Identifying immunological biomarkers of sepsis using cytometry bioinformatics and machine learning

Ross Burton

School of Medicine

Cardiff University

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

October 2022

Dedication

I dedicate this work to my partner Jennie, my parents, Debbie and Chris, and the friends and family that have supported me throughout the completion of this work. Jennie has been a constant source of support and inspiration, never failing to raise me up and keep me on the right track. I want to thank my parents for teaching me that hard work pays dividends. I also want to thank John, Maryam, Will, and Alex, to whom I owe a great deal of gratitude for their guidance and support. Finally, I would like to dedicate this work to Basil Burton, my grandfather. It was your curiosity and stoic nature that planted in me an appreciation for science and the natural world.

Acknowledgements

I want to thank Cardiff University and the School of Medicine who generously funded this research and made this thesis possible.

To my supervisors, Prof. Matthias Eberl, Dr Andreas Artemiou, Dr Matthew Morgan, and Prof. Peter Ghazal, I thank you for your support, encouragement, and guidance throughout my studies. The resilience and professionalism that you have shown during such unprecedented times in the face of a global pandemic have been a source of inspiration. I will forever be grateful for your commitment to the work we have produced here.

I want to thank Ms Sarah Baker and Dr Löic Raffray for your assistance with the analysis of blood samples, and Prof. Bernhard Moser, Dr Teja Rus, and Dr Adriadni Kouzeli for your valuable advice. I would also like to thank Dr Ann Kift-Morgan for your expert cytometry guidance.

Special thanks are required to Dr Simone Cuff, whose endless appetite for debate and discussion regarding the intersection of biology and statistics helped forge the work presented in this thesis. In addition, I would like to thank the staff of Cardiff University, both past and present, especially Dr Alex Greenshields-Watson and Dr Robert Andrews.

I want to thank the invaluable contributions of the intensive care unit at Cardiff and Vale Health Board and the patients recruited for this study. Without the hard work and dedication of the nurses and doctors and the generosity of the patients and their loved ones, the work discussed in this thesis would not have been possible.

Finally, I want to thank the open-source community and the kind strangers willing to contribute to the technologies that allow modern science to exist.

Abstract

Sepsis is a leading cause of mortality and significantly strains healthcare systems worldwide. Improving sepsis care and outcomes depends on appropriate risk stratification and timely identification of the causative pathogen to guide patient management and treatment. Enormous efforts have been made to identify diagnostic and prognostic biomarkers to aid decision making, but to date, they have failed to identify candidates with acceptable accuracy and precision to have an impact in the clinic. Past studies have often focused on individual biomarkers without considering the potential benefit of multi-marker panels incorporating deep immunological phenotyping. This work addressed this issue with a cross-disciplinary approach that integrated sepsis biomarker discovery, cytometry bioinformatics, and supervised machine learning.

Firstly, a novel framework for cytometry data analysis was developed, along with a new ensemble clustering algorithm that reduced the risk of biasing exploratory analyses with the application of a single clustering technique. Secondly, the analysis framework was applied to a study cohort of severe sepsis patients, and their early immunological profile consisting of cellular and humoral parameters (within 36 hours of diagnosis) was determined. The captured immunological parameters were then combined with routine clinical data and lipid plasma concentrations to generate interpretable machine learning models for predicting mortality and the underlying cause of infection. The generated models discriminated between survivors and non-survivors, and between Gram-negative and Gram-positive infections, and identified potential combinations of biomarkers with predictive value.

Keywords— Sepsis - Biomarkers - Unconventional T cells - Clustering - Cytometry - Bioinformatics - Machine learning

Publications and presentations

Publications

2022

RJ Burton, SM Cuff, MP Morgan, A Artemiou, M Eberl. GeoWaVe: Geometric median clustering with weighted voting for ensemble clustering of cytometry data. *Bioinformatics, Volume 39, Issue 1; btac751* (2022).

2021

RJ Burton, R Ahmed, SM Cuff, S Baker, A Artemiou, M Eberl. CytoPy: an autonomous cytometry analysis framework. *PLoS Computational Biology 17 (6), e1009071* (2021).

2020

MJ Ponsford*, **RJ Burton***, L Smith, PY Khan, R Andrews, SM Cuff, L Tan, M Eberl, IR Humphreys, F Babolhavaeji, A Artemiou, M Pandey, SRA Jolles, J Underwood. Examining the utility of extended laboratory panel testing in the emergency department for risk stratification of patients with COVID-19: a single-centre retrospective service evaluation. *Journal of Clinical Pathology 75 (4), 255-262* (2020).

L Raffray*, **RJ Burton***, SE Baker, MP Morgan, M Eberl. Zoledronate rescues immunosuppressed monocytes in sepsis patients. *Immunology 159 (1), 88-95* (2020).

2019

RJ Burton, M Albur, M Eberl, SM Cuff. Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. *BMC Medical Informatics and Decision Making 19 (1), 1-11* (2019).

**Authors contributed equally.*

Oral presentations

2022

A Novel Cytometry Analysis Framework in Python: Application To Real-World Immunophenotyping Of Patients With Severe Sepsis, International Society for Advancement of Cytometry Annual Conference (CYTO), Philadelphia, USA.

Poster presentations

2019

Developing a novel end-to-end cytometry data analysis pipeline., Cardiff University Annual Infection & Immunity Meeting, Cardiff, UK.

Zoledronate rescues immunosuppressed monocytes in sepsis patients., British Society for Immunology Congress, Liverpool, UK.

2018

Using AI to reduce diagnostic workload in one of the largest microbiology laboratories in England., Microbe, Sheffield, UK.

Table of Contents

1 Introduction **2**

1.1 Sepsis . . . 2

1.1.1 Definition . . . 3

1.1.2 Pathophysiology and the immune response in sepsis . . . 6

1.1.2.1 The innate immune response . . . 7

1.1.2.2 Unconventional T-cells . . . 8

1.1.2.3 Immunosuppression . . . 9

1.1.2.4 Coagulopathy and organ failure . . . 11

1.1.3 Biomarkers in sepsis . . . 12

1.1.3.1 Diagnostic biomarkers . . . 13

1.1.3.2 Prognostic biomarkers . . . 14

1.1.3.3 Biomarkers of aetiology . . . 16

1.1.3.4 Multi-parameter biomarker panels . . . 17

1.2 Cytometry bioinformatics . . . 19

1.2.1 The start of a golden age for cytometry bioinformatics . . . 19

1.2.2 The promise of automated gating . . . 22

1.2.3 A supervised approach to classifying cytometry data . . . 22

1.2.4 Clustering continues to improve . . . 24

1.2.5 Seeing is believing: how dimension reduction changed the game . . . 25

1.2.6 Moving forward . . . 26

1.3 Pattern recognition for biomarker discovery . . . 28

1.3.1 Feature selection for pattern recognition . . . 28

1.3.2 Moving beyond the ‘black box’: the promise of interpretable machine learning . . . 30

1.4 The scope of this thesis . . . 32

2 Materials and Methods **34**

2.1 Innate-like T cells in sepsis (ILTIS) study . . . 34

2.1.1 Ethics and consent . . . 34

2.1.2 Patient recruitment . . . 35

2.1.3 Sample and data collection . . . 36

2.1.4 Reagents . . . 37
2.1.5 Isolation of leukocytes, peripheral blood mononuclear cells, and cell-free plasma from whole blood . . . 37
2.1.6 Flow cytometry . . . 39
2.1.7 Luminex™ and ELISA . . . 44
2.1.8 Lipid analysis . . . 46
2.2 Patient immune responses to infection in Peritoneal dialysis (PERIT-PD) study 47
2.2.1 Ethics and consent . . . 47
2.2.2 Patient recruitment . . . 47
2.2.3 Sample and data collection . . . 48
2.2.4 Flow cytometry . . . 48
2.3 Critical assessment of population identification in cytometry data by supervised classification . . . 50
2.4 Statistical analysis . . . 52
2.4.1 Statistical hypothesis testing . . . 52
2.4.2 Statistical machine learning . . . 53

3 Development and validation of CytoPy, an open-source framework for cytometry data analysis in Python 54
3.1 Introduction . . . 54
3.2 Aims . . . 57
3.3 Results . . . 58
3.3.1 Design & implementation . . . 58
3.3.2 Identifying batch effect in blood T cell subsets . . . 62
3.3.3 Autonomous gates . . . 63
3.3.4 Autonomous gates can reliably identify T cell subsets by addressing batch-effect with hyperparameter search and landmark registration . 67
3.3.5 Addressing batch effect with the Harmony algorithm . . . 68
3.3.6 Supervised methods for classification of cytometry data . . . 70
3.3.7 Unsupervised clustering of cytometry data . . . 74
3.3.8 Implementing the CytoPy framework to identify an immune signature that differentiates patients with acute peritonitis from stable controls . . . 76

3.4 Discussion 84

4 Ensemble clustering of cytometry data 87

4.1 Introduction 87

4.2 Aims 90

4.3 Results 91

4.3.1 GeoWaVe: a novel heuristic ensemble clustering algorithm 91

4.3.2 Graph ensemble clustering methods fail to outperform individual clustering algorithms for cytometry data analysis 94

4.3.3 GeoWaVe outperforms graph ensemble methods and improves upon the performance of base clustering algorithms. 107

4.3.4 GeoWaVe outperforms graph ensemble methods for the detection of under-represented populations. 108

4.3.5 GeoWaVe is computationally efficient. 113

4.4 Discussion 116

5 Phenotypes of severe sepsis patients and their relationship with mortality and causative pathogen. 120

5.1 Introduction 120

5.2 Aims 123

5.3 Results 124

5.3.1 Characterising a cohort of acute severe sepsis patients. 124

5.3.2 Insights from routine clinical data in sepsis patients 128

5.3.3 Quantifying soluble biomarkers from plasma of sepsis patients. . . 133

5.3.4 Immune cell profiling in acute severe sepsis patients demonstrates phenotypes that correlate with mortality and causative pathogen. . . 139

5.4 Discussion 165

6 Machine learning models identify biomarker signatures correlated with mortality and causative pathogen in sepsis 172

6.1 Introduction 172

6.2 Aims 175

6.3 Preparing multi-omic clinical data for multivariate modelling 176

6.4 Imputing missing values to maximise available training data 181
6.5 Multicolinearity 186
6.6 Feature selection 188
6.7 Multivariate models identify signatures that correlate with outcome and causative pathogen 193
6.7.1 A T cell dominant signature predict mortality at 90 days after diagnosis of sepsis. 201
6.7.2 Neutrophils, $CD8^+$ T cells, and unconventional T cells form a predictive signature that differentiates Gram-negative and Gram-positive infection in sepsis. 209
6.8 Discussion 216

7 Discussion and future work **223**
7.1 The role of cytometry bioinformatics in biomarker discovery 223
7.2 The immunopathology of sepsis and the role of unconventional T cells. . . 226
7.3 Machine learning models for identifying potential biomarker combinations. 227
7.4 Sepsis heterogeneity 230
7.5 Conclusion 232

Appendix A Appendix **253**

List of Figures

1.1 Evolution of modern sepsis definitions, adapted from Gyawali *et al.* [13]. . 4
1.2 The loss of homeostasis during sepsis and septic shock as a result of pro- and anti-inflammatory response to infection 6
1.3 A timeline of developments in the field of cytometry bioinformatics between 2007 and 2022 21
1.4 Schematic of the general concept behind supervised machine learning. . . . 29

2.1 Schematic of sample processing: isolation of leukocytes, peripheral blood mononuclear cells (PBMCs), and cell-free plasma. 38
2.2 Gating strategy applied with the CytoPy software for the identification of single live T lymphocytes, conventional $CD4^+$ and $CD8^+$ subsets, and unconventional T cells. 42
2.3 Gating strategy applied with the CytoPy software for the identification of single live monocytes and neutrophils. 43

3.1 Overview of the CytoPy framework. 59
3.2 Overview of batch effect observed amongst samples from the PERIT-PD study. 64
3.3 Examples of autonomous gating strategies employed in the CytoPy framework. 66
3.4 Landmark registration can align probability density functions of the CD4 parameter. 68
3.5 Number of events captured by autonomous gates for blood T cell subsets compared to the same subsets as defined by manual expert gates. 69
3.6 Batch correction using the Harmony algorithm. 71
3.7 The CellClassifier class has convenient methods to assess the performance of supervised machine learning models. 72
3.8 Percentage of blood T cell subsets as identified by XGBoost compared to the same subsets as identified by expert manual gates. 74
3.9 Meta-clustering results for FlowSOM and Phenograph when applied to blood T cells after batch effect correction with Harmony. 76
3.10 Percentage of T cell subsets as identified by FlowSOM and Phenograph clustering, compared to the same subsets as identified by expert manual gates. . 77
3.11 Batch correction of samples taken from PERIT-PD study. 79

3.12 Leukocyte subsets in peritoneal effluent as a fraction of $CD45^+$ cells as identified by an XGBoost classifier, Phenograph clustering and FlowSOM clustering. 80
3.13 T cell subsets as a fraction of $CD3^+$ lymphocytes as identified by an XGBoost classifier, Phenograph clustering and FlowSOM clustering. 81
3.14 Feature selection process to reduce variables for predicting acute peritonitis. 83

4.1 Schematic diagram of the GeoWaVe algorithm. 93
4.2 Expression profile of the 13 parameter Levine CyTOF data (*Levine-13*) and the total number of observations for each ground-truth population. 98
4.3 Expression profile of the 32 parameter Levine CyTOF data (*Levine-32*) and the total number of observations for each ground-truth population. 99
4.4 Expression profile of the 39 parameter Samusik CyTOF data (*Samusik*) and the total number of observations for each ground-truth population. 100
4.5 Expression profile of the 28 parameter OMIP-44 spectral flow cytometry data (*OMIP*) and the total number of observations for each ground-truth population. 101
4.6 Expression profile of conventional flow cytometry data (*Sepsis* and *Peritoneal Dialysis (PD)*) and the total number of observations for each ground-truth population. 102
4.7 UMAP density plots show the topology of the six benchmark datasets for the evaluation of ensemble clustering. 103
4.8 Internal metrics for a range of final consensus clusters (k) as generated by HGPA clustering of Levine-13 data. 104
4.9 Adjusted rand index (ARI) for base clustering algorithms, graph ensemble methods, and GeoWaVe ensemble for the six benchmark datasets. 105
4.10 Fowlkes-Mallows index (FMI) and Adjusted Mutual Information (AMI) for base clustering algorithms, graph ensemble methods, and GeoWaVe ensemble for the six benchmark datasets. 106
4.11 Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), and Fowlkes-Mallows Index (FMI) when the number of consensus clusters (k) is varied for HBGF ensemble clustering. 107

4.12 F1 score of base clustering algorithms, graph ensembles and GeoWaVe ensembles, after matching cluster labels to ground-truth labels using the Hungarian linear assignment algorithm. 109
4.13 F1 score, precision, and recall of base clustering algorithms, graph ensembles and GeoWaVe ensembles, after matching cluster labels to ground-truth labels using the Hungarian linear assignment algorithm. 110
4.14 Heatmap of population F1 scores for the *Levine-13*, *Levine-32*, *Samusik*, and *OMIP* data. 111
4.15 Heatmap of population F1 scores for the *Sepsis* and *Peritoneal Dialysis (PD)* data. 112
4.16 UMAP embeddings show the distribution of 10 Gaussian 'clouds' of synthetically generated data points for GeoWaVe runtime benchmarking. . . . 115
4.17 Runtime performance of GeoWaVe algorithm on randomly generated synthetic data. 116

5.1 Overview of ILTIS study data capture. 125
5.2 Stratification of acute severe sepsis patients and available data within each subcategory. 126
5.3 C-reactive-protein (CRP) concentration in blood taken from patients diagnosed with sepsis and enrolled into the ILTIS study. 130
5.4 Lactate concentration in blood taken from patients diagnosed with sepsis and enrolled into the ILTIS study. 131
5.5 Concentration of lymphocytes, neutrophils, and monocytes in blood from patients diagnosed with sepsis and enrolled into the ILTIS study. 132
5.6 Comparisons of variables captured in routine clinical data and their ability to differentiate mortality, culture-positivity, and the Gram status of the causative pathogen in sepsis. 133
5.7 Distribution of biomarkers captured by Luminex multiplex assays before and after batch correction by $\log_2$ and z-score normalisation. 135
5.8 Concentration of soluble analytes in cell-free plasma isolated from sepsis patients, comparing survivors and non-survivors 30 days after sepsis diagnosis. 136
5.9 Concentration of soluble analytes in cell-free plasma isolated from sepsis patients, comparing survivors and non-survivors 90 days after sepsis diagnosis. 136

5.10 Concentration of soluble analytes in cell-free plasma isolated from sepsis patients, comparing those with and without a microbiologically confirmed infection. 137
5.11 Concentration of soluble analytes in cell-free plasma isolated from sepsis patients, comparing those with a Gram-positive and Gram-negative infection, amongst those with a positive bacterial culture. 138
5.12 Proportion of samples above, below, or within the detectable range of the assay used for the measurement of an analyte. 139
5.13 Odds ratios for death within 30 days of enrolement date, death within 90 days of enrolement date, culture-negative sepsis, and Gram-negative causative pathogen. 140
5.14 Comparison of the proportion of T cells, monocytes and neutrophils, and conventional and unconventional T cell subsets in survivors and non-survivors of sepsis 30 days after sepsis diagnosis. 141
5.15 Comparison of the proportion of T cells, monocytes and neutrophils, and conventional and unconventional T cell subsets in survivors and non-survivors of sepsis 90 days after sepsis diagnosis. 142
5.16 Comparison of the proportion of T cells, monocytes and neutrophils, and conventional and unconventional T cell subsets in sepsis patients with and without a microbiologically confirmed infection. 143
5.17 Comparison of the proportion of T cells, monocytes and neutrophils, and conventional and unconventional T cell subsets in sepsis patients with a Gram-positive or Gram-negative infection, where sepsis was microbiologically confirmed. 144
5.18 Before and after correction of batch effects with the Harmony algorithm as applied to single cell flow cytometry data from the ILTIS study. 145
5.19 The distribution of Local Inverse Simpson Index (LISI) for a sample of 10000 events before (blue) and after (orange) the Harmony algorithm was applied to correct batch effect. 146
5.20 UMAP scatterplots show the preservation of population structure before and after the application of the Harmony algorithm to T cells acquired in many batches. 147

5.21 Proportion of manually gated T cell populations (as a percentage of $CD3^+$ cells) before (x-axis) and after (y-axis) Harmony batch correction. 148
5.22 Before and after correction of batch effects with the Harmony algorithm, applied to $V\delta2^+$ $\gamma\delta$ T cells stained for activated subsets and stained for memory subsets, and MAIT cells stained for activation subsets and memory subsets. 149
5.23 GeoWaVe ensemble clustering of monocytes. 150
5.24 The mean fluorescence intensity (MFI) of HLA-DR, CD86, CD46, CD40 and CD62L on monocytes in sepsis. 151
5.25 GeoWaVe ensemble clustering of neutrophils in sepsis. 152
5.26 GeoWaVe ensemble clustering of T cells. 155
5.27 Proportion of GeoWaVe consensus clusters as a percentage of T cells. . . . 156
5.28 Proportion of $CD8^+$ GeoWaVe consensus clusters from PBMCs stained for the identification of memory T cell subsets, as a percentage of total $CD8^+$ T cells. 158
5.29 GeoWaVe ensemble clustering of $V\delta2^+$ $\gamma\delta$ T cells, stained for identification of memory subsets. 159
5.30 GeoWaVe ensemble clustering of $V\delta2^+$ $\gamma\delta$ T cells, stained for identification of activated subsets. 161
5.31 GeoWaVe ensemble clustering of MAIT cells ($CD3^+$ $V\alpha7.2^+$ $CD161^+$ lymphocytes), stained for identification of memory subsets. 162
5.32 GeoWaVe ensemble clustering of MAIT cells ($CD3^+$ $V\alpha7.2^+$ $CD161^+$ lymphocytes), stained for identification of activated subsets. 164
5.33 The mean fluorescence intensity (MFI) of HLA-DR, CD86, CD46, CD40 and CD62L on MAIT cells, with comparisons between sepsis patients with a Gram-positive versus a Gram-negative infection. 165

6.1 Class imbalance amongst target variables for binary classification models. . 177
6.2 Summary of steps taken to reduce the complexity of the feature space prior to development of binary classification models. 180
6.3 Visualisation of missing data in the ILTIS study. 182
6.4 Out-of-bag (OOB) imputation error estimates when imputing missing values with MissForest and MissRanger. 185

6.5 Multicolinearity amongst features visualised using pairwise Spearman's rank correlation coefficient. 186
6.6 Pairwise Jaccard Index measures the overlap of feature sets generated by five independent feature selection algorithms. 192
6.7 Schematic of the modelling pipeline for selecting, comparing, and inspecting classification algorithms. 196
6.8 Partial dependency plot showing the relationship between T cells (% of PBMCs) and the outcome of a Logistic Regression model. 199
6.9 Waterfall plots for a negative and positive prediction by a Logistic Regression model predicting 30-day mortality. 200
6.10 Cross-validation and holdout performance for the top-performing model selected within each classifier family for predicting 30-day mortality in sepsis 202
6.11 Cross-validation and holdout performance for the top-performing model selected within each classifier family for predicting 90-day mortality in sepsis 204
6.12 Complete case analysis for an Extra Random Forest model tasked with predicting 90-day mortality in sepsis. 206
6.13 SHAP (SHapely Additive exPlanations) values for an Extra Random Forest model tasked with predicting mortality at 90 days after diagnosis with sepsis. 208
6.14 Cross-validation and holdout performance for the top-performing model selected within each classifier family for predicting Gram-negative cause in sepsis. 210
6.15 Complete case analysis for a Logistic regression model and a Random Forest model tasked with predicting Gram-negative cause in sepsis. 212
6.16 SHAP (SHapely Additive exPlanations) values for a Logistic regression model tasked with predicting Gram-negative cause in sepsis. 214
6.17 SHAP (SHapely Additive exPlanations) values for a Random Forest model tasked with predicting Gram-negative cause in sepsis. 215
6.18 The proportion of V$\delta 2^+$ $\gamma\delta$ T cells plotted against corresponding SHAP (SHapely Additive exPlanations) values that explain the impact on a Random Forest model tasked with predicting Gram-negative cause in sepsis. . . 216

A.1 Summary of clusters identified from flow cytometry analysis of whole blood samples taken from patients with acute severe sepsis as part of the ILTIS study, described in full within Chapter 5. 253

List of Tables

1.1 Sequential (sepsis-related) organ failure assessment (SOFA) score, adapted from Gyawali *et al.* [13]. 5

2.1 The inclusion and exclusion criteria for recruitment into the ILTIS study. . . 35
2.2 All solution based reagents used including their constituent parts. 37
2.3 Antibody-fluorochrome cocktails applied to PBMCs for identifying subsets of T lymphocytes. 40
2.4 Antibody-fluorochrome cocktail for cell-surface staining of Leukocytes, after red cell lysis, for identifying subsets of monocytes and neutrophils. . . . 40
2.5 Cytokines and chemokines identified in cell-free plasma in this study using either Luminex™ multi-plex assays or single ELISA. 45
2.6 Summary of microbiological culture results for peritoneal dialysis patients with acute peritonitis. 48
2.7 Flow cytometry staining panel for pertioneal leukocytes. 49
2.8 Flow cytometry staining panel for T cell subsets in peritoneal mononuclear cells and PBMCs. 49

3.1 Performance of supervised classification algorithms for identifying cell populations from the FlowCAP competition data. 73

4.1 Description of the benchmark data employed for assessment of ensemble clustering algorithms. 96
4.2 Runtime performance of base clustering algorithms on benchmark data. . . 114
4.3 Runtime performance of graph ensemble clustering algorithms on benchmark data. 115
4.4 Runtime performance of GeoWaVe ensemble clustering algorithms on benchmark data, using Agglomerative hierarchical clustering. 115
4.5 Runtime performance of GeoWaVe ensemble clustering algorithms on benchmark data, using K-means, Mean shift, and Affinity Propagation. 116

5.1 Comparison of survivors and non-survivors at 30 days after diagnosis of sepsis. 127
5.2 Comparison of survivors and non-survivors at 90 days after diagnosis of sepsis. 127

5.3 Comparison of Gram-negative, Gram-positive, and unknown causative pathogen in sepsis patients. 129
5.4 Summary of GeoWaVe T cell cluster phenotypes stained for differentiating memory subtypes. 153

6.1 Holdout performance for the top-performing model selected within each classifier family for predicting 30-day mortality in sepsis. 201
6.2 Holdout performance for the top-performing model selected within each classifier family for predicting 90-day mortality in sepsis. 205
6.3 Holdout performance for the top-performing model selected within each classifier family for predicting Gram-negative cause in sepsis. 211

A.1 Description of routine clinical data available for patients diagnosed with sepsis and enrolled on the ILTIS study. 254
A.2 Description of all variables considered as potential features for machine learning models. Where a variable was later removed, a reason for exclusion is given. 257

