (handbook). Report, 2023.

[322] SR Howie. Blood sample volumes in child health research: review of safe limits. *Bull World Health Organ*, 89(1):46–53, 2011.

[323] European Parliament. Ethical considerations for clinical trials on medicinal products conducted with the paediatric population: Recommendations of the ad hoc group for the development of implementing guidelines for directive 2001/20/ec relating to good clinical practice in the conduct of clinical trials on medicinal products for human use, 2008.

[324] UNICEF. Nepal introduces typhoid vaccine into routine immunisation across the country, 2022.

[325] KP Paudel. Pcv 10 introduction in national immunization program of nepal. *Pediatric Infectious Disease*, 8(2):67–71, 2016.

[326] E Esteban, E Bujaldon, M Esparza, I Jordan, and ME Esteban. Sex differences in children with severe health conditions: Causes of admission and mortality in a pediatric intensive care unit. *Am J Hum Biol*, 27(5):613–9, 2015.

[327] A Yadav, U Shamim, V Ravi, P Devi, P Kumari, R Maurya, P Das, M Somani, S Budhiraja, B Tarai, and R Pandey. Early transcriptomic host response signatures in the serum of dengue patients provides insights into clinical pathogenesis and disease severity. *Scientific Reports*, 13(1):14170, 2023.

[328] D Habgood-Coote, C Wilson, C Shimizu, AM Barendregt, R Philipsen, R Galassini, IR Calle, L Workman, PKA Agyeman, G Ferwerda, ST Anderson, JM van den Berg, M Emonts, ED Carrol, CG Fink, R de Groot, ML Hibberd, J Kanegaye, MP Nicol, S Paulus, AJ Pollard, A Salas, F Secka, LJ Schlapbach, AH Tremoulet, M Walther, W Zenz, M Van der Flier, HJ Zar, T Kuijpers, JC Burns, F Martinón-Torres, VJ Wright, LJM Coin, AJ Cunnington, JA Herberg, M Levin, and M Kaforou. Diagnosis of childhood febrile illness using a multi-class blood rna molecular signature. *Med*, 4(9):635–654.e5, 2023.

[329] T Liu, L Zhang, D Joo, and S-C Sun. Nf-b signaling in inflammation. *Signal Transduction and Targeted Therapy*, 2(1):17023, 2017.

[330] E Avota, E Gassert, and S Schneider-Schaulies. Measles virus-induced immunosuppression: from effectors to mechanisms. *Med Microbiol Immunol*, 199(3):227–37, 2010.

[331] R Liu, AZ Holik, S Su, N Jansz, K Chen, HS Leong, ME Blewitt, ML Asselin-Labat, GK Smyth, and ME Ritchie. Why weight? modelling sample and observational level variability improves power in rna-seq analyses. *Nucleic Acids Res*, 43(15):e97, 2015.

[332] E Granda, M Urbano, P Andrés, M Corchete, A Cano, and R Velasco. Comparison of severity scales for acute bronchiolitis in real clinical practice. *European Journal of Pediatrics*, 182(4):1619–1626, 2023.

[333] AJ Justicia-Grande, J Pardo-Seco, M Cebey-López, L Vilanova-Trillo, A Gómez-Carballa, I Rivero-Calle, M Puente-Puig, C Curros-Novo, J Gómez-Rial, A Salas,

JM Martinón-Sánchez, L Redondo-Collazo, C Rodríguez-Tenreiro, and F Martinón-Torres. Development and validation of a new clinical scale for infants with acute respiratory infection: The resvinet scale. *PLoS One*, 11(6):e0157665, 2016.

[334] N Zivanovic, D Öner, Y Abraham, J McGinley, SB Drysdale, JG Wildenbeest, M Crabbe, G Vanhoof, K Thys, RS Thwaites, H Robinson, L Bont, PJM Openshaw, F Martinón-Torres, AJ Pollard, and J Aerssens. Single-cell immune profiling reveals markers of emergency myelopoiesis that distinguish severe from mild respiratory syncytial virus disease in infants. *Clin Transl Med*, 13(12):e1507, 2023.