List of Abbreviations

ADM ..... Adrenomedullin
AMI ..... Adjusted Mutual Information
APACE ..... Acute Physiology and Chronic Health Evaluation
APCs ..... Antigen Presenting Cells
API ..... Application Programming Interface
ARI ..... Adjusted Rand Index
AUC ..... Area Under Curve
CLES ..... Common Language Effect Size
CNN ..... Convolution Neural Networks
CNS ..... Central Nervous System
CRP ..... C-Reactive Protein
CSPA ..... Cluster-based Similarity Partitioning Algorithm
CV ..... Cross-Validation
DAMPs ..... Damage-Associated Molecular Patterns
DBM ..... Density-Based merging
DCs ..... Dendritic Cells (DCs)
DIC ..... Disseminated Intravascular Coagulation
EHR ..... Electronic Health Records
FCS ..... Flow Cytometry Standard
FiO2 ..... Fraction of Inspired Oxygen
FlowCAP ... Flow cytometry Critical Assessment of Population identification methods
FMI ..... Fowlkes-Mallows Index
FMO ..... Fluorescence-Minus-One
GeoWaVe ... Geometric median clustering with Weighted Voting
HBP ..... Heparin-Binding Protein
HGPA ..... Hyper-Graph Partitioning Algorithm
HLH ..... Haemophagocytic LymphoHistiocytosis
HMB-PP ... (E)-4-hydroxy-3-methyl-but-2-enyl pyrophosphate
HMGB1 .... High Mobility Group Box Protein 1
IL ..... Interleukin
ILTIS ..... Innate-Like T cells In Sepsis
LASSO ..... Least Absolute Shrinkage and Selection Operator
LISI ..... Local Inverse Simpson's Index
LOOCV .... Leave-One-Out Cross-Validation
LBP ..... Lipopolysaccharide-Binding Protein
JSON ..... JavaScript Object Notation

KNN ..... K-Nearest Neighbours

MAIT ..... Mucosal-Associated Invariant T cells

MCLA ..... Meta-CLustering Algorithm

MDR ..... Multi-Drug Resistant

MRMR ..... Minimum Redundancy - Maximum Relevance

MFI ..... Mean Fluorescence Intensity

MHC ..... Major Histocompatibility Complex

MICE ..... Multiple Imputation by Chained Equations

MIF ..... Macrophage migration Inhibitory Factor

MMD ..... Maximum Mean Discrepancy

NETs ..... Neutrophil Extracellular Traps

NFκB ..... Nuclear Factor kappa B

NLR ..... Neutrophil-to-Lymphocyte Ratio

NMRSE .... Normalised Mean Root Squared Error

OOB ..... Out-Of-Bag error

OSM ..... Oncostatin M

PAMPs ..... Pathogen-Associated Molecular Patterns

PBMCs ..... Peripheral Blood Mononuclear Cells

PCA ..... Principal Component Analysis

PCT ..... Procalcitonin

PD ..... Peritoneal Dialysis

pDCs ..... Plasmacytoid Dendritic Cells

PERIT-PD .. Patient immune responses to infection in Peritoneal Dialysis

PHATE ..... Potential of Heat-diffusion for Affinity-based Trajectory Embedding

PMM ..... Predictive Mean Matching

PyPI ..... Python Package Index

RBA ..... Relief-Based Algorithms

RFE ..... recursive feature elimination

ROC ..... Receiver Operating Characteristic

ROS ..... Reactive Oxygen Species

scRNA-seq .. single-cell RNA sequencing

SHAP ..... SHapley Additive exPlanations

SOFA ..... Sequential Organ Failure Assessment

SOM ..... Self-Organising Map

SIRS ..... Systemic Inflammatory Response Syndrome

SVM ..... Support Vector Machine

TCR ..... T Cell antigen Receptor

TLRs ..... Toll-Like Receptors

TNF-α ..... Tumor Necrosis Factor-α

t-SNE ..... t-distributed Stochastic Neighbour Embedding

TRAIL ..... TNF-Related Apoptosis-Inducing Ligand

VIF ..... Variance Inflation Factor

UMAP ..... Uniform Manifold Approximation and Projection

XGBoost .... Extreme Gradient Boosting

1 | Introduction

1.1 Sepsis

Sepsis is a devastating disease with high mortality and lasting effects on individuals, their families, and their communities. Within the UK, a rising incidence of sepsis was observed before the COVID-19 pandemic, and conservative estimates suggested that over 46,000 deaths were attributed to sepsis in 2018 alone. Sepsis-related deaths in 2018 exceeded the estimated number of deaths from breast, prostate, and bowel cancer combined [1]. Internationally, 11 million sepsis-related deaths were estimated to have occurred in 2017, with an in-hospital mortality rate of 27%, increasing to 47% in intensive care patients. Among those who survive sepsis, one in three will die within a year, and one in six will experience significant long-term morbidity [3, 2]. During the coronavirus disease 2019 (COVID-19) pandemic, respiratory failure, septic shock, or multiple organ dysfunction were observed in approximately 5% of symptomatic patients. Clinically these patients met the criteria for sepsis and had high mortality rates, reflecting the burden of sepsis during the global pandemic [4]. – The financial burden of sepsis is significant, with an estimated cost of £15.6 billion per year for the UK economy [5]. In the United States, sepsis is ranked as the most expensive condition to treat, with an aggregated cost of $24 billion in 2013, amounting to 6.2% of all hospitalisation costs in that year [6]. Compared to non-sepsis admissions, survivors of sepsis experience a greater risk of re-hospitalisation, increased risk of infections, a higher prevalence of mental health issues, and a 3-fold increase in the prevalence of moderate to severe cognitive impairment [7, 8]. The range of symptoms that can arise following sepsis has been termed ‘post-sepsis-syndrome’ and is associated with a decline in quality of life [9]. Despite the devastating impact, historically, the public and professionals have a poor understanding of sepsis, and the past decade has seen many efforts to improve awareness. Examples include the Global Sepsis Alliance [10] and the launch of ‘World Sepsis Day’, an annual event running since 2011, the establishment of the UK Sepsis Trust in 2012 [11], and a global initiative from the World Health Assembly to strengthen efforts to recognise, prevent, and treat sepsis [12]. In the face of growing efforts to tackle sepsis, the incidence continues to rise, and there has been little success in reducing mortality [2, 3]. To make significant progress in improving clinical outcomes, translational research must continue to

push the boundaries of our understanding, and this thesis will help contribute to the field of biomarker research for sepsis care.

1.1.1 Definition

Sepsis has been recognised since antiquity, appearing in the Greek poems of Homer with the word "sepo", meaning "I rot.", and discussed amongst the writings of Hippocrates, thought to be the first use of the term "sepsis". The 1800s saw the establishment of 'germ theory' through the work of Koch and Pasteur, and simultaneously 'antiseptic' techniques were pioneered by clinicians such as Semmelweiss and Lister. The first modern definition of sepsis dates back to 1914 from the writings of Hugo Schottmüller, linking the condition to persistent or transient bacteraemia. Research continued through the 20th century, revealing the role of the coagulation system, cytokines, and nitric oxide in the pathophysiology of sepsis. However, it was not until 1991 that the international community agreed on a formal definition of sepsis [14, 13]. The ACCP/SCCM consensus conference committee, recognising the importance of the host immune response to sepsis, defined systemic inflammatory response syndrome (SIRS) and sepsis as a "systemic response to infection, manifested by two or more of the SIRS criteria as a result of infection" [13]. Subsequent advancements in the understanding of cell biology, biochemistry, and the immune response have led to repeated revisions of the sepsis definition (Figure 1.1), culminating in the most recent formal definition, "Sepsis-3". Under Sepsis-3, sepsis is defined as "a life-threatening organ dysfunction caused by the dysregulated host response to infection", accompanied by the sequential (sepsis-related) organ failure assessment (SOFA) score to assist identification of sepsis (Table 1.1). Clinical criteria for sepsis under Sepsis-3 are defined as "suspected or documented infection with an acute increase of $\geq$2 SOFA points". The work presented in this thesis adopts the Sepsis-3 criteria for the recruitment and study of the immunopathology of sepsis.

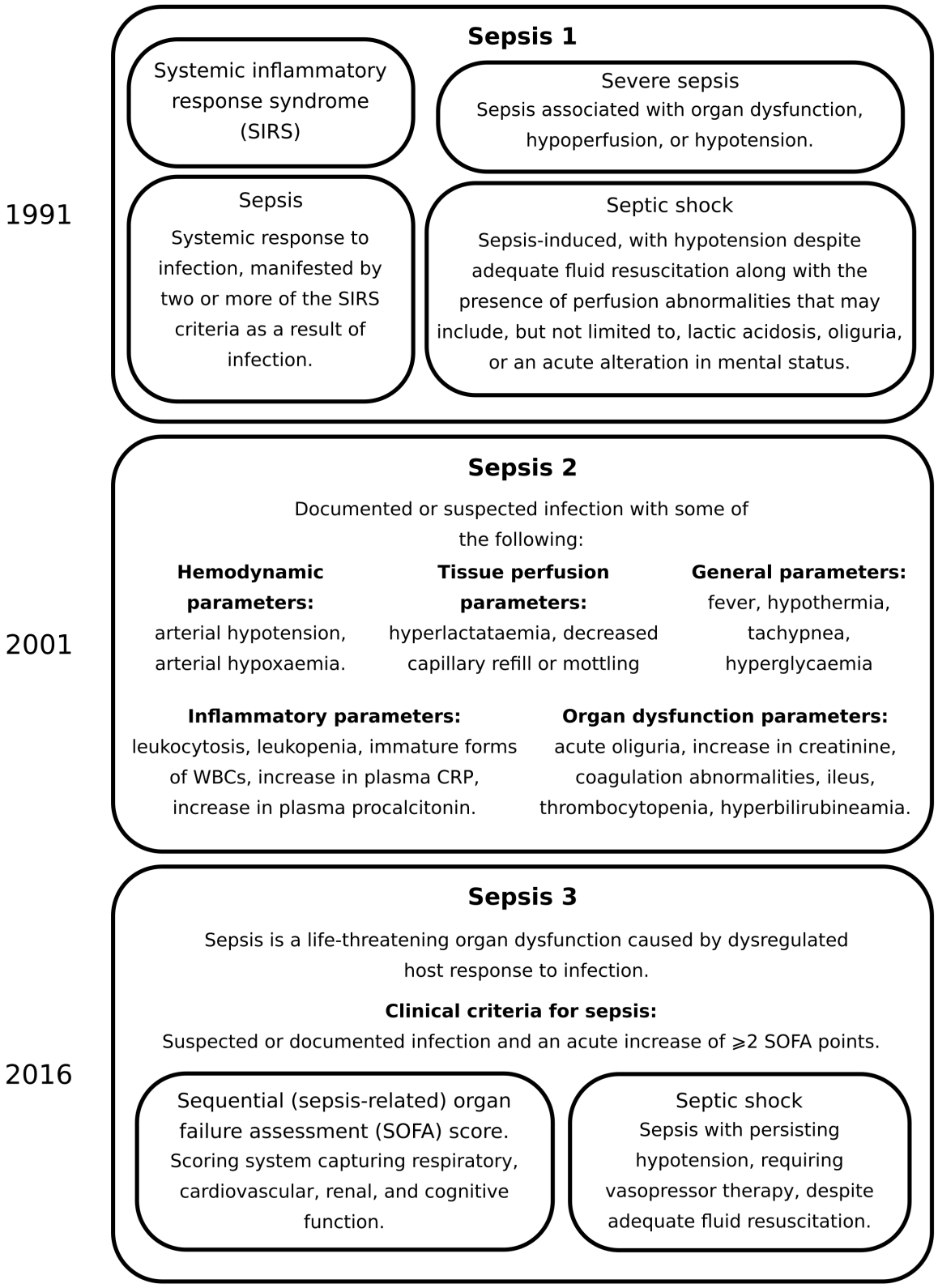

Figure 1.1: Evolution of modern sepsis definitions, adapted from Gyawali *et al.* [13].

System	Score				
	0	1	2	3	4
Respiratory					
PaO_2/FIO_2, mmHg (kPa)	≥400 (53.3)	<400 (53.3)	<300 (40)	<200 (26.7) with respiratory support	<100 (13.3) with respiratory support
Coagulation					
Platelets, $\times 10^3\ \mu L^{-1}$	≥150	<150	<100	<50	<20
Liver					
Bilirubin, mg dL^{-1} (μmol L^{-1})	<1.2 (20)	1.2–1.9 (20–32)	2.0–5.9 (33–101)	6.0–11.9 (102–204)	>12.0 (204)
Cardiovascular	MAP ≥ 70 mmHg	MAP < 70 mmHg	Dopamine < 5 or dobutamine (any dose)*	Dopamine 5.1–15 or epinephrine ≤ 0.1 or norepinephrine ≤	Dopamine > 15 or epinephrine > 0.1 or norepinephrine > 0.1*
Central Nervous System (CNS)					
Glasgow Coma Scale score†	15	13–14	10–12	6–9	<6
Renal					
Creatinine, mg dL^{-1} (μmol L^{-1})	<1.2 (110)	1.2–1.9 (110–170)	2.0–3.4 (171–299)	3.5–4.9 (300–440)	>5.0 (440)
Urine output, mL per day				<500	<200

Table 1.1: Sequential (sepsis-related) organ failure assessment (SOFA) score, adapted from Gyawali *et al.* [13].
PaO_2: partial pressure of oxygen; FIO_2: fraction of inspired oxygen; MAP: mean arterial pressure.
*Catecholamine doses are given as μg kg^{-1} min^{-1} for at least one hour.
†Glasgow Coma Scale scores range from 3 to 15; a higher score indicates better neurological function.

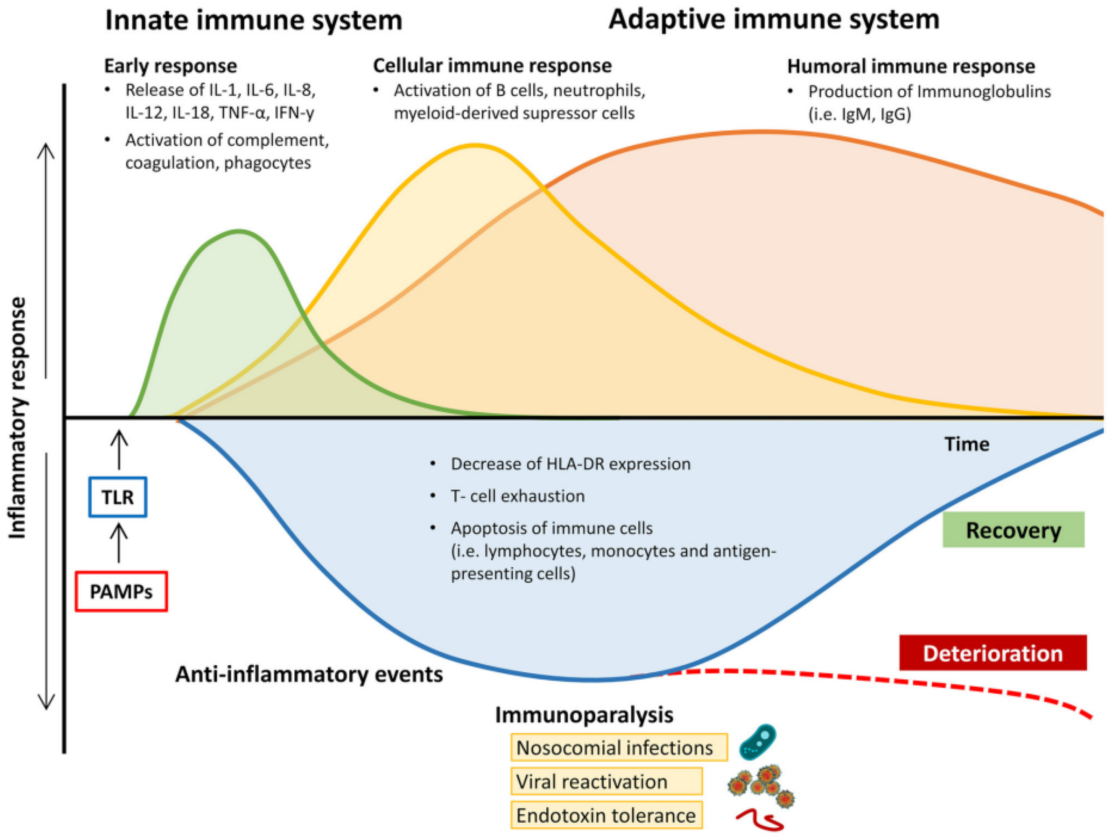

Figure 1.2: The loss of homeostatis during sepsis and septic shock as a result of pro- and anti-inflammatory response to infection; figure taken from [15]. HLA-DR, human leukocyte antigen-D related; IgM/G, immunoglobulin M/G; IL, interleukin; IFN-γ, Interferon γ; PAMPs, pathogen-associated molecular patterns; TNF-α, tumor necrosis factor alpha; TLR, Toll-like receptor.

1.1.2 Pathophysiology and the immune response in sepsis

Our understanding of the molecular pathobiology of sepsis has evolved considerably in recent decades. Once thought to be primarily a hyperimmune response to infection, it is now understood to be a complex dysregulation of the immune response involving both inflammatory and immunosuppressive mechanisms, resulting in a loss of homeostasis [13]. Figure 1.2, adapted from [15], describes the temporally dynamic state of sepsis resulting from the induction of both pro- and anti-inflammatory signalling pathways. Complex molecular cascades drive the loss of homeostasis, partially depending on the virulence factors of the causative pathogen and any pre-existing inflammatory or non-inflammatory co-morbidities, resulting in an immunological profile that remains highly individualised and difficult to diagnose and manage [13, 15, 16].

1.1.2.1 The innate immune response

The series of events in the pathobiology of sepsis starts with the activation of the innate immune response, involving chemical mediators, the complement system, and specialised cellular compartments. Examples of the main innate immune cell populations include monocytes, neutrophils, and natural killer cells. Monocytes migrate from the blood into the site of inflammation, differentiating into either macrophages or dendritic cells (DCs). These cells phagocytose pathogens and particles and are often termed 'professional' antigen-presenting cells (APCs) due to their ability to process and present elements of antigen on their surface along with a class II major histocompatibility complex (MHC). They are also potent producers of specialised molecular signals called 'cytokines'. Another important phagocyte is neutrophils, known (along with eosinophils and basophils) as polymorphonuclear cells (PMNs) due to their distinct lobed nuclei and granulocytes due to the presence of granules in their cytoplasm. Neutrophils are crucial for microbial clearance through phagocytosis, oxidant generation, and the release of neutrophil extracellular traps (NETs), networks of chromatin fibres containing granules of antimicrobial peptides and enzymes [18, 16, 17]. Natural killer (NK) cells are another lymphocyte involved in the innate immune response with both cytotoxicity and cytokine-producing effector functions. NK cells detect ligands on cells in 'distress' and recognise the absence of constitutively expressed 'self' molecules on target cells. In response, they act to destroy compromised host cells and have an important role in the innate immune response to viruses and intracellular bacteria [19].

Activation of the innate immune system occurs as a response to pathogen-associated molecular patterns (PAMPs) such as bacterial exo- and endotoxins, fungal β-glucans, viral DNA/RNA, or in response to host-derived damage-associated molecular patterns (DAMPs), such as ATP, high mobility group box protein 1 (HMGB1), or mitochondrial DNA [13, 15, 16]. PAMPs or DAMPs bind pattern recognition receptors on antigen-presenting cells (APCs) and some epithelial cells. Examples of pattern recognition receptors include Toll-like receptors (TLRs) and C-type leptin receptors found on the surface of innate immune cells, or NOD-like receptors and RIG-1-like receptors found in the cytosol. NOD-like receptor groups can aggregate into larger protein complexes called inflammasomes, which are involved in producing potent cytokines such as IL-1β and IL-18, as well as caspases involved in pyroptosis, a specific form of cell death triggered by proinflammatory signals [13, 15].

The binding of pattern recognition receptors results in up-regulation of early response genes and the release of type-1 interferons and pro-inflammatory cytokines such as TNFα, IL-1β, IL-5, and IL-12, resulting in the activation and proliferation of leukocytes, activation of the complement system, upregulation of adhesion molecules on endothelial cells, and chemokine expression [13, 15, 16].

In sepsis, neutrophil migration is impaired through increased nitric oxide production, known to inhibit neutrophil migration through binding of β2-integrins and selectins [18]. Additionally, impaired recognition of pathogens and reduced antimicrobial functionality of sepsis neutrophils has been reported [20]. Down-regulation of CXCL12 during sepsis leads to a large release of both mature and immature forms of neutrophils from the bone marrow through emergency granulocyte maturation [15]. Immature neutrophils (often termed "band cells") show reduced phagocytosis, oxidative burst capacity, and greater resistance to spontaneous apoptosis [18, 21, 15]. Excessive quantities of immature neutrophils in the peripheral blood of sepsis patients have been associated with worse outcomes [22].

1.1.2.2 Unconventional T-cells

Independent of the mechanisms above are unconventional T cells with innate-like capacity. $\gamma\delta$ T cells expressing a Vγ9Vδ2 receptor are unique to humans and primates and have been shown to expand dramatically in response to infection [23, 24]. They differ from 'conventional' $\alpha\beta$ T cells in their ability to be directly activated by the microbial metabolite (E)-4-hydroxy-3-methyl-but-2-enyl pyrophosphate (HMB-PP), an essential metabolite in most Gram-positive and Gram-negative bacteria [25]. Patients with acute sepsis with a confirmed infection caused by an HMB-PP-producing pathogen are shown to have elevated levels of circulating activated Vγ9Vδ2 T cells [23]. Activated Vγ9Vδ2 T cells produce cytotoxic mediators and interact with monocytes leading to the rapid production of TNF-α and IL-6. Additionally, HMB-PP stimulated monocyte-$\gamma\delta$ T cell co-cultures, compared to controls, displayed increased quantities of chemokines that target monocytes and neutrophils such as CXCL8 (IL-8) and CXCL10 (IP-10), important in the recruitment of effector T cells. Vγ9Vδ2 T cells can also promote the generation of mature dendritic cells via a TNF-dependent mechanism [26]. It is also now widely recognised that circulating V$\delta2^+$ T cells can display flexible APC functions and provide co-stimulatory signals that stimulate $\alpha\beta$ T cell proliferation and differentiation [27].

Mucosal-associated invariant T (MAIT) cells are an innate-like population of T cells, remarkably abundant in human tissues, and characterised by a semi-invariant T cell antigen receptor (TCR) with specificity for microbial riboflavin-derivative antigens presented by HLA-1b major histocompatibility complex (MHC)-related protein 1 (MR1) [28, 24, 29]. MAIT cells exhibit specificity towards microbial vitamin B metabolites [30], have an intrinsic effector-memory phenotype, and are capable of rapidly secreting several pro-inflammatory cytokines [24]. MAIT-deficient mice in experimental sepsis models demonstrated greater mortality and bacterial load. Additionally, MAIT cells isolated from sepsis patients within 48 hours of ICU admission showed a reduced abundance of circulating MAIT cells compared to healthy controls and reduced capacity for IFN-γ production [31].

1.1.2.3 Immunosuppression

Anti-inflammatory cytokine pathways are activated even in the first hours of severe sepsis. Eventually, the inflammatory state is superseded by a prolonged state of immunosuppression driven by overwhelming anti-inflammatory mechanisms (Figure 1.2). Interleukin 10 (IL-10), a cytokine produced by a variety of leukocytes such as $CD4^+$ Th2 cells, monocytes, and B cells, suppresses the production of the pro-inflammatory cytokines IL-6 and IFN-γ and stimulates the production of soluble TNF receptor and IL-1 receptor antagonists. The effect is neutralisation of potent TNF-α and IL-1 signalling [32]. In sepsis, it has been reported that IL-10 production is significantly increased, and concentrations in serum correlate with severe outcomes and mortality [33]. Another anti-inflammatory cytokine indicated in sepsis immunosuppression is transforming growth factor beta (TGF-β). TGF-β regulates several different immune cells and, importantly, is necessary for the induction of thymic and peripheral regulatory T cells. TGF-β levels in plasma have been reported to be increased in sepsis and associated with adverse outcomes [34].

A well-documented phenomenon in sepsis is T cell depletion, largely a result of T cell apoptosis [16, 13, 15]. Postmortem studies of patients that succumbed to sepsis demonstrated a global depletion of $CD4^+$ and $CD8^+$ T cells, most notably in the lymphoid organs such as the spleen, and the remaining splenocytes also showed a reduced capacity for cytokine stimulation. Amongst the remaining T cell population, exhaustion and functional defects of sepsis T cells were reported [13]. Post-septic $CD4^+$ T cells exhibit a global state of anergy and decreased ability to proliferate. There are also marked changes in the composition of

$CD4^+$ T cell subsets, with a pronounced increase in circulating Treg cells [35]. The loss of T cell populations and suppression of functionality ultimately result in a condition that severely affects the patient's ability to combat secondary infections.

A hallmark of immunosuppression in sepsis is reduced expression of the major histocompatibility complex (MHC) class II cell molecule HLA-DR on the surface of circulating monocytes. The family MHC molecules consist of specialised host-cell glycoproteins responsible for delivering antigens to the cell surface for recognition by T cell receptors (TCRs). The TCRs of $\alpha\beta$ T cells respond to short, continuous amino acid sequences, often buried within the native structure of the target protein. Consequently, the processing and presentation of antigens by MHC molecules are vital for antigen recognition. MHC class II molecules differ from MHC class I molecules in their peptide-binding cleft, and whilst MHC class I molecules are recognised by $CD8^+$ T cells, MHC class II molecules are recognised by $CD4^+$ T cells [17]. Expression of HLA-DR on APCs is a sign of immune competence, and the decreased expression of HLA-DR on septic monocytes reflects a reduced antigen presentation capacity. This disrupts the Th1- and Th2-mediated response, and the inability to restore HLA-DR expression is associated with endotoxin tolerance in the early stages of sepsis [15]. Now widely accepted as a marker of severity in sepsis and risk of secondary infection [36], monocyte HLA-DR (mHLA-DR) is being adopted for monitoring sepsis patients in the clinic [38, 37] and is being implemented as a screening tool for enrolment in sepsis clinical trials [39]. Another key feature of septic monocytes is immune reprogramming, and the impaired capacity to produce pro-inflammatory cytokines [40, 41]. For example, Reyes *et al.* described a $CD14^+$ monocyte population they named MS1, which displayed an immunosuppressive phenotype and, upon *ex vivo* stimulation with LPS, were unable to activate NF-κB and produce TNF-α [42].

Our understanding of the immunosuppressive phenotype observed in sepsis has been historically driven by the application of cytometry to characterise immune populations in the days and weeks that follow the initial hyperinflammatory response. As previously mentioned, mHLA-DR is progressively becoming a popular immunomonitoring tool, but flow cytometry is also being adopted to characterise immature neutrophil subsets, myeloid-derived suppressor cells, and alterations in regulatory lymphocytes with implications in the clinic [38]. It is worth noting, however, that these efforts often focus on quantifying the abundance of cell populations that are progressing through state changes in a dynamic system. Therefore

the effectiveness of such monitoring will be highly dependent on the patient's progression and their current immune state, something that is subject to natural variation. Technological advances in the field of single-cell RNA sequencing (scRNA-seq) and cell trajectory analysis could arise as a solution to this problem, allowing for more accurate identification of the precise course of a patients septic pathway [42, 43].

1.1.2.4 Coagulopathy and organ failure

Dysregulation of the coagulation pathway and the profound alterations to the endothelium have multifactorial aetiology, but ultimately contribute to tissue damage and subsequent organ failure. A study of 1895 patients from Japan emphasised the extent of coagulopathy, showing that 29% of sepsis patients were diagnosed with sepsis-induced coagulopathy, with a clinical picture synonymous with disseminated intravascular coagulation (DIC) [44].

Hypercoagulation is driven by the release of tissue factors from damaged endothelial cells, resulting in systemic activation of the coagulation cascade. The process is exaggerated by the collateral damage caused by a hyperinflammatory innate immune response, resulting in the release of reactive oxygen species (ROS), such as the hydroxyl radical and nitric oxide, which can damage cellular proteins, lipids, and DNA. Activation of the complement system further increases the generation of ROS. The release of intravascular tissue factor, combined with NETs and tissue factor expression by monocytes in the blood, results in 'immunothrombosis' trapping invading pathogens and attracting activated leukocytes and can impair microvascular function and cause organ injury [16, 15].

There is also a decrease in plasma levels of protein C and antithrombin in sepsis. Activated protein C has potent anti-inflammatory effects by inhibiting pro-inflammatory cytokines such as TNF-α and IL-6 and by limiting endothelium adhesion of monocytes and neutrophils. Reduction in protein C levels causes failure to control the coagulation cascade. Simultaneously, a reduction in fibrinolysis is observed, resulting from increased TNF-α and IL1β production that induce the release of tissue plasminogen activators and subsequently increases the production of plasminogen activator inhibitor type 1 (PAI-1). The cascading effects are a reduction in fibrinolysis and fibrin removal, further worsening the effects of hypercoagulation [16, 13].

DIC, tissue damage, and multiple organ dysfunction are some of the most devastating effects of sepsis. They are responsible for high mortality and a reduction in the quality of life for those that survive [45, 16]. The resulting tissue damage from DIC causes a breakdown of endothelial and epithelial barriers in the lungs, gastrointestinal tract, and liver. The increased permeability of these tissues causes more systemic inflammation and further exacerbates the condition.

1.1.3 Biomarkers in sepsis

With a greater understanding of the pathophysiology comes the hope of diagnostic and prognostic biomarkers that will help identify sepsis earlier, offer personalised care and triage patients to maximise hospital resources, and identify the underlying cause to reduce the risk of empiric broad-spectrum antimicrobial therapy. The primary objective of this thesis was to conduct a broad observational study of patients with sepsis and leverage immunophenotyping combined with statistical pattern recognition techniques to identify informative biomarkers that correlate with the outcome and underlying cause. In this section, a summary of the existing body of work around identifying biomarkers for sepsis will be discussed. The current state of sepsis biomarker research will be broadly summarised into diagnostic biomarkers, prognostic biomarkers that correlate with mortality or increased hospital stay, and finally, those biomarkers that indicate the aetiology of the disease.

Throughout this summary, reference will be made to the area under the receiver operating characteristic curve (AUC) score. It is common for biomarkers to be reported according to their AUC score, which captures the relationship between the false positive rate (one minus the specificity) and the true positive rate (sensitivity). An AUC of 0.5 indicates that a biomarker is no better than a random classifier. In contrast, a higher AUC indicates that the biomarker has both good sensitivity and specificity, with a maximum score of 1.0 representing a perfect biomarker. It is generally accepted that an AUC score of between 0.7 and 0.8 is fair, between 0.8 and 0.9 is good, whereas an AUC greater than 0.9 is ideal and is suggestive of a very promising biomarker [46].

1.1.3.1 Diagnostic biomarkers

The positive acute-phase proteins (hepatic derived inflammatory mediators), C-reactive protein (CRP) and procalcitonin (PCT) are the two most widely studied biomarkers in severe infectious disease and sepsis. They are actively used in a clinical setting. CRP is a non-specific marker of inflammation and although it retains a high sensitivity for identifying bacterial infection, its low specificity makes it unsuitable for differentiating infection from noninfectious causes of inflammation [47]. PCT concentrations in serum increase significantly in the first hours of a bacterial infection [48]. Although it has been previously suggested as a diagnostic biomarker in sepsis, revised recommendations suggest it only has prognostic value [49].

Other than acute-phase proteins, pro-inflammatory cytokines have been suggested as possible diagnostic biomarkers. The most widely studied one for its diagnostic potential in sepsis is IL-6. Experimental models of sepsis in mice demonstrated that IL-6 could be an early marker of inflammation due to infection, albeit with poor sensitivity despite the promising specificity [50]. Subsequent human studies showed that IL-6 could discriminate sepsis from healthy controls with an AUC score of between 0.83 and 0.94. IL-6 could also differentiate septic shock from sepsis, albiet with a slightly reduced AUC score of between 0.71 and 0.89 [51]. In a multi-centre observational study of 306 patients presenting with suspected infectious illness, IL-6 was employed to predict those with confirmed infection, resulting in an AUC score of 0.71 [52].

Along with soluble components of the innate immune response, the expression of activation markers on cellular components has also been of interest. The most promising of these for sepsis diagnosis is CD64 expression of neutrophils. CD64, also known as Fc-γ receptor 1 (FcγR1), binds IgG with high affinity and increased expression on neutrophils is considered an early activation marker. A prospective observational study of over 500 patients found that amongst the 103 diagnosed with sepsis, there was a higher expression of CD64 on neutrophils upon hospital admission. Neutrophil CD64 mean fluorescence intensity (MFI) was able to identify sepsis with a sensitivity of 89% and specificity of 87% [53]. Such evidence has encouraged some authors to suggest that CD64 expression of neutrophils should be considered a diagnostic biomarker, arguing that superior sensitivity and specificity compared to biomarkers such as CRP warrants adoption [54].

During acute inflammation, the N-terminus of CD14 is cleaved and secreted as soluble CD14 (sCD14), also known as 'presepsin'. sCD14 is thought to play a role in bacterial phagocytosis and lysosomal cleavage of invading pathogens and could form another immune-cell-derived diagnostic biomarker. Notably, sCD14 levels in plasma are elevated before those of PCT or IL-6. A multi-centre study of 207 suspected sepsis patients found that sCD14 had a greater AUC score than IL-6 and procalcitonin for predicting sepsis [55]. Other studies have reported sCD14 as significantly different between infected and non-infected groups with a diagnostic accuracy greater than PCT, IL-6, and high-sensitivity CRP [56]. Multiple meta-analyses highlighted the potential of this sCD14 as a diagnostic biomarker [57, 58].

1.1.3.2 Prognostic biomarkers

As with diagnostic biomarkers, the acute-phase proteins CRP and PCT have been the most popular candidates for prognostic biomarkers. Lee *et al.* conducted a study of over 500 patients admitted to the emergency department with suspected sepsis and measured their admissions biomarkers to study their ability to identify early (within five days of admission) or late (between 6 to 30 days after admission) mortality [59]. They found that the AUC score for levels of CRP in plasma was relatively poor, with an AUC score of 0.68 for early mortality and 0.63 for late mortality. PCT levels showed slightly better performance with an AUC score of 0.76 for early mortality and 0.70 for late mortality. On the other hand, multivariate analysis that included age, SOFA score, and PCT, suggested that PCT plasma levels are an acceptable prognostic biomarker with a favourable odds ratio of 2.004 and initial PCT plasma level was significantly higher within the group of non-survivors compared to survivors [49]. A meta-analysis from 2015 supports this finding, reporting that plasma levels of PCT were significantly lower in the early stages of sepsis amongst survivors but found this difference was lower when observing severe sepsis and septic shock. Conclusive evidence was also hindered by considerable heterogeneity amongst the studies investigated [60].

Another acute-phase protein with potential is pentraxin-3 (PTX-3), which is expressed by various cells of the innate immune system, such as dendritic cells, monocytes, and neutrophils, in response to IL-6, TNF-α, IL-1, and interferons. PTX-3 plasma level had the highest AUC of 0.798 amongst diagnostic biomarkers when comparing sepsis to a healthy group [61, 62] and significantly correlated with the degree of organ dysfunctions [63]. PTX-3 plasma level has also shown a reasonable AUC score of 0.78 for predicting 28-day mortality,

greater than both procalcitonin and lactate levels [64], and both initial and subsequent PTX-3 plasma levels measured during ICU stay were significantly higher in non-survivors [51].

A study of multiple acute phase proteins and cytokines in 47 critically ill patients sampled within 24 hours of their sepsis diagnosis found that CD64 expression on circulating neutrophils and CXCL8 (IL-8) levels in plasma were the only biomarkers that could differentiate sepsis, severe sepsis, and septic shock. CXCL8 and neutrophil CD64 were also significantly associated with 28-day mortality in a multivariate logistic regression analysis, and neutrophil CD64, CXCL8, and IL-6 correlated with Acute Physiology and Chronic Health Evaluation II (APACHE-II) severity score [65]. IL-6 plasma concentration, as well as neutrophil-to-lymphocyte (NLR) ratio, correlate with APACHE II and SOFA scores. In cox-regression models, IL-6 and NLR predicted mortality with an odds ratio of 1.017 and 1.281, respectively [66]. Other potential prognostic cytokines include the family of IL-1, which play an essential role in immune regulation and inflammatory response. Excessive production of IL-1 cytokines has been linked to hypotension, shock, and multi-organ failure in sepsis, SIRS, and septic shock [67, 68].

sCD14 was identified as a possible prognostic biomarker in the Albumin Italian Outcome Sepsis (ALBIOS) trial [69], which enrolled 997 patients with severe sepsis or septic shock and randomised treatment with albumin or crystalloids. They found that baseline sCD14 positively correlated with SOFA score and frequency of organ dysfunction, and increasing concentrations of sCD14 from day 1 to day 2 predicted 90-day mortality [69].

Monocyte HLA-DR has been cited in over 200 publications for its potential as a prognostic biomarker and has been adopted in the clinic as an indication of increased mortality and risk of secondary infections [70]. HLA-DR expression on monocytes has seen the greatest success of all the biomarkers studied for their prognostic potential. Other worthy mentions with growing evidence of their application in sepsis prognosis include adrenomedullin (ADM), TNF-related apoptosis-inducing ligand (TRAIL), and heparin-binding protein (HBP). ADMs are produced mainly by endothelial cells and vascular smooth muscle cells and help mediate vasodilation and systemic circulation. Mid-regional proadrenomedullin (MR-proADM) has been identified in several studies as a predictor of mortality in sepsis and septic shock [71]. TRAIL helps regulate the immune response in sepsis by inducing apoptosis of activated cells. In three independent cohorts of critical care patients, lower concentrations of TRAIL

in plasma were associated with increased mortality [72]. Finally, HBP has been linked to neutrophil-derived induction of vascular leakage and is a potential marker for severe outcomes in sepsis. A study of over 500 emergency department patients found that HBP was a good predictor of either admittance to ICU, death, or persistently high SOFA scores, with an AUC of 0.87 (95% CI 0.77–0.99).

1.1.3.3 Biomarkers of aetiology

The early identification of the causative pathogen is important for directing therapy and source control. The current gold standard for identifying causative pathogen is bacterial culture, which can take days to yield a positive result [73], culture conditions and low bacterial load can negatively impact the quality of results, and bacterial culture has reduced specificity when anti-microbial treatments are employed before sampling [74, 75, 76]. The percentage of suspected sepsis yielding negative culture results can range from 28 to 89% [77]. Although molecular techniques offer the potential for rapid identification of the causative pathogen, many proposed solutions either require some bacterial growth on culture media or are expensive and require technical expertise [73]. Early intervention with empiric broad-spectrum antibiotics is recommended in treating suspected sepsis [76] with the risk of mortality increasing with each hour that antibiotics are delayed [78]. However, initial antibiotics were found to be inappropriate in up to a third of patients diagnosed with sepsis and were associated with an increased likelihood of mortality and more extended hospital stay [79, 80]. Furthermore, inappropriate antibiotic use is of concern as the incidence of multi-drug resistant (MDR) pathogens continues to grow [81].

To this end, biomarkers that could reliably identify causative pathogens in sepsis prior to lengthy bacterial culture would be of great value. The causative pathogen in sepsis results in distinct molecular characteristics in the immune response [82] and therefore, one can theorise that a host-derived biomarker (or set of biomarkers) might exist that correlates with the underlying cause.

PCT plasma levels have been associated with differentiating bacterial infection from other sources of inflammation [84, 83] and as a possible biomarker for Gram-negative bacteremia [87, 83, 85, 86]. PCT levels have also been suggested as a means to guide antibiotic de-

escalation, however, a recent systematic review and meta-analysis warned against successful claims of this application given low certainty in the evidence presented [88].

In a study of immunocompromised patients, PCT in plasma, combined with CRP and sCD14, was found to accurately identify invasive fungal infections with an AUC score of 0.962 (95% CI 0.868 to 0.995) [89]. Admission sCD14 plasma levels in patients with suspected sepsis were significantly lower in those with a nonbacterial infectious disease compared to those with a confirmed bacterial pathogen [55]. sCD14 levels were also found to decrease during ICU stay in patients with negative blood cultures and those with positive blood cultures and appropriate antibiotic therapy [69].

There have been mixed reports regarding the clinical value of cytokines for identifying bacterial infections. Oever *et al.* [90] reported that lipopolysaccharide-binding protein (LBP), PCT, IL-6, IL-18, or soluble triggering receptor expressed on myeloid cells-1 (sTREM-1) combined with CRP offered no additional improvement in differentiating bacterial and viral infection amongst emergency department admissions with suspected sepsis. Meanwhile, a prospective study comparing Gram-negative, Gram-positive, and fungal bloodstream infections found utility in IL-3 plasma levels for identifying Gram-positive infections [91] and a study of 132 patients with fever found that IL-6 and CXCL8 levels were significantly higher in bacterial infection compared to viral infection [92].

LBP, an acute-phase protein, has also been suggested as a diagnostic biomarker for infection. Although LBP levels in plasma were found to be significantly higher in patients with infectious endocarditis compared to noninfectious heart valve diseases [93] and was increased in bacterial gastroenteritis compared to viral cause [94], it failed to demonstrate diagnostic value in post-operative sepsis patients [95].

1.1.3.4 Multi-parameter biomarker panels

Despite the progress in the study of biomarkers in sepsis, no biomarker has been approved with specific application to sepsis diagnosis, and the accurate prognosis is still a challenging task [76]. A recent review found that many studies had considerable limitations. Only 26 biomarkers had been evaluated in populations of greater than 300 patients, and the definition for sepsis varied greatly [96]. The same review also identified that the number of new biomarkers discovered has decreased despite the increase in studies dedicated to iden-

tifying biomarkers for sepsis. Therefore, new strategies must be explored, including an increased effort in exploring multi-parameter panels that combine the benefits of the individual biomarkers whilst offering flexibility to account for the heterogeneity in sepsis.

Some studies have attempted to combine multiple markers, starting with Kofoed *et al.* [97] who showed a linear combination of soluble urokinase-type plasminogen activator (suPAR), sTREM-1, macrophage migration inhibitory factor (MIF), CRP, PCT, and neutrophil count produced a desirable AUC score of 0.88 (95% CI 0.81 to 0.92), significantly greater than the AUC of the individual markers. A more recent study identified optimal thresholds for PCT, sCD14, galectin-3, and soluble suppression of tumorigenicity 2 (sST2) using ROC analysis and combined markers to predict 30-day mortality with an AUC score of 0.769 (95% CI 0.695–0.833) [98]. Taneja *et al.* deployed machine learning algorithms combining multiple novel biomarkers (IL-6, nCD64, IL-1ra, PCT, MCP1, and G-CSF) with routine electronic health data, identifying sepsis patients in early to peak phase of sepsis (classified on clinical judgement) using a support vector machine and reported an AUC score of 0.81 [99]. The same authors demonstrated in 2021 that a random forest model combining PCT, CRP, and IL-6 with routine electronic health data could identify patients with sepsis (according to the sepsis-3 criteria) with an AUC score of 0.83 [100]. Several studies in recent years have tried to capitalise on the power of statistical machine learning models to identify combinations of biomarkers that could predict sepsis or outcomes from sepsis. Due to the quantity of existing data, most studies in this domain focus on electronic health records, with the hope that combinations of informative biomarkers are present in existing data. Although some successful reports exist, with AUC scores far exceeding what is observed in traditional biomarker studies, results must be interpreted with caution due to low comparability, reproducibility, and a lack of conformity in sepsis definitions used [102, 101]. For those less familiar with machine learning, a comprehensive introduction is provided later in this chapter (see 1.3).

1.2 Cytometry bioinformatics

In the previous sections, I discussed how the pathophysiology of sepsis had been dissected over the decades, giving rise to the study of numerous biomarkers to try and help identify sepsis, determine its cause, and predict outcomes. A key player in this progress has been the application of cytometry, a technology widely used for the numeration and characterisation of biological material. The main application is the study of the immune system. When investigating immune cells by cytometry, whether with traditional flow cytometry or the recently introduced advanced techniques of mass cytometry and spectral flow cytometry, large single-cell data are generated with hundreds of thousands to millions of observations per sample. Historically, such data have been analysed using manual gating strategies in software such as FlowJo [103], and FCS Express [104]. The manual gating process involves observing data in two-dimensional plots and hand-drawing polygons (referred to as 'gates') around data regions to separate out data of interest. Gates can be generated in complex sequences (termed 'gating strategies') by alternating the variables of two-dimensional plots and adding additional gates to selected data regions. Whilst gating has been highly effective in the past, the knowledge of the immune system has grown substantially, and cytometry technology expanded to allow more parameters to be measured in a single experiment. Consequently, more extensive gating strategies formed and the process of manual gating has grown to a state of exceedingly high labour per experiment, introducing risks of subjective bias and poor reproducibility [105]. As a result, the field of cytometry bioinformatics has arisen, where modern computational algorithms and statistical machine learning methods have been applied to circumvent the need for manual gating, address quality control concerns, and improve our ability to make sense of high dimensional cytometry data.

1.2.1 The start of a golden age for cytometry bioinformatics

The rise of cytometry bioinformatics is a culmination of simultaneous advances in open-source software development, data science, advances in machine learning, and cross-discipline collaboration. Although attempts to use computer algorithms as a replacement for manual gates date back to the early 1990s [106], it is only since 2007 that progress has been made in delivering practical solutions.

Figure 1.3 provides a summary of open-source solutions published over the past decade. The font size given to each tool/algorithm reflects the impact measured by uptake and citations. The areas of development in cytometry bioinformatics overlap significantly with other single-cell technologies. They can be broadly categorised into infrastructure and frameworks, automated gating, supervised classification, clustering, and dimension reduction for data visualisation.

It can be argued that Raphael Gottardo and Ryan Brinkman started a new age in cytometry bioinformatics with the introduction of flowCore [108], one of the first comprehensive R programming libraries to offer data structures specific for cytometry data analysis. Early success in automated analysis followed with algorithms such as flowClust, utilising mixtures of t-distributions following a box-cox transformation for automatic selection of subpopulations [109]. FlowMeans followed this, an adaption of the traditional K-means clustering algorithm, modified to handle concave cell population and implemented a change point detection algorithm to detect the optimal number of subpopulations [110]. Around this time, SPADE [111] was developed, which deployed agglomerative clustering to density-dependent downsampled data to ensure that underrepresented cell populations are not merged into larger populations. SPADE also offered a minimum-spanning tree of identified clusters to help visualisation and exploratory analysis. Other notable examples were FLOCK, which used a grid-based density clustering algorithm [112] and samSpectral, a spectral clustering algorithm designed to handle large cytometry data [113].

In what appears as a compounding effect, additional algorithms were developed, often improving on previous attempts that utilised mixture models [114], or K-means clustering [115]. The developments in the field culminated in the first-ever critical assessment of automated cytometry analysis titled “The Flow Cytometry: Critical Assessment of Population Identification Methods (FlowCAP)" [116], extending two challenges to the field: firstly, present algorithms that could accurately replicate the labels generated by expert manual gating, and secondly, demonstrate analysis pipelines that could accurately predict external variables such as clinical outcomes using only cytometry data as input. Multiple algorithms reported F1 scores of about 0.85 and on some challenges as high as 0.98, with the authors concluding that automated methods had reached a significant level of maturity and accuracy for more widespread use. Despite the optimism, the data used in FlowCAP were not representative of the extensive high-dimensional data that modern instruments obtain. The largest

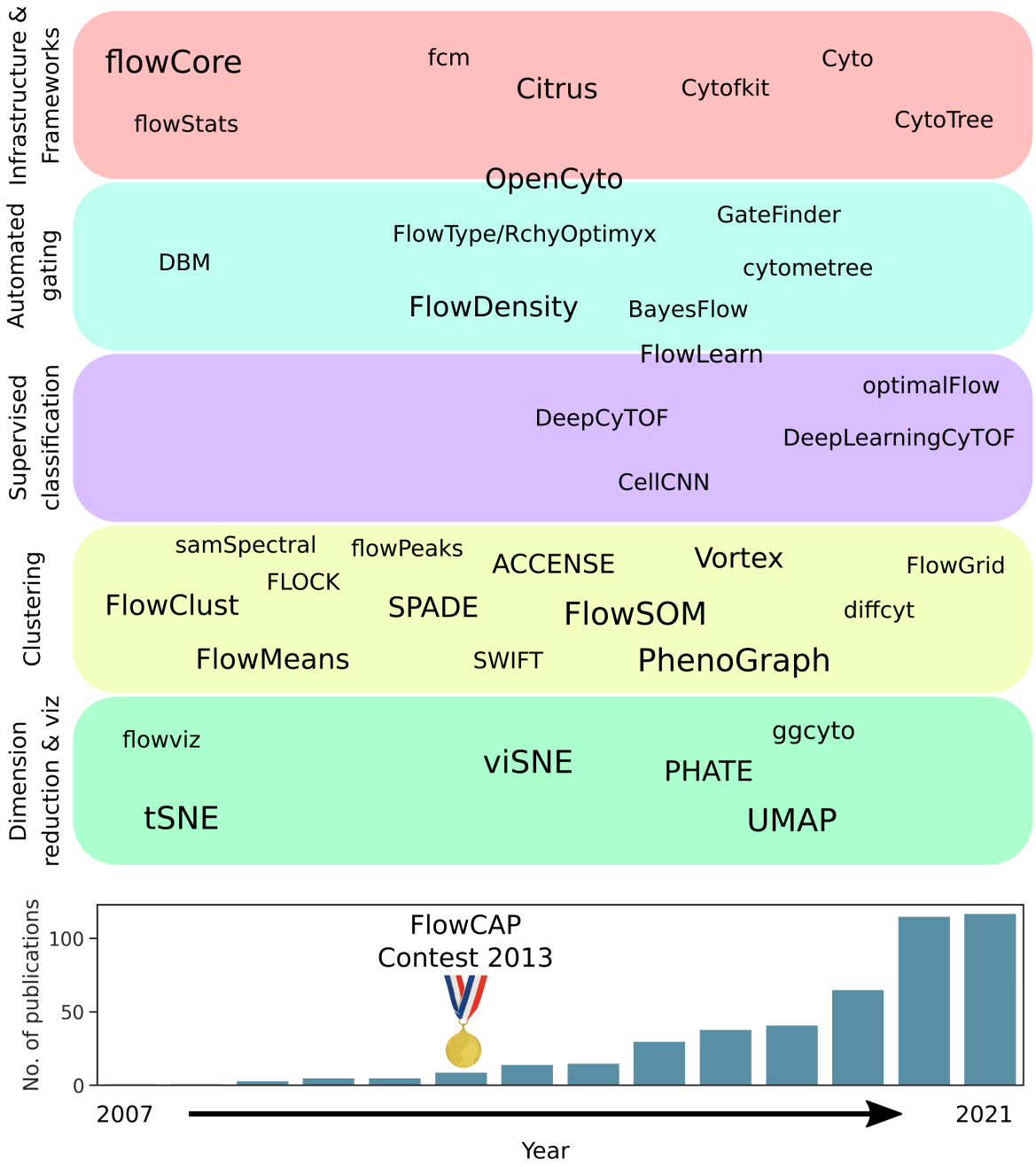

Figure 1.3: A timeline of developments in the field of cytometry bioinformatics between 2007 and 2022. Tools and algorithms are categoried into: infrastructure and frameworks, automated gating, supervised classification, clustering, and dimension reduction & visualisation. The font size given to each tool or algorithm reflects the impact measured by uptake and citations. The bar plot at the bottom of the figure shows the number of publications per-year using the search term “Machine learning and cytometry". Data were obtained from the PubMed search engine [107].

FlowCAP dataset consisted of only 100,000 events, and no more than ten parameters were captured in any single FlowCAP dataset.

1.2.2 The promise of automated gating

The field continued to develop, with significant progress around 2014-2015 in developing more approachable technologies that aimed to directly replicate manual gating in a way that could be recognised by traditional analysts. The flowDensity package [117] was a similar approach to the earlier Density-Based merging (DBM) [118] and generated gates in one- or two-dimensional space by first estimating the probability density function, then applying a peak finding algorithm, before applying the bounds of a gate based on regions of high and low density within the data. Around the same period, a similar approach was introduced by flowType used in combination with the RchyOptimyx algorithm [119]. The approach used flowMeans clustering to search for positive and negative subsets in two-dimensional space with an exhaustive search across all possible combinations of markers. RchyOptimyx measured the importance of these cell types by the correlation of cell population abundance with external outcomes such as disease state.

Significant impact was made with the first substantial autonomous gating framework, OpenCyto [120], built upon the flowCore architecture. OpenCyto presented the first end-to-end analysis framework emulating common gates found in manual analysis, explicitly designed for autonomous gating frameworks in one- or two-dimensional space. In the years that followed, the field continued to develop, with tools such as BayesFlow [121] and flowLearn [122], addressing issues of global variance, GateFinder [123], a back-gating algorithm that presents the optimal gating strategy for identifying a population found in N-dimensional space, and cytometree [124], an exhaustive search strategy that utilises Gaussian mixture models.

1.2.3 A supervised approach to classifying cytometry data

Autonomous gating might be easy to comprehend and validate but is still quite labour-intensive. It often requires optimising for each dataset due to its sensitivity to biological variation and technical noise. Cytometry data lend themselves well to another strategy: supervised machine learning. In brief, supervised machine learning algorithms are trained on

labelled data to predict the labels of unlabelled data. The 'training' part of this process often involves optimising a function (more specifically, the parameters of that function) that differentiates the labels. In the context of cytometry data, example data can be labelled with manual or automated gating and presented to a supervised machine learning algorithm to learn the association between the data and the labels. As a result, a method is generated whereby populations can be easily identified in new data.

Success in this area has mostly been found by applying multi-layer neural networks. DeepCyTOF [125] introduced a solution that first accounted for inter-sample variation with a residual neural network that tried to minimise the maximum mean discrepancy (MMD; a measure of similarity between two distributions) between a reference sample and subsequent batches. The reference sample was assumed to be manually gated and was also used to train a deep neural network in a supervised manner to classify cell populations. New data were aligned to the reference sample and labelled using the trained neural network. The SAUCIE algorithm [126] built on the concept of batch correction by minimising MMD with an additional autoencoder reconstruction penalty, forcing the preservation of the original structure in each sample.

An alternative approach was presented with the CellCNN algorithm [127], negating the need for manually gated reference data. CellCNN combined multiple instance learning and convolution neural networks (CNN) in a supervised representation learning approach. Data from all samples were pooled, and each event in the cytometry data was labelled by some external label of interest, such as disease state or patient outcome. A CNN (a particular type of neural network initially developed for analysing image data) was trained to predict the external label. The learnt filter weights corresponded to the molecular profiles of relevant cell subsets that are important for predicting the external label. CellCNN identified populations of cells of importance to broader questions rather than characterising all cell populations in a more traditional sense. The approach presented by CellCNN was extended further to demonstrate robust classification in the face of high variability between data obtained from different studies [128].

1.2.4 Clustering continues to improve

Clustering analysis has arisen as the most popular automated approach to cytometry analysis, primarily because it offers an exploratory aspect to analysis, allowing the investigator to categorise events by their phenotypic similarity and then perform hypothesis testing on the acquired groups. The approach is similar to traditional manual gating but is not biased by prior assumptions about the expected cell populations and is less labour-intensive.

After the FlowCAP competition, innovations continued, starting with ACCENSE [129], a peak detection algorithm applied to a kernel density estimate of a reduced feature space using t-distributed stochastic neighbour embedding (t-SNE). The development of dimension reduction technologies was a significant contributor to the field of cytometry bioinformatics and is discussed in length in the next section (see 1.2.5).

Arguably, the most significant leap forward in the development of cytometry clustering algorithms came in 2015 with the publication of FlowSOM [130] and PhenoGraph [131]. FlowSOM clustered data using self-organising maps (SOM), a specific type of unsupervised neural network. A SOM distributes a grid of nodes in N-dimensional space (often randomly initiated), where each node represents a point in the given feature space. During clustering, a data point is classified with the node that is its nearest neighbour. The grid is trained so that the nodes closely connected via the observed data resemble each other more than nodes connected via a longer path, capturing topological information about the data. PhenoGraph instead utilised graph theory to describe populations in single-cell data. First, a nearest neighbours graph was constructed, and then the problem of density detection was addressed by identifying communities of similar cells as highly connected regions within the graph. The idea was borrowed from the field of social network research and utilised the Louvain community detection algorithm to partition the graph in a way that maximises modularity.

FlowSOM and PhenoGraph offered exceptional computational efficiency compared to their earlier counterparts and performed well on single-cell data. The impact of these methods is reflected in the thousands of citations accumulated by both and the adoption of these techniques into traditional software as plugins. The computational performance of the FlowSOM algorithm stands out with its ability to cluster millions of data points in a few minutes but requires the user to define the expected number of clusters and is, therefore, best suited to tasks where the investigator has some prior knowledge of the expected populations. Phenograph,

on the other hand, does not require the desired number of clusters to be known and does an excellent job of describing small populations, but is computationally expensive and can result in fragmented clusters [132, 133]. Despite the success of FlowSOM and PhenoGraph in solving many of the challenges in clustering cytometry data, development did not cease, and new tools that optimised nearest neighbour density estimation and selection of an optimal number of clusters followed [134], as did novel ways of framing the analysis of clustering results, such as differential abundance and differential state analysis, as seen in the diffcyt package [135].

1.2.5 Seeing is believing: how dimension reduction changed the game

No single technology can be credited as solely responsible for the maturation of cytometry bioinformatics, but the development of dimension reduction technologies certainly comes close. The ability to visualise multi-dimensional cytometry data has offered a tool of communication that has helped bridge the gap between bioinformatics and other disciplines. The story starts with t-SNE [136], a manifold learning technique that modelled the similarity between pairs of data points as joint probabilities. A similar probability distribution was then constructed in low-dimensional space, and the t-SNE algorithm tried to minimise the Kullback-Leibler divergence between the two probability distributions with respect to the location of the original data points.

The original t-SNE algorithm was computationally expensive, and in 2013 viSNE was published, a fast distributed implementation of the t-SNE algorithm improved for single cell analysis. A notable divergence from the original t-SNE algorithm was the absence of principal component analysis (PCA) in the pre-processing steps. However, the fundamental approach to t-SNE was criticised for its computational complexity, stochastic behaviour, and loss of global structure in favour of conserving local structure. The Uniform Manifold Approximation and Projection (UMAP) algorithm [137] addressed these issues using Riemannian geometry, resulting in significantly faster runtimes and improved conservation of global structure. At the same time, the Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE) algorithm was developed [138]. Unlike UMAP, which was developed for general purposes, PHATE was specifically designed to analyse biological data. PHATE acknowledged that biological data tends to contain progressive branching structures, which are often non-linear and reflect underlying biological processes. PHATE preserved the pro-

gressive structure in data by modelling the problem with heat-diffusion processes to compute cell-cell affinities. A diffusion-potential geometry captured the high-dimensional trajectory structures within the data. The resulting embeddings accurately depicted complex trajectories and data distances.

1.2.6 Moving forward

Cytometry bioinformatics has come a long way in the past 15 years, bringing not just autonomous means by which to identify cell populations but also dedicated programming frameworks for specific applications [123, 127, 139, 140, 141] and tools to improve the quality of our raw data prior to analysis [142, 143, 144]. However, widespread adoption of the more advanced techniques in cytometry bioinformatics is still lacking. Broad open-source programming frameworks that offer structure and guidance to analysis help newcomers to the field and reduce the barrier of entry for those lacking experience with programming languages [108, 120].

Despite the ongoing efforts, much work must be done to increase the accessibility of tools further. Most cytometry bioinformatics solutions are implemented in the R programming language. However, in recent years, there has been a spike in the popularity of other programming languages, notably Python. Python is ranked amongst the most desired programming languages in recent surveys of developers [145, 146, 147], is popular across domains (meaning guidance and resources are abundant) and offers improved debugging and a simpler syntax to R. It is no surprise that Python has rapidly been adopted in other areas such as genomics [148] and single-cell RNA sequencing analysis [149]. In recent years more cytometry-specific tools have arisen that were developed in Python [125, 127, 128, 131, 138, 150, 151] and this reflects the adoption of deep learning neural networks, where Python is the preferred programming language due to the availability of open-source tools such as TensorFlow [152], and PyTorch [153].

More work is required to deliver low-code interfaces for the tools developed in the past decade. Additionally, the overwhelming choice of technologies will require a pathway to quick validation and consolidation of results that benefit from the advantages of each tool without overlooking their limitations. In improving access to these tools, cytometry bioinformatics will also substantially impact biomarker work. The exploratory analysis that clus-

tering offers opens the potential for identifying novel biomarkers that have been overlooked in the past. In combination with statistical machine learning technology, combinations of biomarkers could potentially be found to overcome the limitations that individual biomarkers of sepsis have presented.

1.3 Pattern recognition for biomarker discovery

Traditionally, biomarker discovery in sepsis has followed deductive reasoning driven by the observations of the pathophysiology of sepsis derived from experimental models and clinical trials. The failure to identify individual biomarkers in sepsis [155, 154, 96], combined with the overwhelming amount of data that can be obtained from the combination of electronic health records (EHR) and multi-omic platforms, questions whether biomarker discovery is better placed in an inductive framework of pattern recognition. The inductive pattern recognition approach does not assume that a single biomarker is capable of prediction. Instead, a combination of biomarkers and their interaction with one another is required for suitable accuracy. A limitation of an inductive approach is that the logic depends on the quality of the observations and how well they generalise to a broader population. Therefore, the objective should be to generate a hypothesis such as "does the identified combination of N biomarkers accurately predict X" from the data observed. The resulting hypothesis should be tested on larger populations to see if the chosen biomarker pattern generalises.

The next question is, how do we identify such a pattern? Pattern recognition or 'classification' is fundamental to supervised machine learning. As mentioned in the previous section, supervised machine learning allows us to 'train' a function using some labelled data and then apply the optimised function to new data to classify unknown events (Figure 1.4). The training step involves an optimisation algorithm that searches for the parameters of a function that obtain the best predictions in the training data. Many types of decision functions and algorithms can be used, ranging from simple linear models such as ordinary least squares logistic regression to more complex models such as support vector machines, tree-based classifiers, and deep neural networks.

1.3.1 Feature selection for pattern recognition

In the field of biomarker study, the objective is to ascertain the most relevant variables from N input variables. Such a task is synonymous with a process known as feature selection in the statistical literature. Feature selection is necessary for several reasons. Firstly, the practical application of biomarkers in the clinic requires the minimal possible combination for classification accuracy to reduce cost and simplify the interpretation of results. Secondly, it

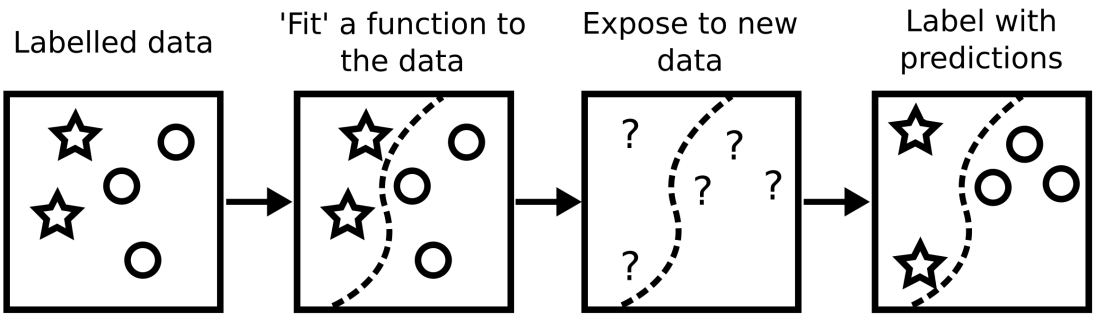

Figure 1.4: Schematic of the general concept behind supervised machine learning.

is well known that reducing the number of variables can help improve the performance of classifiers [156, 157] and reduce the risk of overfitting. Overfitting occurs when a model performs well on training data but does not generalise well when exposed to new data, and therefore performance will be poor. Overfitting is especially important for biomarker studies since clinical research tends to reduce the size of available training data. Reducing the variable space helps reduce the complexity of models and lowers the risk of overfitting. Finally, a reduced number of variables can lower the likelihood that variables are highly correlated. Multi-colinearity amongst the input variables can break the assumption of independence required for some models, negatively impact classifier performance, and create issues with the interpretation of variable importance in model decisions [158].

Feature selection can be broadly categorised into filter, wrapper, and embedded methods [159]. Filter methods include t-tests, information gain, correlation testing, and more advanced methods like Relief algorithms. Filter methods are model-independent and tend to be computationally simple, offering fast performance. Most filter methods have the disadvantage of ignoring feature dependencies but are independent of the downstream model and can therefore be seen as unbiased [159, 161, 160].

Wrapper and embedded methods are classifier-dependent, and the retention or elimination of a feature is driven by the impact that the feature has on the classifier's performance. Wrapper methods perform some form of iteration around a classifier, modifying the input variables according to changes in model performance. Examples include backward and forward feature elimination, genetic algorithms, and randomised hill climbing. Wrapper methods take into account feature interactions but are more prone to overfitting and are computationally intensive. Embedded methods offer better runtimes than wrapper methods. In embedded

methods, the characteristics of a classifier are employed for the task of feature selection. Examples include feature importance in tree-based learning algorithms, the weight vector of support vector machines and regularisation methods such as Least Absolute Shrinkage and Selection Operator (LASSO) [159, 161, 160].

The use of feature selection is quickly becoming the norm amongst biomarker studies exploring high-dimensional data, and this trend will likely continue as more extensive proteomic, genomic, and transcriptomic data are acquired [162, 163]. Examples already exist for the successful application of feature selection for identifying biomarkers in sepsis. A whole-genome transcriptomic analysis of messenger RNA isolated from urine employed four independent feature selection techniques to identify an optimal subset of probes for differentiating sepsis from non-infected controls, with an AUC score of 0.86 (0.77–0.93) [164]. Parthasarathy *et al.* identified novel deferentially expressed immature neutrophil subsets in sepsis patients, using LASSO for feature selection prior to generating a Random Forest model to differentiate sepsis patients and healthy controls [165]. Lukaszewski *et al.* applied the Boruta algorithm (a wrapper method for Random Forest) to select a gene expression signature that could differentiate sepsis from non-infected controls with an AUC of 0.897.

1.3.2 Moving beyond the 'black box': the promise of interpretable machine learning

As reliance on more complex statistical models becomes necessary to make sense of the high-dimensional data explored for biomarker signatures, the risk of creating 'black box' models arises. When a model's decisions cannot be interpreted, the risk of unchecked bias in the training data occurs, and the ability to explain model outputs is completely diminished. Therefore, it is of the utmost importance that steps are taken to generate transparent and well-understood models.

Interpretable machine learning can be achieved through several techniques. The most straightforward is to ensure that the models are naturally interpretable, an example being linear models. In a linear model, the prediction is a weighted sum of the variable inputs, and the optimised weights can be interpreted as the contribution a variable makes to the prediction. Simplistic models tend to underfit when there are complex non-linear relationships between variables and the target class, encouraging the application of more complex models that

are harder to describe. Some complex models still offer insight into their decision-making; examples include ensembles of tree-based learners, like Random Forest. Since decision trees split data at nodes using a chosen variable, the impurity measure at each node can be weighted by the probability of reaching that node, giving a value for the importance of a variable; this is commonly referred to as feature importance. For other complex models, model-agnostic methods can be used, such as partial dependence plots, permutation testing, and local surrogate models [166].

A recent advance in the development of interpretable machine learning technologies has been the introduction of SHapley Additive exPlanations (SHAP) [167]. The SHAP framework uses Shapely values to estimate the contribution of input variables to a model. The Nobel prize-winning economist Lloyd Shapely first formulated Shapely values to answer the question "given a coalition of actors that generate some output, what is the contribution of each actor?". The problem is complicated by interactions between individual members of the coalition. Shapely values are computed for each coalition member to find an acceptable answer to this question whilst considering the interaction terms. A Shapely value is calculated by first taking a sample of the coalition that contains the member of interest and then compare to the same permutation with that member removed. The outcome value is calculated for both permutations, and the difference between these values represents the marginal contribution of the missing member to the coalition where this member is absent. The procedure is repeated across all possible coalition permutations, and each permutation's marginal contributions are calculated. The mean marginal contribution is the Shapely value for that member. The SHAP authors took this concept and applied it to the context of machine learning, treating the input variables of a function as members of the coalition. To account for the computational intensity of computing Shapely values, they introduced the Shapely kernel to approximate Shapely values through much fewer permutations [166, 167]. SHAP values were used recently to explain the outputs of an in-hospital mortality prediction algorithm for critically ill patients with sepsis, identifying the importance of Glasgow Coma Score, blood urea nitrogen, respiratory rate, urine output, and age to an XGBoost model [168]. SHAP values are computationally inexpensive, offer transparent explanations relative to the average prediction, and can be extended to global model interpretations. SHAP values will be explored further in Chapter 6 to explain the choice of biomarker signatures in sepsis.

1.4 The scope of this thesis

This thesis demonstrates comprehensive immunophenotyping followed by the application of supervised machine learning models for identifying prognostic and diagnostic patterns in sepsis. Novel contributions to cytometry bioinformatics are demonstrated and then applied to a small yet complex dataset of patients sampled within 36 hours of their sepsis diagnosis. The software developed as part of this work and the findings that I present have implications beyond the study of sepsis and the identification of multi-biomarker panels but are broadly applicable to all researchers performing cytometry data analysis. The work demonstrates how to characterise the immune response with cytometry bioinformatics and use those insights as input for a supervised machine learning framework. The final results chapter provides an end-to-end solution for utilising supervised machine learning algorithms to identify combinations of informative biomarkers when faced with challenges such as class imbalance, missing data, and multicollinearity. Descriptive analysis and statistical modelling focused on identifying biomarkers that correlate with patient mortality and the underlying cause of infection. The outputs presented here suggested new concepts for stratifying patients, directing care, and delivering improved prognoses. Additionally, recognising biomarker combinations that predict the causative pathogen within the first 36 hours of sepsis diagnosis has implications for delivering personalised care. The thesis delivered these elements in the following sections:

Chapter 2 provided an overview of the methods used throughout this thesis. The Innate-like T cells in sepsis (ILTIS) study was described. ILTIS was a comprehensive observational study of sepsis patients identified according to the Sepsis-3 criteria and was the main subject of this thesis. Additional data utilised for the validation of new methodologies was also described.

Chapter 3 introduced CytoPy, a novel cytometry data analysis framework developed in the Python programming language. CytoPy formed a foundation for reliable and practical analysis of cytometry data in Python, creating data structures specifically designed for autonomous analysis of cytometry data. Features included automated gating, batch correction, and clustering analysis. CytoPy was validated on data from dialysis patients diagnosed with acute peritonitis.

Chapter 4 builds on the previous chapter by introducing a novel ensemble clustering algorithm for cytometry data analysis. There are currently an overwhelming number of clustering algorithms for cytometry analysis, each with its benefits and disadvantages. These algorithms also include complex hyperparameters, and the outputs can differ significantly between models. Inspired by the 'wisdom of crowds' approach, this chapter introduced the geometric median clustering with weighted voting (GeoWaVe) algorithm, a novel ensemble clustering algorithm to combine multiple clustering results.

Chapter 5 described the early immunological changes in severe sepsis patients and their electronic health record data captured within 48 hours of their enrolment in the ILTIS study. CytoPy and GeoWaVe were applied in an exploratory analysis that describes the phenotypes of neutrophils, monocytes, conventional CD4 and CD8 T cells, and the unconventional T cell populations of MAITs and V$\delta 2^{+}$ $\gamma\delta$ T cells. The descriptive and univariate statistical analysis explored the relationship between immune system variables and patient outcomes, as well as the underlying cause of infection.

Chapter 6 detailed the creation of advanced statistical machine learning models that combine all data from the ILTIS study to predict mortality and the underlying cause of infection. A comprehensive machine learning pipeline was generated, including multiple imputation of missing values, feature selection, model selection and evaluation, and interpretation of model predictions with SHAP values. The work here demonstrated how biomarker signatures can be identified from small yet complex data of severe sepsis patients and help generate new hypotheses regarding potential combinations of biomarkers.

2 | Materials and Methods

2.1 Innate-like T cells in sepsis (ILTIS) study

The "Innate-Like T cells In Sepsis" (ILTIS) study was the primary focus of this work and the subject of Chapter 5 and Chapter 6. Data generated from this study also appear in Chapter 4 for bench-marking ensemble clustering methods. The principal investigator was Professor Matthias Eberl, and patients were recruited by the clinical lead Dr Matthew Morgan and the research team at the critical care directorate within the Cardiff and Vale University Health Board. Sample processing was performed by Dr Löic Raffray, Ms Sarah Baker, and myself. Data acquisition in the laboratory, electronic data collection from the clinic, and data analysis were performed by myself.

2.1.1 Ethics and consent

Recruitment of sepsis patients was approved by the Health and Care Research Wales Research Ethics Committee under reference 17/WA/0253, protocol number SPON1609-17 and IRAS project ID 231993, and conducted according to the principles expressed in the Declaration of Helsinki. All participants provided written informed consent for the collection of samples and their subsequent analysis. A waiver of consent system was used when patients were unable to provide prospective informed consent due to the nature of their critical illness or therapeutic sedation at the time of recruitment. In all cases, retrospective informed consent was sought as soon as the patient recovered and regained capacity. In cases where a patient passed away before regaining capacity, the initial consultee's approval would stand. Recruitment of healthy adult volunteers was approved by Cardiff University's School of Medicine Research Ethics Committee under reference 18/04.

Inclusion criteria	Exclusion criteria
Acute severe sepsis patients	
Age > 18 years.	Pregnant, breastfeeding or females of childbearing age in whom a pregnancy test has not been performed.
Diagnosis of sepsis according to the 'Sepsis-3' criteria.	Severe immune deficiency, for example: a diagnosis of AIDS, anti-rejection transplant drugs, long-term high dose corticosteroid treatment (> 10mg prednisolone/-day or equivalent).
Cared for in the intensive care unit.	Severe liver failure (Childs-Pugh III or worse).
Within 96 hours of presumed onset of infection.	Patient judged by admitting clinician unlikely to survive for 3 days regardless of treatment.
Patient already has or will require arterial cannulation as part of standard treatment.	Patients admitted post-cardiac arrest.
Healthy volunteers	
Age > 18 years.	Pregnant, breastfeeding or females of childbearing age in whom a pregnancy test has not been performed.
	Any long term chronic disease or medication use.
	Currently suffering from an acute illness however minor.

Table 2.1: The inclusion and exclusion criteria for recruitment into the ILTIS study.

2.1.2 Patient recruitment

Sepsis patients over the age of 18 years old with a diagnosis of sepsis, according to the Third International Consensus Definitions for Sepsis and Septic Shock ('Sepsis-3'), were cared for in the intensive care unit at the University Hospital of Wales in Cardiff and were recruited within 36 hours of the presumed onset of infection when they already had or would require arterial cannulation as part of standard treatment. Healthy controls were recruited through institutional advertisement, with all donors signing a consent form and being presented with a participant information leaflet. The inclusion and exclusion criteria for acute severe sepsis patients and healthy donors are detailed in Table 2.1.

2.1.3 Sample and data collection

A day 1 sample of 30 ml of peripheral blood was taken from an arterial line within the first 36 hours of sepsis diagnosis with suspected infection. Samples were taken in an EDTA vacutainer collection tube and transported to the laboratory on ice.

Patient demographics, body mass index, and the unit outcome was obtained from the Cardiff and the Vale University Health board critical care Ward Watcher software. All clinical data were collected retrospectively from the Cardiff and the Vale Health Board Clinical Portal, including patient mortality, captured as either death within 30 days or 90 days after enrolment.

Haemotology, biochemistry, blood gas analyser, and other point-of-care testing data were extracted from the web interface as HTML files and processed following the Data Protection Act (DPA) 2018, General Data Protection Regulation (GDPR), and Cardiff and the Vale Health Board data protection policy. Consent to access electronic medical records was obtained during patient recruitment, and only data relevant to this study were accessed and retained for analysis. Data were extracted in a secure environment within the hospital computer network and anonymised prior to analysis. Python version 3.8 [169] and the Beautiful Soup package [170] was used to process electronic health records and generate an anonymised tabular database. Electronic health records from seven days before and after enrolment were included in this database. A patient could have multiple physiological and biochemical measurements for the same variable within the 14-day window. Therefore, to avoid biasing our observations by including events outside the episode of sepsis, data were summarised as follows:

- The median value within 48-hours prior to enrolment and 8-hours after enrolment.
- The value obtained closest to the enrolment date and time.

Microbiology data were captured within the hospital laboratory information management system (LIMS) and accessed through the patient's electronic medical record. The primary causative pathogen was obtained from a positive microbiological culture of pure growth or positive virology in any sample from 72 hours preceding recruitment to 72 hours following recruitment in the critical care unit. The microbiological techniques used in the study hospital for diagnostics included standard microscopy and culture, Matrix Assisted Laser

Reagent	Constituents
FACS buffer	2% *v/v* foetal calf serum (Invitrogen) 0.02% Sodium azide (Fisher Scientific) PBS (1x)
Blocking buffer	1% Human Purified IgG (Kiovig; Baxter) FACS buffer (1x)

Table 2.2: All solution based reagents used including their constituent parts.

Desorption/Ionization Time-of-flight Mass Spectrometry (MALDI-TOF) for bacterial identification, viral PCR studies, and urine Legionella antigen testing. The causative pathogen was checked against clinical notes and discharge summaries for clarification and confirmed by a critical care consultant. Patients whose infectious source and causative pathogen could not be obtained were labelled culture negative. Infections were grouped into Gram-positive and Gram-negative groups.

2.1.4 Reagents

The solution-based reagents used are described in Table 2.2. All reagents were kept under the conditions specified by the manufacturer and added constituents filtered using a 0.22 μm pore size hydrophilic polyethersulfone membrane filter. Stock tests were performed on new batches before use in experiments.

2.1.5 Isolation of leukocytes, peripheral blood mononuclear cells, and cell-free plasma from whole blood

The procedure for isolating leukocytes, peripheral blood mononuclear cells (PBMC), and cell-free plasma is detailed in Figure 2.1. For the analysis of monocytes and neutrophils, 3 ml of peripheral blood was subjected to red blood cell lysis: whole blood was exposed to 1:10 RBC lysis buffer (eBiosciences - Thermofischer, ref 00-4300-54), gently mixed, incubated for 12 minutes at room temperature, and centrifuged at 400 $\times$ g at room temperature. The supernatant was discarded, and the leukocyte fraction was retained and washed with PBS.

The remaining sample of whole blood was centrifuged at 300 × g at room temperature, and the supernatant was removed, centrifuged again, and then stored at -70°C. This cell-free plasma was analysed at a later time point for the quantification of cytokines and chemokines. PBMCs were isolated from the remaining content of the samples using density centrifugation: whole blood was layered onto 15 ml of Lymphoprep™ density gradient medium (Stem-Cell Technologies, ref 07801), centrifuged, and the mononuclear layer carefully harvested using a pipette and washed with PBS.

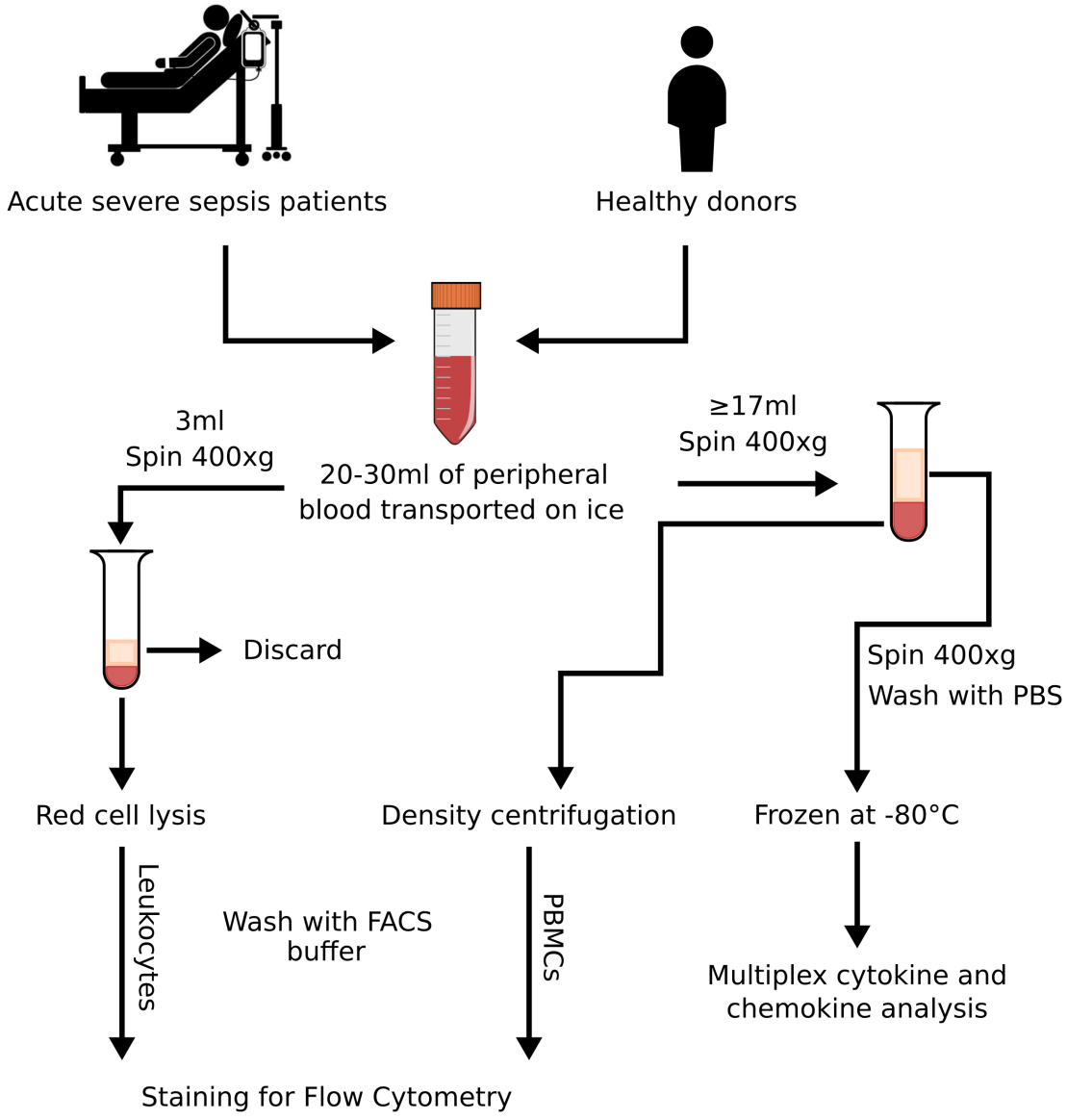

Figure 2.1: Schematic of sample processing: isolation of leukocytes, peripheral blood mononuclear cells (PBMCs), and cell-free plasma.

2.1.6 Flow cytometry

Three monoclonal antibody-fluorochrome staining panels were applied for cell surface staining. Two staining panels were applied to PBMCs for the identification of T lymphocyte subsets (Table 2.3): classical T cells and non-classical T cells (MAITs and V$\delta 2^+$ $\gamma\delta$ T cells). The first of these two staining panels (labelled 'T1') identified memory subsets, and the second (labelled 'T2') identified activated subsets. The third staining panel (Table 2.4) was applied to the lymphocyte fraction of whole blood after the removal of erythrocytes by RBC lysis and identified monocytes and neutrophils and activated subsets of each.

The T cell panels (Table 2.3) included CD3, CD4, CD8, Pan-$\gamma\delta$, and V$\delta 2$ for the identification of conventional CD4 and CD8 T cells, as well as V$\delta 2^+$ $\gamma\delta$ T cells. The surrogate markers CD161 and V$\alpha 7.2$ were included to identify MAIT cells. In the dump channel (V500 conjugate), CD14 and CD19 were included to eliminate monocytes and B cells. The markers CD45RA, CCR7, and CD27 enabled memory subsets of T cells to be differentiated, and CD57 was included as a marker of T cell senescence. The markers CD25, CD69, HLA-DR, and CXCR3 allowed the identification of activated cell states.

The panel for the identification of monocytes and neutrophils (Table 2.4) consisted of CD14 and CD15, which, when combined with forward and sideward scatter, allowed for the differentiation of monocytes and neutrophils. The costimulatory receptor CD86 combined with the MHC class II molecule HLA-DR served as activation and antigen-presenting capability markers. CD11b (also known as ITGAM) modulates cell adhesion, migration, and phagocytosis, and its expression is increased upon activation of monocytes and neutrophils. CD62L (also known as L-selectin) is a cell adhesion molecule important in leukocyte trafficking and is shed upon activation [17]. These markers, combined with CD40 (a member of the TNF receptor superfamily) and CD64 (an early activation marker), were used to characterise activated states of monocytes and neutrophils and evaluate their antigen-presenting capabilities.

For each staining panel, a pellet of 2×10^6 cells was stained with 3μl live/dead stain (fixable Aqua; Invitrogen) and incubated at room temperature for 15 minutes in the dark. Following incubation, cells were washed with PBS and re-suspended in FACS buffer. Cells were blocked for non-specific antigen binding using 1% human IgG (Kiovig; Baxter) diluted in FACS buffer and incubated for 15 minutes on ice in the dark. The cells were rewashed with FACS buffer before staining with the relevant monoclonal antibodies.

Antigen	Conjugate	Clone	Isotype	Manufacturer
CD3*	APC/FIRE	SK7	Mouse IgG1, κ	Biolegend
CD4*	PE-Cy5.5	S3.5	Mouse IgG2α, κ	Life Tech (Thermo Fisher)
CD8a*	BV711	RPA-T8	Mouse IgG1, κ	Biolegend
CD14*	V500	M5E2	Mouse IgG2α, κ	BD
CD19*	V500	HIB19	Mouse IgG1, κ	BD
CD25†	PE-Cy7	M-A251	Mouse IgG1, κ	BD
CD27^	PE-Cy7	M-T271	Mouse IgG1, κ	Biolegend
CD45RA^	PE Dazzle	HI100	Mouse IgG2b, κ	Biolegend
CD57^	FITC	NK-1	Mouse IgM, κ	BD
CD69†	PE-CF594	FN50	Mouse IgG1, κ	BD
CD161*	APC	191B8	Mouse IgG2α, κ	Miltenyi
CD197 (CCR7)^	BV421	G043H7	Mouse IgG2α, κ	Biolegend
CXCR3†	FITC	49801	Mouse IgG1, κ	R&D
HLA-DR†	BV421	G46-6	Mouse IgG2α, κ	BD
TCR-pan-$\gamma\delta$*	PE-Cy5	IMMU510	Mouse IgG1, κ	Beckman Coulter
Vα7.2*	BV605	3C10	Mouse IgG1, κ	Biolegend
Vδ2*	PE	B6	Mouse IgG1, κ	BD

Table 2.3: Antibody-fluorochrome cocktails applied to PBMCs for identifying subsets of T lymphocytes.
*Linage markers included in both staining panels (T1 & T2).
^ Memory and effector markers included in T1 staining panel.
†Activation markers included in T2 staining panel.

Antigen	Conjugate	Clone	Isotype	Manufacturer
CD11b	BV421	ICRF44	Mouse IgG1, κ	Biolegend
CD14	PE-Cy7	M5E2	Mouse IgG2α, κ	Biolegend
CD15	BV605	W6D3	Mouse IgG1, κ	Biolegend
CD19	V500	HIB19	Mouse IgG1, κ	BD
CD40	PE	MAB89	Mouse IgG1, κ	Beckman Coulter
CD62L	PE-Cy5	DREG-56	Mouse IgG1, κ	BD
CD64	APC-H7	10.1	Mouse IgG1, κ	BD
CD86	FITC	2331	Mouse IgG1, κ	Biolegend
HLA-DR	BV711	L243	Mouse IgG2α, κ	Biolegend

Table 2.4: Antibody-fluorochrome cocktail for cell-surface staining of Leukocytes, after red cell lysis, for identifying subsets of monocytes and neutrophils.

Cells were acquired using an 16-colour BD LSRFortessa™flow cytometer (BD Biosciences, Wokingham, UK) and the BD FACSDiva™software. The flow cytometer was calibrated using BD FACSDiva™CS&T research beads (BD Biosciences, ref 650622) prior to acquisition to verify optical path and stream flow. Compensation for spectral overlap was accounted for using BD CompBeads (BD Biosciences, positive control ref 51-90-9001229 and negative

control ref 51-90-9001291) and a spillover matrix generated using the FACSDiva™software. Compensation was checked for errors using the FlowJo software (TreeStar) prior to analysis.

The CytoPy software, described in full in Chapter 3, was used for all subsequent analyses of cytometry data. T cells, monocytes, and neutrophils were gated on their appearance in side and forward scatter area/height and exclusion of live/dead staining (fixable Aqua; Invitrogen). This pre-processing step was performed using a mixture of manual and autonomous gating. An example of the gating strategy for T cells is given in Figure 2.2, and for monocytes and neutrophils in Figure 2.3. The exclusion of monocytes and B cells from PBMCs in T cell staining was ensured by the inclusion of CD14 and CD19 in the live/dead staining channel. When identifying monocytes and neutrophils in the leukocyte fraction of whole blood, T cells were removed by their appearance in side and forward scatter area/height, and B cells were excluded by including CD19 in the live/dead staining channel.

Before analysis with the CytoPy software, quality control checks of the flow cytometry standard (FCS) files were performed using the FlowAI software [142] in the R programming language version 4.0. The FlowAI package removes unwanted events from cytometry data by detecting abrupt changes in flow rate, instability of signal acquisition, and outliers in the lower and upper limits of the dynamic range. The FlowAI software was run using the default parameters, and each file was checked manually using the R shiny interface.

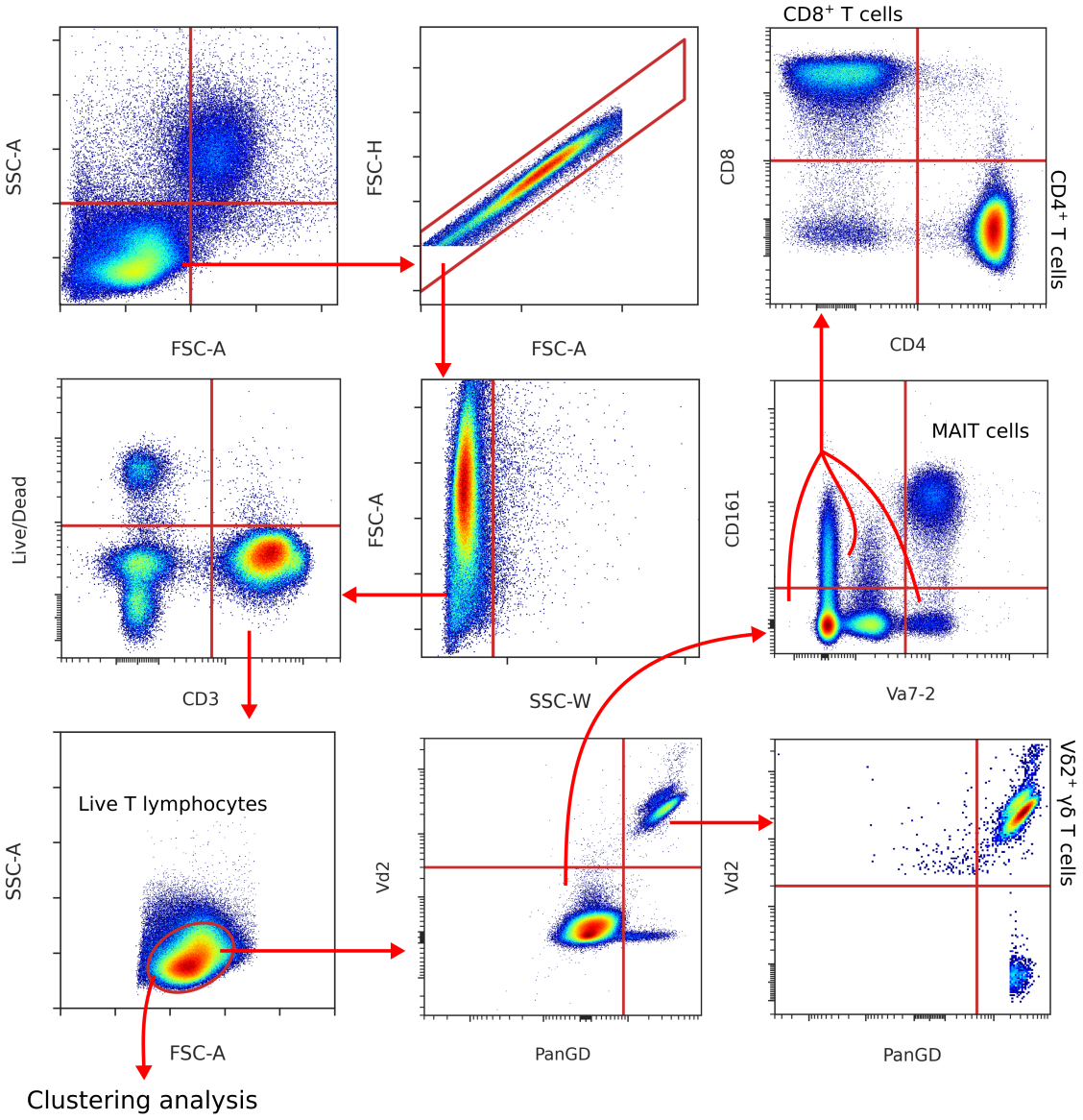

Figure 2.2: Gating strategy applied with the CytoPy software for the identification of single live T lymphocytes, conventional $CD4^+$ and $CD8^+$ subsets, and unconventional T cells. Gates were generated using autonomous gating as discussed in chapter 3.3.3. T lymphocytes are provided as input for downstream clustering analysis.

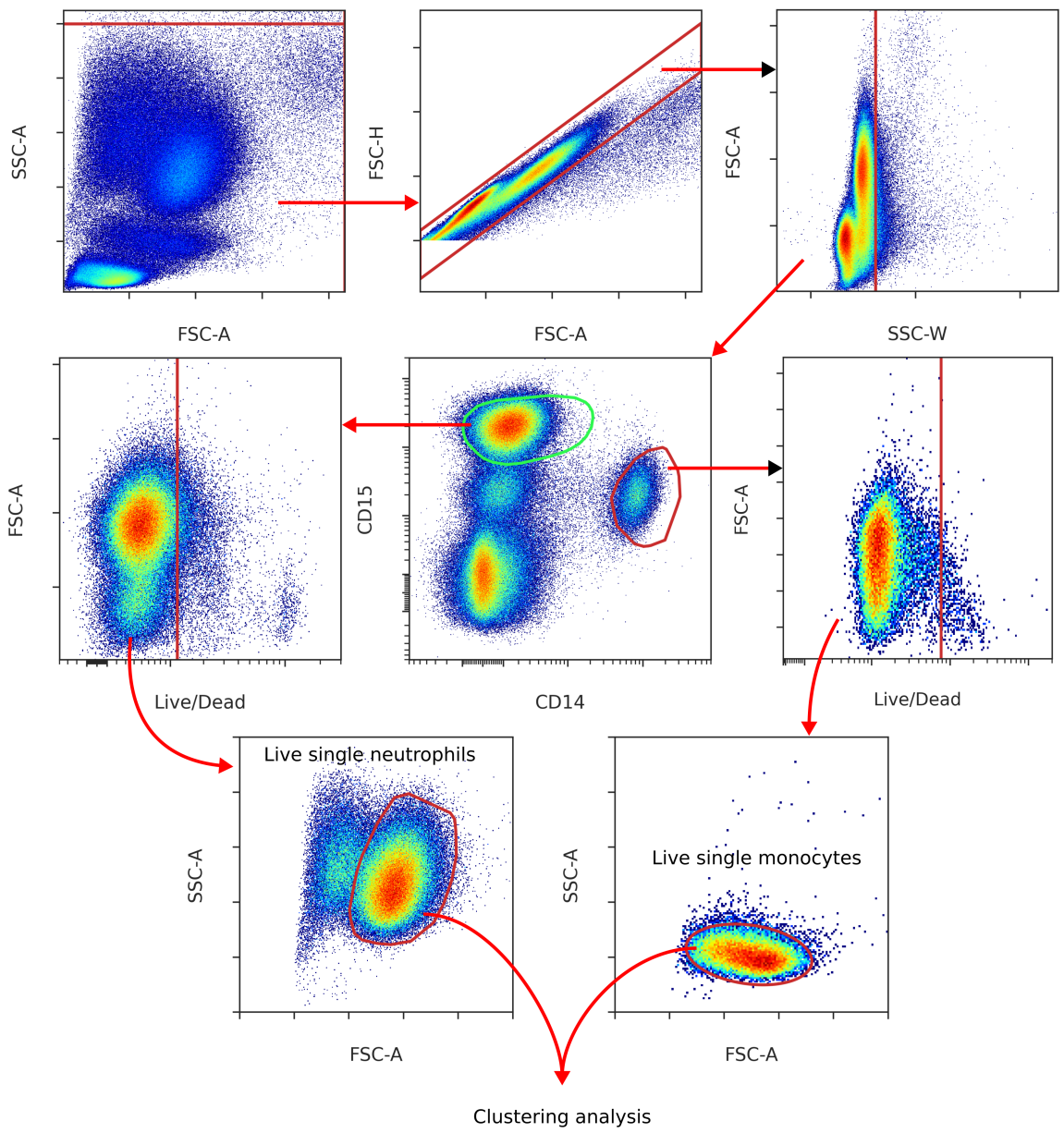

Figure 2.3: Gating strategy applied with the CytoPy software for the identification of single live monocytes and neutrophils. Gates were generated using autonomous gating as discussed in chapter 3.3.3. Monocyte and neutrophils populations were then combined and provided as input for downstream clustering analysis.

2.1.7 Luminex™ and ELISA

As described in 2.1.3, frozen cell-free plasma was obtained from whole blood and thawed in two batches. Cytokines and chemokines (Table 2.5) were quantified according to manufacturer guidelines using Luminex™standard sensitivity magplex assays based on the xMAP (multi-analyte profiling) technology. This technology uses micro-sphere beads labelled with monoclonal-fluorochrome conjugate antibodies to capture multiple analytes simultaneously. A compact flow cytometer is then used to quantify beads with bound analytes. Data were acquired on a Luminex 200™compact analyser. A standard panel of cytokines, chemokines, and acute-phase proteins was constructed as per the manufacturer's guidelines. The chosen analytes reflect anti-inflammatory and pro-inflammatory markers of interest and acute-phase proteins previously implicated in sepsis.

Concentrations were obtained by fitting a standard curve using the Python programming language [169], and SciPy version 1.7.1 [171]. A five-parameter logistic fit using the generalised hill equation for standard dose-response curves [172] was used per the manufacturer's guidelines. Data quality was assessed by observing the coefficient-of-variation and standard recovery. Where less than three observations fall within the standard range, the majority of observations (>50%) had a CV greater than 50%, or the standard recovery was outside a range of 75% to 125%, analytes were deemed of poor quality and excluded from subsequent analysis (reasons for exclusion are detailed in Table 2.5).

TNF-α (eBioscience; ref:88-7346), IFN-γ (eBioscience; ref:88-7316), and IL-6 (R&D Systems; ref: DY206) were measured using single ELISA, as per manufacturer guidelines. Unlike with the multi-plex Luminex™ assays, ELISAs were performed in a single batch.

Batch effect in Luminex™ multi-plex assay experiments was addressed with posthoc correction and data alignment. Using the Python programming language [169] data were log base 2 transformed and values replaced with a z-score, as previously described by Tomic *et al.* [173] and Whiting *et al.* [174].

Analyte	Method	Avg. standard recovery (%)	Reason for exclusion (if excluded)
CXCL8	Luminex™	101.2	
CXCL10	Luminex™	108.2	
CXCL13	Luminex™	97.4	
CCL2	Luminex™	98.7	
CCL3	Luminex™	-	All observations outside standard range
CCL5	Luminex™	100.2	
CCL7	Luminex™	-	All observations outside standard range
CCL11	Luminex™	-	> 90% of observations outside standard range
CXC3CL1	Luminex™	-	Average coefficient of variation > 50%
FLT3 Ligand	Luminex™	99.2	
G-CSF	Luminex™	100.1	
GM-CSF	Luminex™	-	All observations outside standard range
IFN-γ	Single ELISA	125.1	
IL-1α	Luminex™	99.5	
IL-1β	Luminex™	-	All observations outside standard range
IL-2	Luminex™	-	> 90% of observations outside standard range
IL-4	Luminex™	98.7	
IL-6	Single ELISA	100.0	
IL-10	Luminex™	98.9	
IL12 p70	Luminex™	-	All observations outside standard range
IL-15	Luminex™	99.9	
IL-17	Luminex™	-	> 90% of observations outside standard range
IL-21	Luminex™	-	All observations outside standard range
MMP-8	Luminex™	96.7	
MMP-9	Luminex™	96.7	
Oncostatin M	Luminex™	100.4	
Procalcitonin	Luminex™	99.0	
TNF-α	Single ELISA	100.0	
TNF-β	Luminex™	-	All observations outside standard range
VEGF	Luminex™	99.3	
PD-L1	Luminex™	100.5	
Ferritin	Luminex™	98.6	
Lactoferrin	Luminex™	100.1	

Table 2.5: Cytokines and chemokines identified in cell-free plasma in this study using either Luminex™ multi-plex assays or single ELISA. Where an analyte is excluded from downstream analysis, reasons are given in this table.

2.1.8 Lipid analysis

Lipid concentrations in thawed cell-free plasma were captured by mass spectrometry. Analysis was performed by Ms Linda Moet and kindly provided for inclusion in statistical machine learning models described in Chapter 6. Normalised concentrations for the following lipids were provided: C4 carnitine, C6 carnitine, C8 carnitine, C10 carnitine, C12 carnitine, C2 carnitine, C14 carnitine, C16 carnitine, C18 carnitine, C3 carnitine, C18:1 carnitine, C12-2OH/3OH, C22:6, C18:2, C18:3, C20:5, C18:1, C8:0, C10:0, C12:0, C20:4, C14:0, C16:0, C18:0.

A calibration curve was prepared with concentrations of lipids in the range expected to be encountered in the samples. The calibration curve included internal standards at the same concentration as when added to the samples. Samples were extracted batch-wise in randomized batches. Per batch, five blanks (10 μl methanol) were extracted in parallel with the samples. 10 μl of the sample were mixed with 250 μl of methanol containing internal standards. Samples were sonicated in iced water for 1 minute, vortexed at 1400 rpm for 10 minutes at 4°C, and then centrifuged at 13000 rpm for 10 minutes. 100μl of supernatant were derivatized by adding 50μl 3-NPH solution and 50 μl EDC and pyridine solution and subsequent incubation for 30 minutes at 40°C in a water bath. The reaction was quenched by adding 100μl 0.5% formic acid solution in 75% methanol and incubating for 30 minutes at 40°C in a water bath. Until measurement samples were stored at -20°C. Before quantification on the machine, samples were spun down, and the supernatant was transferred to be measured. A QC sample was prepared by pooling aliquots from all extracted samples. The QC sample was measured after every ten other measurements on the machine.

Peaks were integrated using the MultiQuant software (SCIEX), and concentrations in samples were calculated based on internal standards using Microsoft Excel. For lipids with a corresponding deuterated lipid included in the analysis, concentrations were calculated based on the lipid to IS ratio and internal standard concentration. A calibration curve of the lipid to a similar deuterated lipid ratio was used for other lipids. All further analysis was conducted using R and GraphPad Prism version 8.4.3 for Windows (GraphPad Software, San Diego, California USA) and MetaboAnalyst 4.0 and higher (metaboanalyst.ca).

2.2 Patient immune responses to infection in Peritoneal dialysis (PERIT-PD) study

Data obtained as part of the “patient immune responses to infection in peritoneal dialysis (PERIT-PD)" study were used in this work as a validation for the CytoPy software described in Chapter 3 and a novel ensemble clustering algorithm described in Chapter 4. The principal investigator was Prof. Matthias Eberl, and patients were recruited by the clinical lead, Dr Kieron Donovan. Sample processing was performed by Dr Raya Ahmed, Ms Sarah Baker, and Dr Simone Cuff. Electronic data collection was performed by Dr Raya Ahmed and Dr Simone Cuff, and data analysis by myself.

2.2.1 Ethics and consent

All methods were carried out in accordance with relevant guidelines and regulations and written informed consent was obtained from all subjects. Recruitment of peritoneal dialysis (PD) patients was approved by the South East Wales Local Ethics Committee under reference number 04WSE04/27, and conducted according to the principles expressed in the Declaration of Helsinki. The study was registered on the UK Clinical Research Network Study Portfolio under reference numbers #11838 “Patient immune responses to infection in Peritoneal Dialysis" (PERIT-PD).

2.2.2 Patient recruitment

The study cohort comprised 21 adult individuals receiving peritoneal dialysis (PD) who were admitted between October 2016 and October 2018 to the University Hospital of Wales, Cardiff, on day 1 of acute peritonitis, before commencing antibiotic treatment (47.6% female; median age 53.0 years, range 30.0-86.0 years). 30 age and gender-matched individuals receiving PD and with no previous infections for at least three months served as stable, non-infected controls (53.3% female; median age 59.7 years, range 39.7-84.3 years). Subjects known to be positive for HIV or hepatitis C virus were excluded. Clinical diagnosis of acute peritonitis was based on abdominal pain and cloudy peritoneal effluent with >100 white blood cells/mm3. According to the microbiological analysis of the effluent by the routine Microbiology Laboratory, Public Health Wales, episodes of peritonitis were defined as

infections caused by Gram-positive or Gram-negative organisms. Cases of fungal infection and negative or unclear culture results were excluded from this analysis. A summary of the bacterial culture results for patients with peritonitis is shown in Table 2.6.

Culture result	N
Coagulase-negative Staphylococcus	6
Alpha-haemolytic Streptococcus	3
Staphylococcus aureus	1
Escherichia coli	1
Streptococcus agalactiae	1
Corynebacterium amycolatum	1
Pseudomonas aeruginosa	1
Yeast	1
Mixed growth	2
No growth/unknown	4

Table 2.6: Summary of microbiological culture results for peritoneal dialysis patients with acute peritonitis.

2.2.3 Sample and data collection

Peritoneal leukocytes were harvested from overnight dwell effluents and processed as previously described [175]; samples were treated with DNase (Sigma; 1:2,500 dilution) when excessive debris was visually apparent. Mononuclear cells from peritoneal effluent and PBMCs from whole blood were obtained with density gradient centrifugation using Ficoll (Ficoll-Paque PLUS; Fisher Scientific).

2.2.4 Flow cytometry

Peritoneal leukocytes were stained using monoclonal antibodies against CD1c, CD3, CD14, CD15, CD16, CD19, CD45, CD116, HLA-DR and Siglec-8 (Table 2.7) and identified as $CD45^+$ immune cells, $CD3^+$ T cells, $CD19^+$ V cells, $CD15^-CD14^+$ monocytes/macrophages, $CD15^+$ neutrophils, $CD15^-CD14^{\pm}CD1c^+$ dendritic cells, and $CD15^-$SIGLEC-8^+ eosinophils. T cell subsets in peritoneal mononuclear cells and PBMCs were identified using monoclonal antibodies against CD3, CD4, CD8, TCR-Vα7.2, TCR-Vδ2, TCR-pan-$\gamma\delta$, CD45RA, CCR7, and CD27 (Table 2.8). Cell acquisition by flow cytometry was performed using a 16-colour BD LSR Fortessa cell analyser (BD Biosciences). Live single cells were gated based on side and forward scatter area/heigh, and live/dead staining (fixable Aqua; Invitrogen).

Autonomous gating, supervised classification, and clustering analysis discussed in Chapter 3 were compared to expert-driven manual gating. Gating was performed using FlowJo v10.7 (TreeStar) by two independent experts. The total number of events for each gate of interest were exported as a CSV file. The average number of events between the two independent analysts was used for comparison to autonomous methods.

Antigen	Conjugate	Clone	Manufacturer
CD45	Alexa Fluor 700	2D1	Biolegend
CD14	FITC	63D3	Biolegend
CD16	Per-CP Cy5.5	3G8	Biolegend
CD3	APC/Fire	UCHT1	Biolegend
SIGLEC-8	APC	7C9	Biolegend
CD1c	BV421	L161	Biolegend
CD15	BV605	SSEA-1	Biolegend
HLA-DR	BV711	L243	Biolegend
CD116	PE	4H1	Biolegend
CD19	PE-Cy7	HIB19	Biolegend

Table 2.7: Flow cytometry staining panel for pertioneal leukocytes.

Antigen	Conjugate	Clone	Manufacturer
CD3	APC/Fire	UCHT1	Biolegend
CD4	PE-Cy5.5	OKT4	Biolegend
CD8	BV711	RPA-T8	Biolegend
CD161	APC	191B8	Miltenyi Biotec
Vα7.2	BV605	3C10	Biolegend
TCR-pan-$\gamma\delta$	PE-Cy5	IM2662	Beckman Coulter
Vδ2	PE	B6 RUO	BD Biosciences
CCR7	BV421	G043H7	Biolegend
CD27	PE-Cy7	M-T271	Biolegend
CD45RA	PE Dazzle	HI100	Biolegend

Table 2.8: Flow cytometry staining panel for T cell subsets in peritoneal mononuclear cells and PBMCs.

2.3 Critical assessment of population identification in cytometry data by supervised classification

Supervised classifiers discussed in Chapter 3.3.6 were compared using the Critical Assessment of Population Identification Methods (FlowCAP) challenge data [176]. The FlowCAP-I data consist of four human studies (graft-versus-host disease, diffuse large B-cell lymphoma, symptomatic West Nile virus infection, and healthy donors) and one mouse study (hematopoietic stem cell transplant). Data were labelled and pre-processing performed (removal of debris, dead material, and fluorescence compensation applied) at the source laboratory responsible for acquiring the original data. Here, classifiers were trained on 25% of data and classification performance was tested on the remaining 75%. Performance was reported as the average of macro F1 scores across all five datasets, where the F1 score for data with |C| set of possible classes is given as:

$$macro\ F1\ score = \frac{2}{|C|} \sum_{c \in C} \frac{precision_c \cdot recall_c}{precision_c + recall_c} \tag{2.1}$$

Six supervised machine learning algorithms, housed within the CytoPy software, were compared without hyperparameter tuning:

1. Logistic regression with balanced class-weights; implemented in Scikit-Learn version 0.24 [177]
2. Linear discriminant analysis without any shrinkage and number of components equal to either the number of classes or number of features, depending on which is minimum; implemented in Scikit-Learn version 0.24 [177]
3. Support vector machine with a radial basis function kernel without regularisation and γ as $\frac{1}{n}$ where n is the number of available features; implemented in Scikit-Learn version 0.24 [177]
4. K nearest neighbours classifier with k equal to 30; implemented in Scikit-Learn version 0.24
5. XGBoost using default parameters; implemented in xgboost version 1.2 [178]

6. Feed-forward neural network with three hidden layers of size 12, 6, and 3 nodes, L2 penalty of 1×10^{-4}, softplus activation function on the hidden layers, softmax activation function of the outer most layer, and categorical cross-entropy as the loss function; implemented in Tensorflow Keras version 2.4 [152]

2.4 Statistical analysis

2.4.1 Statistical hypothesis testing

Throughout this thesis, statistical analysis was performed using the Python programming language version 3.8. Hypothesis testing was performed using the Scipy (v1.7) [171] and Pingouin (v0.5) [179] libraries. Graphical representation of data and illustrations were generated using Matplotlib (v3.5) [180], Seaborn (v0.12) [181], and edited using Inkscape (v1.2; inkscape.org). Where required, the R programming language (version 4.1) was employed, and its use is described in the text where appropriate.

All variables were visualised using quantile-quantile plots and tested for univariate normality by the Shapiro-Wilk test. Non-parametric testing was employed for univariate comparisons of survivors and non-survivors, culture-positive and culture-negative sepsis, and Gram-negative cause versus Gram-positive cause in sepsis, in Chapter 5. The Mann–Whitney U test for comparison of independent samples was used to test for statistical significance. Throughout this thesis, a p-value equal to or less than 0.05 was considered significant, but original p-values are reported for transparency. Where the number of comparisons was low (less than 15), p-values were adjusted to control the family-wise error rate using Bonferroni–Holm adjustment. This step-wise procedure reduces the risk of a type I error (falsely rejecting the null hypothesis, i.e. false positives). The Bonferroni–Holm is a conservative method and offers lower statistical power when making many comparisons [182]. Therefore, the Benjamini-Hochberg adjustment was adopted for the analysis presented in Figure 5.6, 5.8, 5.9, 5.10, and 5.11. The Benjamini-Hochberg procedure controls the false discovery rate when performing many comparisons. It reduces the risk of a type II error at a slightly increased risk of type I errors (i.e. it reduces the risk of false negatives at a slightly increased risk of more false positives).

Fisher's exact test was used for comparing proportions of patient categories in sepsis in Chapter 5.3.1. The Fisher's exact test was also used for generating odds ratios when comparing cytokines and chemokines above and below detection thresholds in Chapter 5.3.3. Where only two groups were compared, the Scipy implementation was used. Otherwise, the $fisher.test$ function in R was employed.

2.4.2 Statistical machine learning

The statistical machine learning models described in Chapter 6 were developed using Python's Scikit-Learn (v1.1) library. Models were compared by cross-validation and holdout performance and tested for a significant difference in variance using the non-parametric Friedman test. Where significant, pairwise post hoc analysis was performed using the Nemenyi test. Exact training, testing, and evaluation methodologies are described in full within the results section of Chapter 6.

3 | Development and validation of CytoPy, an open-source framework for cytometry data analysis in Python

3.1 Introduction

Cytometry data analysis has undergone a paradigm shift in response to the growing number of parameters observed in any one experiment. As the field evolves, the traditional method of manual gating by sub-setting single-cell data into populations and encircling data points in hand-drawn polygons in two-dimensional space has proven laborious, subjective, and difficult to standardise. This limitation was realised during data collection for the ILTIS study discussed in the methodology section. The ILTIS study consists of three flow cytometry staining panels, each with 12 or more parameters, collected over three years to characterise the innate immune response in severe sepsis. The complexity of this study presents a significant challenge for timely and accurate data analysis. Other researchers have identified equivalent challenges resulting in a cross-disciplinary effort often termed "cytometry bioinformatics" that addresses such concerns. This new discipline seeks to leverage complex computer algorithms and machine learning to automate analysis and improve the investigator's ability to extract meaning from high-dimensional data.

Where cytometry is used for data acquisition, the typical objective is to discern differences between groups of subjects or experimental conditions or to identify a phenotype that correlates with an experimental or clinical endpoint. To this end, a computational approach to the analysis of cytometry data can take one of two strategies: to group events based on similarity (e.g. cell populations), which then form the variables (often descriptive statistics of the obtained groups) the investigator uses to test their hypothesis, or directly model the acquired multidimensional distribution with respect to a chosen endpoint. Classification strategies can be further subdivided: autonomous gating replicates traditional gating through the use of algorithms (flowDensity [117], OpenCyto [120]); high-dimensional clustering groups events according to their individual phenotypes (FlowSOM [130], PhenoGraph [131], Xshift [134], SPADE [111]); and supervised classification where training on an ex-

ample of manually gated data produces a classifier capable of distinguishing cell populations (FlowLearn [122], ACDC [183], DeepCyTof [125]). Modelling strategies have been successfully adopted in applications such as ACCENSE [129] CellCNN [127], CytoDX [184] and in the work described by Hu *et al* [128]. This approach has the benefit of removing any subjectivity and can be considered truly automated but requires the pooling of sample data and is, therefore, sensitive to batch effects.

In addition, various pieces of software have been developed for data handling, transformation, normalisation and cleaning (e.g. flowCore, flowIO, flowUtils, flowTrans, reFlow, flowAI), visualisation (e.g. ggCyto, t-SNE, UMAP, PHATE), and pipelines for specific applications (e.g. Citrus, MetaCyto, flowType/RchyOptimyx) [186, 132, 105, 185]. However, there is no widespread adoption of these methods yet, nor is there a consensus on applying such techniques, with much of the analysis pipeline left to the individual investigator to establish. This inconsistency results in projects amassing collections of custom scripts and data management that are not standardised or centralised, making reproducing results difficult and making for a daunting landscape for newcomers to the field.

The aforementioned difficulties were faced when addressing the large quantities of single-cell data generated in the immunophenotyping of patients with severe sepsis. In response to this, I developed "CytoPy", a novel analysis framework that aims to address these issues whilst granting access to state-of-the-art machine learning algorithms and techniques widely adopted in cytometry bioinformatics. CytoPy is developed and maintained in the Python programming language, which prides itself on readability and a beginner-friendly syntax. CytoPy incorporates popular data science and machine learning libraries such as Pandas [187], Scikit-Learn [177], and Tensorflow [152], with an application programming interface (API) designed to help expand cytometry bioinformatics in the Python ecosystem. In addition, CytoPy provides convenient access to algorithms that have already gained popularity amongst the cytometry community, such as Phenograph [131], UMAP [137] and FlowSOM [130].

In this chapter, the design and implementation of CytoPy are discussed. The performance of supervised classification of cell populations is benchmarked using the Flow Cytometry: Critical Assessment of Population Identification Methods competition (FlowCAP), a collection of data curated for the assessment of cytometry bioinformatics methods previously used for

validating supervised methods [116]. The FlowCAP competition provides example data that have been heavily pre-processed and is not representative of data encountered in extensive clinical studies. Therefore, CytoPy was also challenged with identifying T cell subsets in PBMCs obtained as part of the “Patient immune responses to infection in Peritoneal Dialysis" (PERIT-PD) study led by Prof. Matthias Eberl at Cardiff University. Data from this study was generated by Dr Raya Ahmed, Dr Simone Cuff and Ms Sarah Baker. This study was chosen because of local expertise regarding the data, prior publications from the group forming a ‘ground truth’ for comparison of findings, and the challenging data with a mixture of staining artefacts and batch effect. After validating individual components, the CytoPy framework is applied in its entirety to characterise the local immune response of patients from the PERIT-PD study.

3.2 Aims

1. Design and implement a novel programming framework for cytometry data analysis with the following qualities:
    (a) A data-centric design with a dynamic database that can scale, and facilitates an iterative analytical environment that tracks the output of multiple complex tasks for improved data standardisation
    (b) Facilitates the implementation of state-of-the-art machine learning algorithms for characterisation of cells based on cell-surface marker expression
        i. Must allow for the use of multiple methodologies and handle the results in such a manner that they can be saved and compared
        ii. Generate an analytical environment that promotes transparency of autonomous results and provides the necessary tools for exploring and criticising results
    (c) Provides seamless integration of clinical/experimental metadata into exploratory data analysis
    (d) Provides a "low-code" interface that reduces the analytical burden of complex cytometry analysis
    (e) Demonstrate that this novel programming framework is capable of fundamental tasks such as handling *.fcs files, applying compensation, and cleaning data of cellular debris and artefacts
2. Provide a strategy for assessing experimental batch effect and methods for reducing its impact on autonomous analysis
3. Critically assess the capabilities of: autonomous gating, supervised classification, and unsupervised classification of cell populations
4. Validate this novel programming framework on real-world data, demonstrating that:
    (a) This new framework meets all the specifications laid out in aim (1)
    (b) This new framework can, at a minimum, confirm the findings of multiple prior immunological studies, thus validating its use

3.3 Results

3.3.1 Design & implementation

Data analysis depends on reliable data management, ensuring reproducible findings and fostering collaboration. A typical cytometry project consists of many Flow Cytometry Standard (FCS) files, clinical or experimental metadata, and additional information generated throughout the analysis (*e.g.* gating, clustering results, cell classification, sample-specific meta-data). A further complication is the iterative nature of cytometry data analysis. To account for this, CytoPy is anchored on a document-orientated database, MongoDB [188]; in this database, data are stored in JavaScript Object Notation (JSON)-like documents in a tree. This design choice has many advantages, including a simplified design, dynamic structure (*i.e.* database fields are not 'fixed' and therefore resistant to unforeseen future requirements), and ease to scale horizontally, thereby improving integration into web applications and collaboration. In this respect, CytoPy depends upon MongoDB being deployed either locally or via a cloud service, and MongoEngine [189], a Document-Object Mapper based on the PyMongo driver.

An overview of the CytoPy framework is given in Figure 3.1 including the recommended pathway for analysis and the pattern followed in subsequent analysis in this thesis (although the modular design allows for individual elements of CytoPy to be used independently). CytoPy follows an object-orientated design with a document-object mapper for commitments to and collection from the underlying database. The user interacts with the database using an interface of classes, each designed for one or more tasks. To accommodate expansion and changing requirements, CytoPy is data-driven whilst algorithm-agnostic, meaning new autonomous gating, supervised classification, clustering, or dimensionality reduction algorithms can be introduced to this infrastructure and applied to cytometry data using one of the appropriate classes. CytoPy makes extensive use of the Scikit-Learn [177] and SciPy [171] ecosystems. Throughout an analysis, whenever single-cell data are retrieved from the database, they are stored in memory as Pandas DataFrames that are accessible for custom scripting at any stage.

Following the steps in Figure 3.1, a typical analysis in CytoPy would be performed as follows (functions shown in italics and class names are shown in italics with title-case):

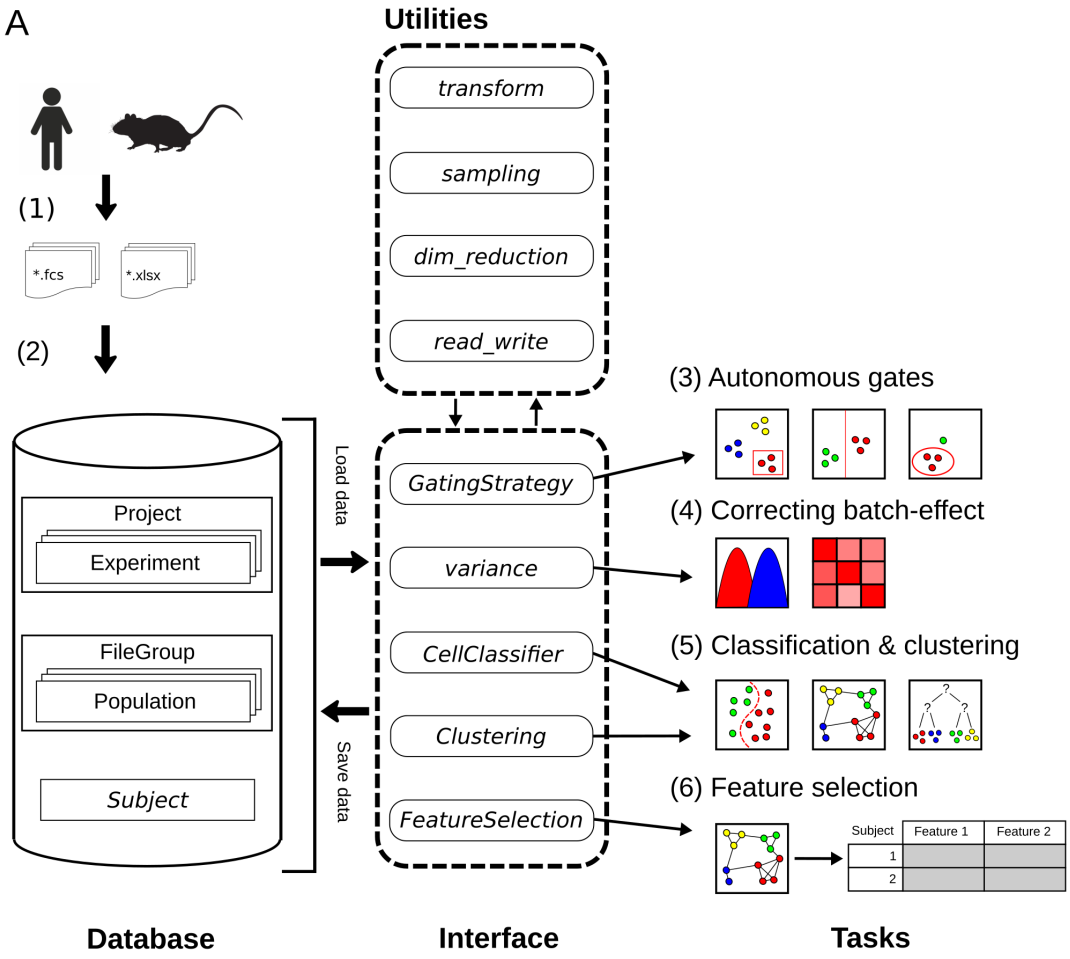

Figure 3.1: Single cell data and experiment/clinical metadata (1) are used to populate a project within the CytoPy database (2). The CytoPy database models analytical data in MonogDB documents (cylinder), and an interface of CytoPy classes retrieves and commits data to this database (dotted rounded rectangle). Utility modules perform regular tasks such as data transformations and sampling throughout the framework. The components of this interface can be used independently, but the recommended workflow is as follows: (3) autonomous gates identify a "clean" population of interest from where to start analysis, (4) batch effect is visualised, quantified and corrected using the Harmony algorithm, (5) supervised and unsupervised algorithms classify cells into groups of similar phenotype, and finally (6) a feature space of cell population descriptive statistics is generated and feature extraction/selection methods deployed to identify a predictive signature that characterises an endpoint of interest.

1. Data are generated and exported from the cytometer; CytoPy supports FCS files version 2.0, 3.0, and 3.1, but additionally supports the introduction of data using a Pandas DataFrame object, therefore supporting wider formats, although this requires that the end user generate this object with suitable formatting. Experimental and clinical metadata are collected in tabular format (Microsoft Excel document or Comma Separated Values (CSV) files), with the only requirement being that metadata be in 'tidy' format.

2. A *Project* is defined and populated with the cytometry data and accompanying metadata. A *Project* houses one or more *Experiment* documents, each defining a set of staining conditions. Each subject (*e.g.* a patient, cell line, or animal) and their associated meta data are contained in a *Subject* document; this document type is dynamic with no restriction's applied to the data they store. A *Subject* can be associated to one or more *FileGroup* documents. The *FileGroup* contains one or more FCS files (or *DataFrames*) associated to a single biological sample collected from the subject; biological samples often have primary stains and then multiple controls such as isotype or Fluorescence-Minus-One (FMO) staining controls, the *FileGroup* is used as an entry point to these related data. Along with the event data, the *FileGroup* stores 'gated' populations, clusters, and meta-information that attains to a single 'sample', including the spill-over matrix for compensation. It should be noted that data are stored on a linear scale with a variety of transformations available during subsequent analysis; this provides flexibility in analysis as the user can compare the effects of different transformations, including the commonly used biexponential (logicle) and hyperbolic arcsine transformations (transformations are implemented using the FlowUtils package [190]).

3. The first step in any cytometry data analysis is cleaning data of debris and artefacts. The FlowAI [142] package provides a preliminary step for removing artefacts and its use is described in full in Flow Cytometry section of the Materials & Methods (2.1.6); throughout this thesis FlowAI is applied prior to using CytoPy. Within CytoPy, manual or autonomous gates can be employed to identify cell populations in one or two-dimensional space, replicating traditional manual analysis conducted with tools such as FlowJo™. The autonomous gates implemented in CytoPy can reduce the time required to perform this initial cleaning and provide a starting population for downstream analysis. Autonomous gates are applied with the *GatingStrategy* class and cell populations are then stored within the database as *Population* documents embedded

within a *FileGroup*. These *Population* documents record the index of events belonging to a population, detail how they were identified, and the conditions in which they were identified such as transformations applied to linear space *e.g.* biexponential or inverse hyperbolic sine transformation of axis.

4. A complication in large studies collecting biological material over a lengthy period is batch effects, which must be addressed prior to analysis. If the batch effect can be minimised by the experimental protocol then the investigator can consider pooling data and modelling the distribution of the event data directly. If batch effects are considerable and cannot be avoided (*e.g.* material is collected over months or years and the integrity will be compromised by freeze-thawing) then computational methods must be used to alleviate batch effect. CytoPy provides tools for visualising and addressing batch effect in the *Variance* module.

5. Classification of events based on a common phenotype can be achieved through a variety of strategies. Methods such as autonomous gating and supervised classification are biased by the training data provided whereas high-dimensional clustering is an unsupervised method that groups cell populations according to their phenotype but can be computationally expensive and outputs are dependent on complex hyperparameters. CytoPy offers a framework where multiple strategies can be applied, contrasted, and compared. The *CellClassifier* class provides an entry point for supervised classification whereas the *Clustering* class offers popular clustering algorithms. These classes are algorithm-agnostic, allowing any function to be applied to data derived from *FileGroup*'s providing they follow specific signatures. Many convenient methods are also provided from visualising and critiquing results; this includes but is not limited to, cross-validation, learning curves, heatmaps, plotting with dimension reduction, and common metrics. Importantly, the results of either strategy generate common *Population* documents that are committed to the database and can be used as input to additional analysis and visualisations.

6. Once cells have been classified, the user can test their hypothesis. Data are summarised into a 'feature space' with summary statistics describing *Populations*. Additional meta-data can be introduced through the *FileGroup* and associated *Subject* documents thanks to the database design. This generates a large number of variables,

many of which will be either uninformative or redundant. Filter and wrapper methods are available through the *feature_selection* module finding only those variables that are important for predicting a biological or experimental endpoint. This module deploys methods from the discipline of interpretable machine learning such as L1-regularised linear models and feature importance derived from ensembles of decision trees.

3.3.2 Identifying batch effect in blood T cell subsets

To validate the individual components of CytoPy, I sought to identify T cell subsets in PBMCs from 14 individuals from the PERID-PD study (chosen based on available data); 4 presented with symptoms of acute peritonitis, whereas the remainder were stable and asymptomatic, providing a mixture of inflammatory and stable immune landscapes, and a class imbalance. The objective was to identify T cells (single live $CD3^+$ cells) in the first instance then subsequently identify $CD4^+$ T helper cells, $CD8^+$ cytotoxic T cells, $V\alpha\ 7.2^+$ $CD161^+$ mucosal-associated invariant T (MAIT) and $V\delta2^+$ $\gamma\delta$ T cell subsets. These populations were chosen to test a range of functionality: the ability to identify large and easy-to-distinguish cell populations ($CD4^+$ and $CD8^+$ T cells) and more complex cell types that can be rare in some patients and difficult to identify reliably in two-dimensional space ($V\delta2^+$ $\gamma\delta$ T cells and MAIT cells). Performance was compared to manual gates decided by user expertise (see Methods & Materials section 2.1.6).

Before any extensive analysis can be conducted, it is important to check for batch effects that could influence the results. Batch effects are best addressed in the experimental protocol, ideally by reducing the number of batches performed, achieved by either processing on the same day or freezing material for bulk analysis. In the PERIT-PD data, batch effects were suspected, given that data were collected over 24 months by multiple personnel. Observation of individual fluorochromes (Figure 3.2A) show "drift" in fluorescent intensity of multiple channels and UMAP plots (Figure 3.2B) demonstrates how this extends into the multivariate space. In each plot, data are transformed into the same space and compared to a reference (blue); the reference was chosen by computing the pairwise Euclidean distance of the set of variance matrices for each sample and selecting the sample with the smallest average distance to all others [125]. The UMAP plots revealed common structures shared between patients but a lack of alignment, suggesting noise infiltration from technical variation. Batch

effects, like the example in Figure 3.2, have impeded autonomous cytometry data analysis and must be addressed with a data transformation step.

3.3.3 Autonomous gates

The most straightforward approach to the autonomous analysis of cytometry data is a replication of traditional manual gating by applying computer algorithms to data in sequences of one or two-dimensional space. Such methodologies have been demonstrated previously [120, 117] and are improved upon in CytoPy. The *Gate* object is used to implement a single algorithm for identifying one or more populations in one or two dimensions. *Gate* objects can then be 'stacked' within a *GatingStrategy*, saved to the database, and applied in sequence to subsequent data. Each *Gate* is defined using example data, and an algorithm is chosen that best encapsulates the population of interest. Figure A gives an example of a polygon gate, which can leverage any clustering algorithm (*e.g.* K-Means, hierarchical clustering, DBSCAN etc.) or probabilistic models that can divide data into components (*e.g.* mixture models). The example in Figure 3.3 (left) uses K-means clustering to define five polygon gates, the red gate is chosen, and its information is saved within the *Gate* and committed to the database. The shape that forms this gate is created by computing the α shape of the cluster. The α shape is a straight line graph that captures the 'crude shape' of a finite set [191] and the behaviour of this graph can be modified by changing the α parameter (Figure 3.3B); a value of 0 creates a convex-hull, equivalent to wrapping an elastic band around the points of a cluster, but as α increases, the shape takes a 'tighter' fit. By default, CytoPy will set α to 0, which helps prevent biasing the shape formed by the reference data.

Upon exposure to new data (Figure 3.3A; right), K-means is reapplied, polygon gates are generated, and the gate most similar to the original reference is chosen. The similarity is measured by comparing the Hausdorff distance between the reference gate and newly generated gates and selecting the gate of minimal distance to the reference:

$$\min_{G_1 \ldots G_n} (h(R, G_n)) \tag{4.1}$$

Where R is the reference gate (a gate being a set of two-dimensional coordinates defining the polygon) and G is a newly generated gate, of which there can be n.

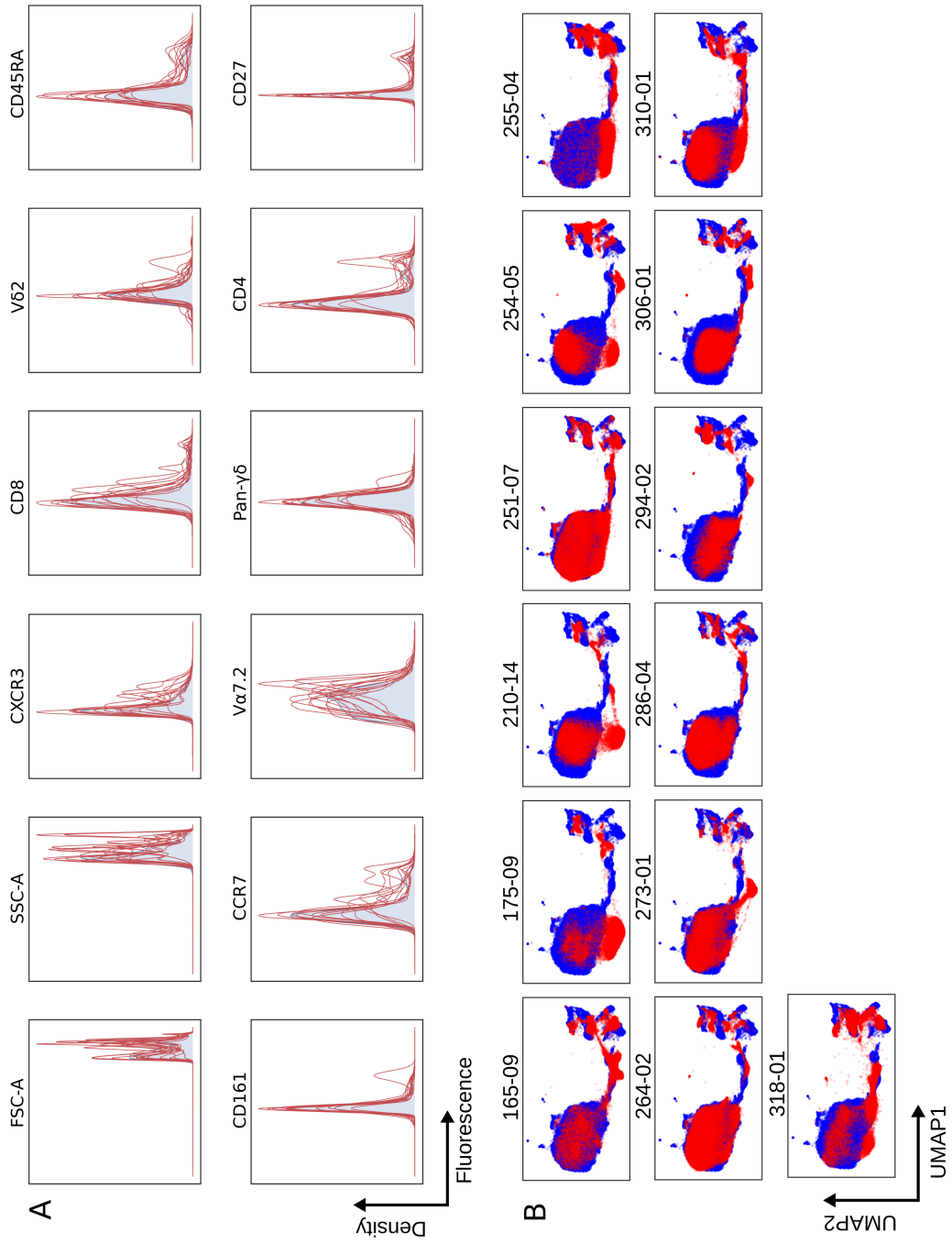

Figure 3.2: (A) Probability density function for individual fluorochromes is shown for 13 subjects (red) compared to a chosen reference subject (blue). The multivariate space for the same 13 patients are shown as UMAP plots in (B), overlaid on the reference subject (blue).

$$h(R, G_n) = \max_{r \in R}(\min_{g \in G}(\sqrt{\sum(r-g)^2})) \tag{4.2}$$

The Hausdorff distance, given in equation 4.2, defines the distance between the reference gate and some other gate as the maximum distance of the reference set to the nearest point in the comparison set; where the euclidean distance is used to compare two points.

Figure 3.3C provides an example of a reference gate (left), and the same gate overlaid on two gates defined using new data (right; green and black). Choosing the gate that best fits the population of interest, as defined in the reference data, is a complex task; the green gate is the obvious choice, yet computationally both shapes overlap the reference gate, and the centroid of both are comparable to the reference. However, by minimising the Hausdorff distance as described above (1.596 for the green gate and 3.574 for the black; units reported after transformation of the space by inverse hyperbolic sine), the most suitable gate is chosen.

As an alternative to polygon gates, CytoPy also implements threshold gates (Figure 3.3D) that divide data within one or two-dimensional space based on properties of the probability density function (PDF; as estimated using a fast convolution-based kernel density estimation [192]). CytoPy uses an adaption of flowDensity [117]. After estimating the PDF, a peak finding algorithm identifies major landmarks (this can be tuned by hyperparameters that control the peak detection limit) and applies a threshold at either the local minima between two peaks or the inflection point on either side of a peak (controlled by a hyperparameter) if only one peak is identified. If more than one peak is identified, the PDF is smoothed using a Savitzky-Golay filter [171] until two or fewer peaks are identified. Similar to the polygon gate, the threshold gate is defined using reference data. The algorithm is run when exposed to new data, generating new thresholds. The resulting populations are matched to the reference based on definition, e.g. a population right of a threshold is labelled “positive”. Therefore, a population right of the threshold in new data would also be labelled “positive”.

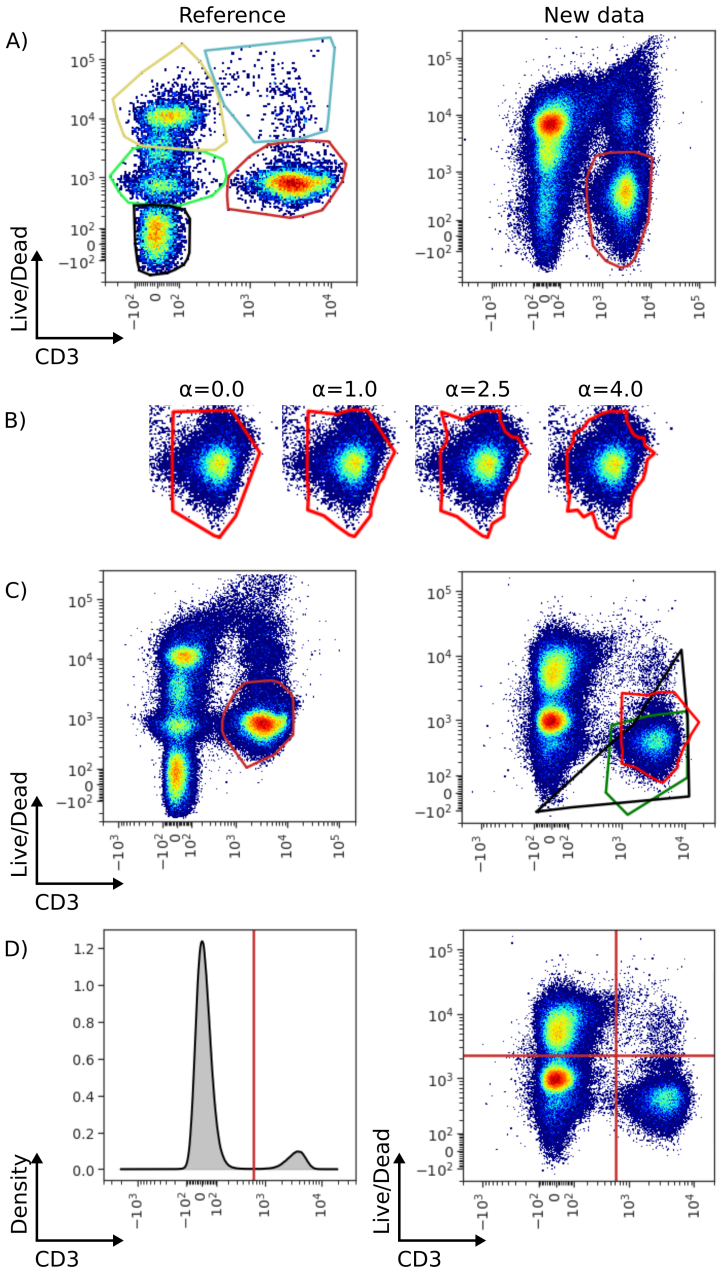

Figure 3.3: (A) An example of autonomous gating using polygon gate's generated by a K-Means clustering algorithm. Reference data provides a template for expected populations (left) and one or more are chosen to match the identification of the equivalent population in new data (right). Polygon gates are defined by the α shape of the clustered set and the α parameter can be adjusted (B) to control how tight a gate fits to a population. (C) Example of a polygon gate (left) and the same gate (right; red) alongside two gate's covering a similar region (right; green and black) that are compared using Hausdorff distance. (D) Threshold gate's are an alternative method that can be used to identify similar populations to polygon gates but divide events purely on properties of their probability density functions.

3.3.4 Autonomous gates can reliably identify T cell subsets by addressing batch-effect with hyperparameter search and landmark registration

A challenge when defining autonomous gates is the choice of hyperparameters that will generalise beyond the chosen example data; batch effects further exacerbate this. CytoPy employs two techniques to overcome this issue: hyperparameter search and landmark registration. Hyperparameter search allows the user to specify a range of hyperparameters when a *Gate* is applied to new data. An exhaustive search is performed across all permutations of chosen hyperparameters resulting in a set of populations. In the case of a polygon gate, each population's α shape is computed and matched to reference α shapes by minimising the Hausdorff distance. Regarding a threshold gate, the euclidean norm is computed for all pairs of reference populations and populations generated by hyperparameter search. Reference populations are then matched to populations where this norm is minimum.

Batch effects can introduce significant variation in the distribution of populations in the one or two-dimensional space where a *Gate* is applied. A strategy for mitigating batch effect during autonomous gating was proposed by Hahne *et al.* [193]. They describe the use of landmark registration, a technique in functional data analysis that can align two functions according to some shared landmark(s). Following this example, landmark registration was implemented in CytoPy to align data to some common reference data prior to applying a gate. Landmarks are identified as points of maximum density and grouped by a K means algorithm [193]. Once typical landmarks are identified between the target and reference data, a warping function is found using monotonic cubic interpolation (Figure 3.4) and function composition used to 'adjust' the data. Hahne *et al.* [193] applied landmark registration to all available data and then gated populations. However, this can distort smaller sub-populations, so CytoPy follows the method described by Finak *et al.* [194]; a localised approach, with landmark registration, applied prior to applying each gate.

Autonomous gates were applied to identify T cell subsets in PBMCs, whilst employing landmark registration and hyperparameter search to address variation between biological specimens (Figure 3.5). The number of events identified by autonomous gates (Figure 3.5B; x-axis) was compared to the same population identified by manual gates (Figure 3.5B; y-axis);

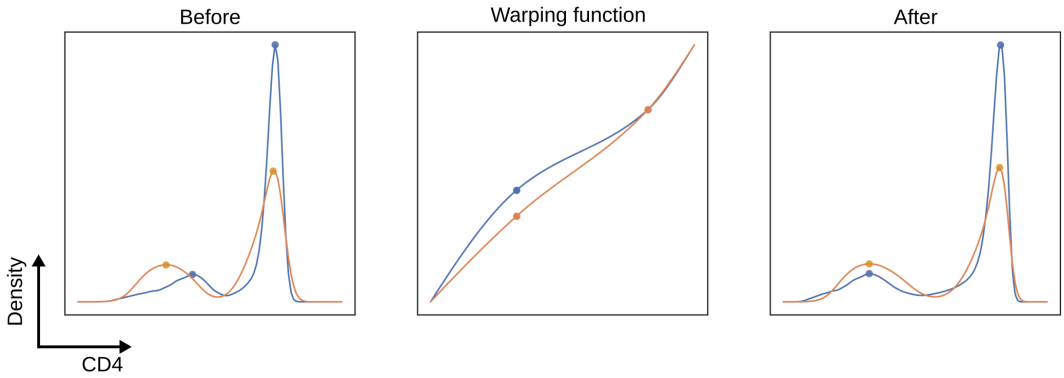

Figure 3.4: The original PDFs (left) show two landmarks that points of maximum density, located at the $CD4^-$ and $CD4^+$ populations. Landmark registration identifies a warping function (middle) that, when taken in composition with the original functions, generates a aligned distributions (right) that mitigate batch effect.

three experts gated populations, and the average events per population was taken as the manually gated result. Each data point is an individual patient. Autonomous gates showed good conformity with manual gates for live $CD3^+$, $CD8^+$, and $CD4^+$ T cell subsets and reasonable performance for $V\delta2^+$ $\gamma\delta$ T cells and MAIT cells, although more varied.

3.3.5 Addressing batch effect with the Harmony algorithm

Despite the success of autonomous gates for identifying T cell subsets in the wake of significant batch effect, they are heavily biased by choice of example data when defining *Gate* objects and by choice of reference for landmark registration. An alternative approach to addressing batch effects is to align cell populations between individual subjects in high dimensional space prior to analysis. Several methods have been proposed with this objective [195], most prominently applied to single-cell RNA sequencing data, although some examples such as SAUCIE [126] demonstrated application to cytometry data.

The Harmony algorithm [196, 197] algorithm was chosen for implementation in CytoPy, given its ability to scale to data of modest size ($\approx 10^6$ events on a personal laptop) and its transparent hyperparameters. Harmony was initially described as being applied to low-dimensional embeddings. Embedding with methods such as PCA is necessary for RNA sequence data where the number of available features can be in the thousands or tens of thousands but is not necessary for cytometry data with only a dozen or more parameters. Therefore, in CytoPy, the original data were exposed to Harmony after removing debris,

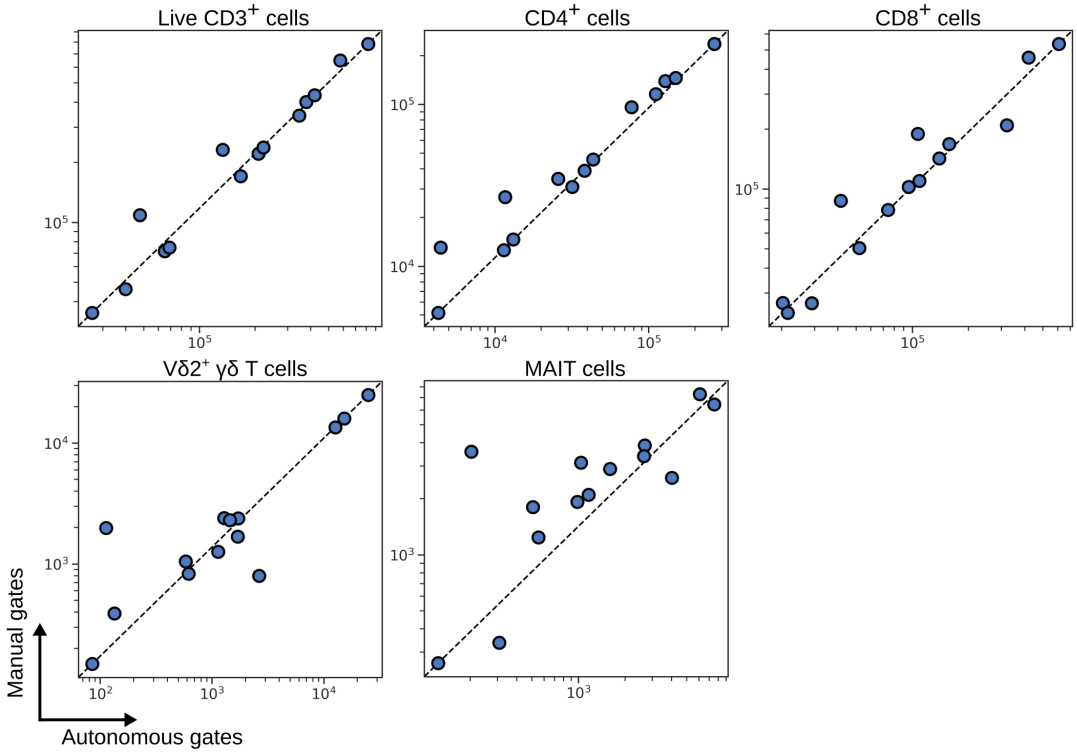

Figure 3.5: Each symbol depicts results obtained with cells from an individual patient.

doublets, and dead cells. Biexponential (logicle) transformation followed by scaling each parameter to unit variance (subtracting the mean and dividing by standard deviation) was performed prior to batch effect correction.

Harmony then attempts to mix batches within the same space whilst maintaining the purity of cell types within each population. Harmony achieves this through an iterative algorithm of soft clustering, centroid identification, correction, and data movement based on soft clustering membership. The original authors propose the local inverse Simpson's Index (LISI) to quantify the integration of batches, which defines the effective number of batches represented in a local neighbourhood; a value of 1 would indicate that cell neighbourhoods consist of a single batch (poorly integrated) whereas a value of say 5, would indicate that the neighbourhood is a mixture of 5 batches (better integration).

The performance of Harmony when applied to our T cell population (as identified by autonomous gates) from PBMCs is shown in Figure 3.6. Harmony has a range of hyperparameters that influence its behaviour. The default values for most of these parameters provided good performance but varying σ further improved performance; this hyperparameter influences the entropy regularisation term of the soft-clustering step of the algorithm, and as it

approaches zero, clustering is more alike to hard K means clustering. For the T cell data discussed here, an optimal value of 0.2 for σ was chosen whilst limiting the number of iterations to 5. The quality of batch correction was determined by visual inspection of batches in embedded UMAP space before and after correction and by comparing the LISI before and after correction (Figure 3.6A). The objective here was to redistribute LISI such that the local neighbourhood around a cell contains a greater representation of different batches without over-correcting and distilling biological variation that differentiates groups of subjects.

The UMAP plots in Figure 3.6A show that large communities of cells consist of single batches before applying Harmony. In contrast, these communities are diffused after application yet maintain a topology of separate cell populations. The concern with batch correction is over-correction that disrupts the biological meaning within the data. However, batch corrected data embedded in UMAP plots was coloured by fluorescent intensity for markers that can identify T cell subsets (Figure 3.6B), and not only do large populations such as $CD4^+$, and $CD8^+$ T cells remain identifiable but smaller and harder to distinguish subsets such as $V\delta2^+$ $\gamma\delta$ T cell and MAIT cells are also visible.

3.3.6 Supervised methods for classification of cytometry data

Cytometry instruments are capable of generating millions of data points for each experiment. Such extensive data offers the opportunity to leverage supervised machine learning techniques that require significantly large training data sets. The limitation of this method is that data must be accurately labelled to provide an objective for the learning algorithm. A labelled example can be produced through autonomous or manually gating. However, the resulting model will not generalise if there are significant batch effects, so data was corrected using Harmony before training a supervised model.

Supervised classification of cytometry data is available in CytoPy through the *CellClassifier* class. Objects of this class can accept any classifier that conforms to/supports the Scikit-Learn API (such as XGBoost), or a Keras [198] model. Many convenient methods are pre-built into those objects (including methods for evaluating classifier performance such as cross-validation, learning curves, and confusion matrices; Figure 3.7), and predictions can be saved as *Population* objects, providing compatibility with all other tools in the CytoPy framework.

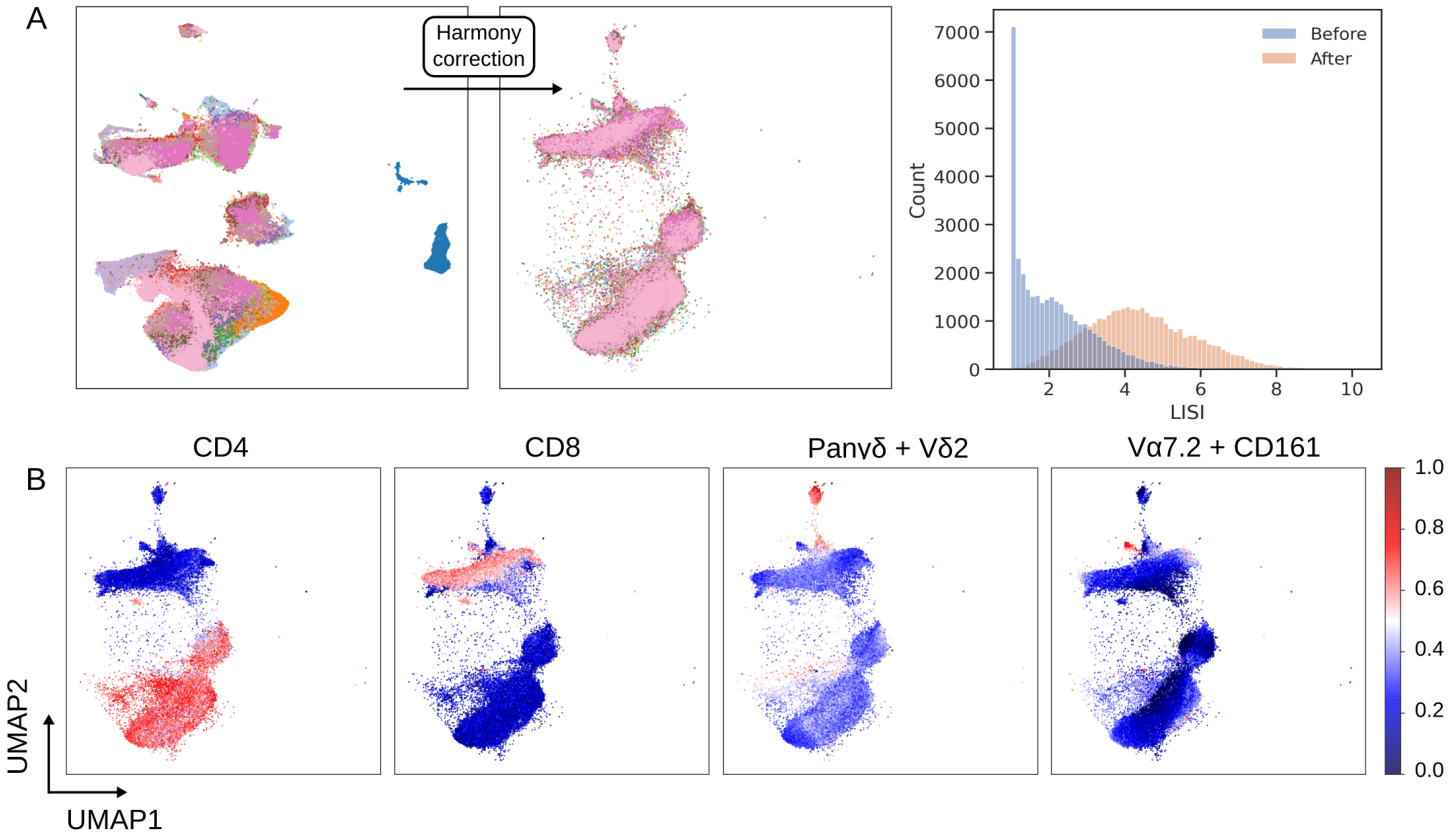

Figure 3.6: (A) Single cell UMAP plots are coloured by cell origin, where each colour represents a unique patient. Shift in batch membership in the local neighbourhood of cells is shown by the change in the UMAP plot after Harmony is applied and by the shift in LISI distribution. (B) Cell population structure is conserved after correction as shown by the shape of latent variables UMAP1 and UMAP2, and the distribution of the cell surface markers CD4, CD8, the linear combination of Pan-$\gamma\delta$ and Vδ2 (to identify Vδ2+ $\gamma\delta$ T cells), and the linear combination of CD161 and V7.2 (to identify MAIT cells).

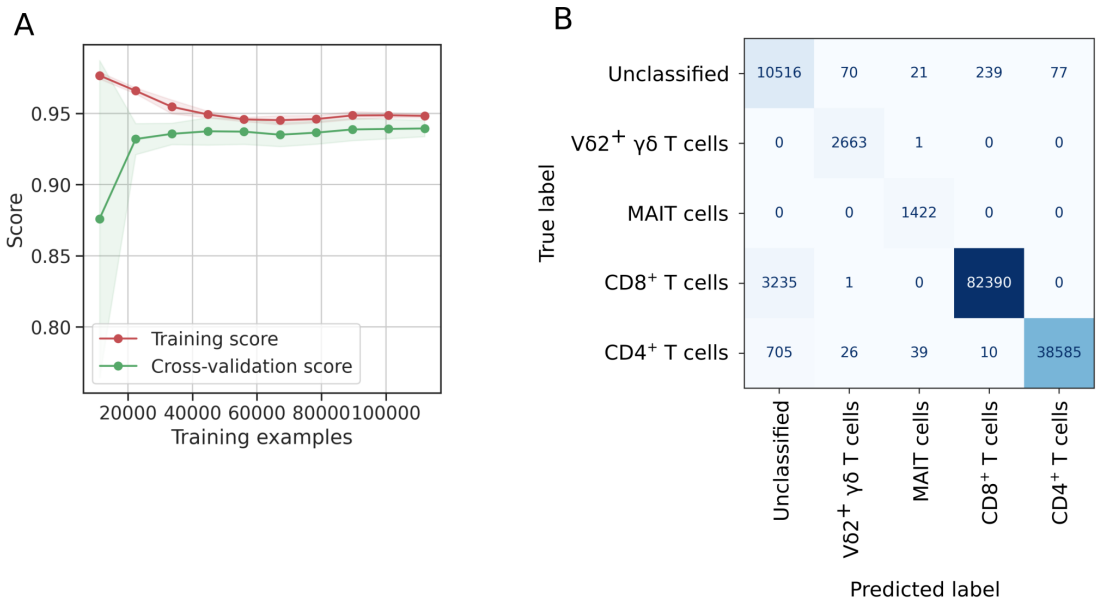

Figure 3.7: Learning curves (A) compare training and cross-validation score, giving insight into potential over-fitting and the benefit brought by introducing more training data. Confusion matrices (B) help identify cell subsets that are miss-classified, a problem exacerbated by class imbalance (rare cell subsets).

To benchmark this supervised approach, classifiers were applied using CytoPy and compared to those reported in the Flow Cytometry: Critical Assessment of Population Identification Methods competition (FlowCAP) [116]. Using CytoPy, four native classifiers from Scikit-Learn (logistic regression, linear discriminate analysis, support vector machine with radial kernel, and K-nearest neighbours), XGBoost [178], and a deep feed-forward neural network (architecture as described by Huamin *et al.* [125]) were chosen for comparison. Algorithms were chosen from a range of classifier families based on their popularity in the literature. Table 3.1 reports the F1 score weighted by support (the number of true instances) for each classifier across the five example datasets from FlowCAP. The deep neural network showed good performance as previously reported [125]. However, XGBoost was the superior method when applied to FlowCAP data and highlights the ability of this classifier to generalise to a wide range of use-cases.

I then tested the utility of XGBoost on 'real-world' data by classifying T cell subsets and comparing the outputs to expert manual gating. Since Harmony accounted for batch effects, data were pooled from all available samples to generate training data manually labelled using the gating infrastructure within CytoPy. Figure 3.8 demonstrates the capability of XGBoost to identify T cell subsets. Since the computational complexity of batch effect correction with

Classifier	GvHD	DLBCL	HSCT	WNV	ND	Mean
RadialSVM*	0.89 (0.83, 0.95)	0.84 (0.80, 0.87)	0.98 (0.96, 0.99)	0.96 (0.94, 0.97)	0.93 (0.92, 0.94)	0.92
flowClust/Merge*	0.92 (0.88, 0.95)	0.92 (0.89, 0.94)	0.95 (0.92, 0.97)	0.84 (0.82, 0.86)	0.89 (0.88, 0.90)	0.90
randomForests*	0.85 (0.78, 0.91)	0.78 (0.74, 0.83)	0.81 (0.79, 0.83)	0.87 (0.84, 0.90)	0.94 (0.92, 0.95)	0.85
FLOCK*	0.82 (0.77, 0.87)	0.91 (0.89, 0.93)	0.86 (0.76, 0.93)	0.86 (0.82, 0.89)	0.86 (0.77, 0.92)	0.86
CDP*	0.78 (0.68, 0.87)	0.95 (0.93, 0.97)	0.75 (0.71, 0.78)	0.86 (0.84, 0.88)	0.83 (0.80, 0.86)	0.80
Ensemble clustering*	0.91	0.94	0.95	0.92	0.94	0.93
Logistic regression	0.93 (0.92, 0.95)	0.94 (0.93, 0.95)	0.97 (0.96, 0.97)	0.91 (0.90, 0.92)	0.82 (0.81, 0.83)	0.91
Linear discriminant analysis	0.94 (0.92, 0.97)	0.98 (0.97, 0.98)	0.95 (0.93, 0.97)	0.92 (0.9, 0.94)	0.81 (0.80, 0.82)	0.92
Radial SVM	0.97 (0.96, 0.98)	0.98 (0.97, 0.98)	0.97 (0.96, 0.97)	0.97 (0.96, 0.97	0.91 (0.90, 0.91)	0.96
K nearest neighbours	0.95 (0.93, 0.97)	0.97 (0.96, 0.98)	0.94 (0.93, 0.96)	0.95 (0.94, 0.95)	0.89 (0.89, 0.90)	0.94
XGBoost	0.99 (0.98, 0.99)	0.98 (0.97, 0.98)	0.99 (0.99, 0.99)	0.99 (0.98, 0.99)	0.99 (0.98, 0.99)	0.99
Feed-forward deep neural net	0.96 (0.93, 0.97)	0.93 (0.91, 0.95)	0.97 (0.96, 0.98)	0.98 (0.98, 0.99)	0.92 (0.92, 0.93)	0.95

Table 3.1: FlowCAP data (GvHD, DLBCL, HSCT, WNV, and ND) are described in full in Materials & Methods section 2.3.
*Performance from the original competition as reported by Aghaeepour N *et al* [116]; all other algorithms are implemented through CytoPy.

Harmony requires a down-sampling step, comparisons are shown as the percentage of T cells observed by manual gates vs populations identified by XGBoost. The output of XGBoost is comparable to manual gating and the results obtained from autonomous gates, although there are some discrepancies for V$\delta 2^+$ $\gamma\delta$ T cell and MAIT cells; this reflects an inability for supervised models to generalise in some instances, especially where abnormal staining or artefacts disrupt the distribution of populations.

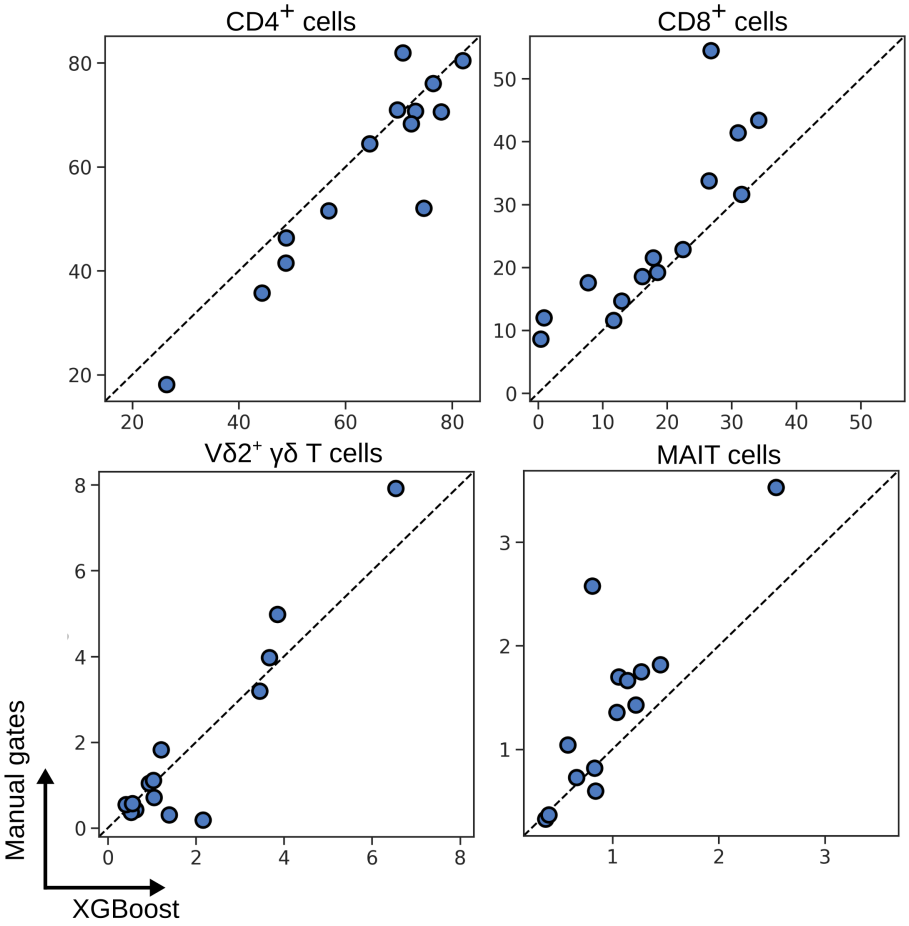

Figure 3.8: Each symbol depicts results obtained with cells from an individual patient.

3.3.7 Unsupervised clustering of cytometry data

Autonomous gates and supervised classification can identify known populations of interest but are biased by the investigator's understanding and expectations of the immune landscape. CytoPy encourages using unsupervised techniques where discovery and exploratory analysis are the objectives to diminish this bias. However, nothing stops an investigator from using both directed and undirected analysis. The data-driven design of CytoPy is such that multiple methods can be executed in parallel and generate comparable *Population* objects.

Unsupervised clustering has gained popularity in single-cell data analysis, and methods such as FlowSOM [130], and Phenograph [131] appear regularly in cytometry data analysis. Similar to the *CellClassifier* class, a *Clustering* class offers an algorithm-agnostic infrastructure for the application of clustering algorithms and generates the same common *Population* objects for compatibility with other tools in the framework. FlowSOM and Phenograph are implemented within CytoPy for convenience, but any algorithm that follows the signatures of the Scikit-Learn ecosystem can be applied, future-proofing CytoPy; this design choice reflects the rapidly changing landscape of cytometry bioinformatics and was chosen so new techniques can easily be integrated into existing infrastructure.

The *Clustering* class offers two different approaches to analysis: the data from multiple subjects can be pooled, and the clustering algorithm applied to this joint space or data from each subject can be clustered independently and clusters matched with a meta-clustering approach as described by [131]; in brief, the centroid is found for every cluster from every subject, and then centroid's are clustered further to generate the final clustering consensus. Clustering requires that the entire data be held in memory during computation. Since meta-clustering offers a per-subject clustering step, this is more accessible to those with modest computers with limited memory. The cost of this approach is reduced sensitivity because of information loss when reducing clusters to centroids.

Unsupervised clustering using the CytoPy software was validated by identifying T cell populations in batch corrected data using FlowSOM and Phenograph. The results of meta-clustering of Harmony corrected T cells are shown in Figure 3.9. The UMAP plots (left) show individual clusters as obtained from individual subjects but plotted in the same two-dimensional space and coloured by meta-cluster membership; the data point size corresponds to the proportion of events as a percentage of T cells in each individual.

A comparison of the proportion of cells obtained by FlowSOM and Phenograph to the same cell type identified by manual gates (Figure 3.10) showed that Phenograph and FlowSOM could reliably identify $CD4^+$ and $CD8^+$ T cells for the majority of instance but struggled with rare cell populations such as MAITs and $\gamma\delta$ T cells; MAITs are under-represented by FlowSOM and Phenograph overestimated the proportion of Vδ2+ $\gamma\delta$ T cells in several patients.

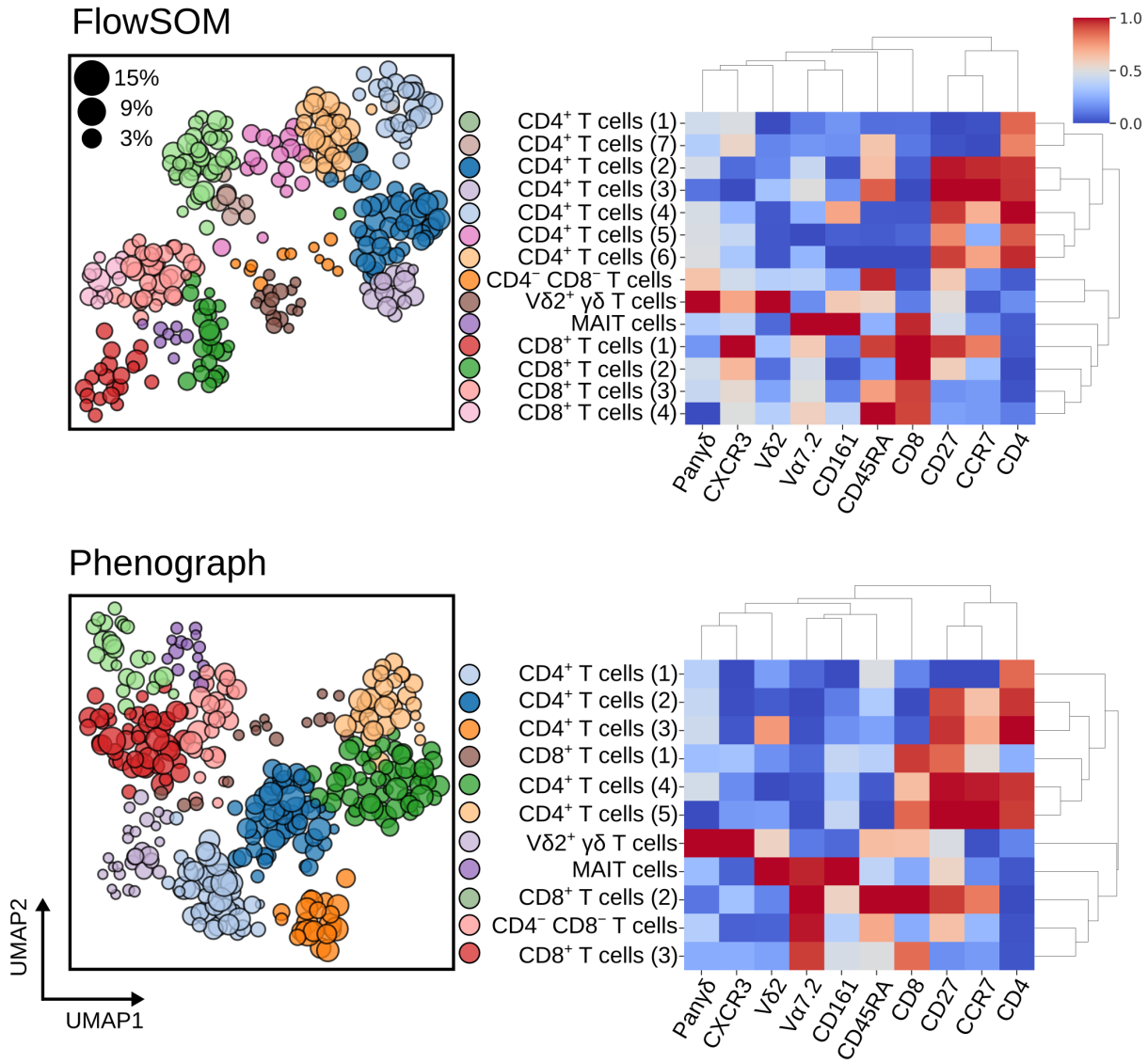

Figure 3.9: Meta-clustering results for FlowSOM (top) and Phenograph (bottom) when applied to blood T cells after batch effect correction with Harmony. Heatmaps show the normalised expression of cell surface markers for meta-clusters (clustered centroids of individually clustered patient samples). In the neighbouring UMAP plots, clusters from all patients are shown in the same embedded space and coloured by their meta-cluster membership. The size of each data point corresponds to the percentage of T cells this cluster represents in the patient it was derived from.

3.3.8 Implementing the CytoPy framework to identify an immune signature that differentiates patients with acute peritonitis from stable controls

The application of CytoPy to an immunophenotyping project was demonstrated with the investigation of the peritoneal effluent of patients undergoing peritoneal dialysis, some of

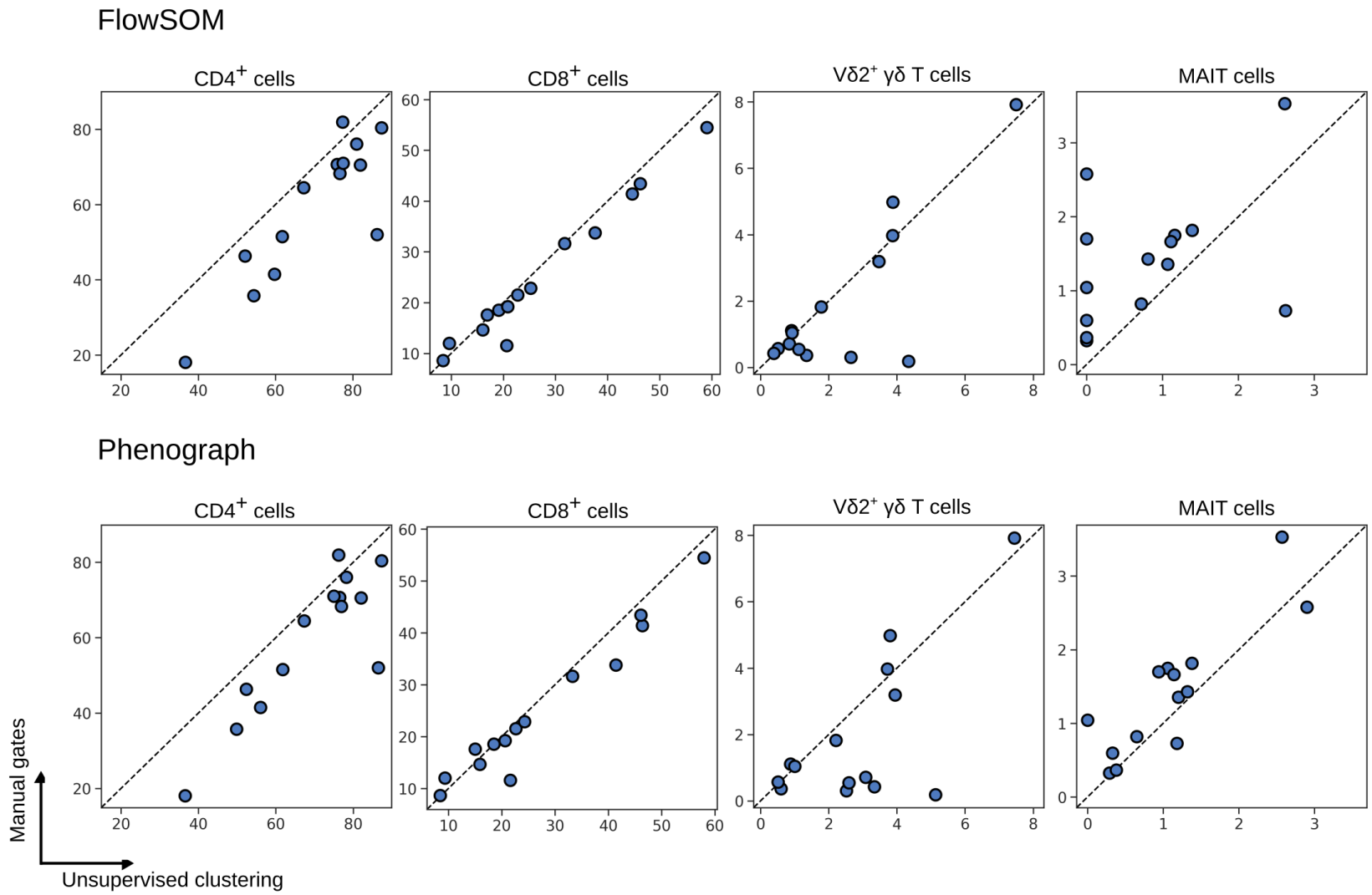

Figure 3.10: Percentage of T cell subsets as identified by FlowSOM (top) and Phenograph clustering (bottom), compared to the same subsets as identified by expert manual gates. Each symbol depicts results obtained with cells from an individual patient.

whom presented with symptoms of acute peritonitis, with the objective to distinguish patients with acute peritonitis from stable controls based on their peritoneal immune signatures. The data were chosen based on our group's long-standing expertise and published findings demonstrating the significance of the local immune response in peritonitis, recognising pathogen-specific infection patterns [199], and the correlation between changes in myeloid populations and treatment failure [200].

Peritoneal effluent was stained using two flow cytometry panels (see Materials & Methods section 2.2.4) to quantify major leukocyte subsets and, more specifically, T cell subsets. $CD45^+$ fraction of cells from total effluent and T cells were obtained by autonomous gating prior to batch correction with Harmony (Figure 3.11). Following batch correction, XGBoost classification using manually gated training data and FlowSOM and Phenograph clustering were performed to quantify cell subsets. All three methods agreed on significant differences in the proportion of neutrophils, monocytes, and T cells in peritoneal effluent when comparing stable controls with those with peritonitis (Figure 3.12). The proportion of T cell subsets was not significantly different between stable controls and those presenting with acute peritonitis (Figure 3.13).

The live $CD45^+$ fraction (for T cells, B cells, monocytes, neutrophils, and eosinophils) and T cell fraction for $CD4^+$, $CD8^+$, $V\delta2^+$ $\gamma\delta$ T cells, and MAIT cells) across the three classification methods were pooled and averaged using the *feature_selection* module to generate a feature space representative of the local immune profile of the peritoneum. Age and gender were included in this feature space as potential confounding variables. High collinearity was observed between the fraction of $CD4^+$ and $CD8^+$ T cells, monocytes and DCs, and T cells and B cells (Figure 3.14A). $CD8^+$ T cells, DCs, and B cells showed low variability and were therefore removed from the analysis. With the remaining features, principal component analysis (PCA) was performed, showing that patients with acute peritonitis were highly discernible from stable controls along the axis of the first principal component (Figure 3.14B). The absolute value of the coefficients for this component showed that neutrophils contributed the most to the observed variation. I generated a linear support vector machine with an L1 regularisation term to confirm these findings. The regularisation parameter, C, was varied, and the coefficient of each feature was plotted; as the value of C decreases, a sparse model is encouraged, eliminating features that do not contribute to the prediction. Figure 3.14C demonstrates that the neutrophil fraction is the only feature to persist in a constrained model.

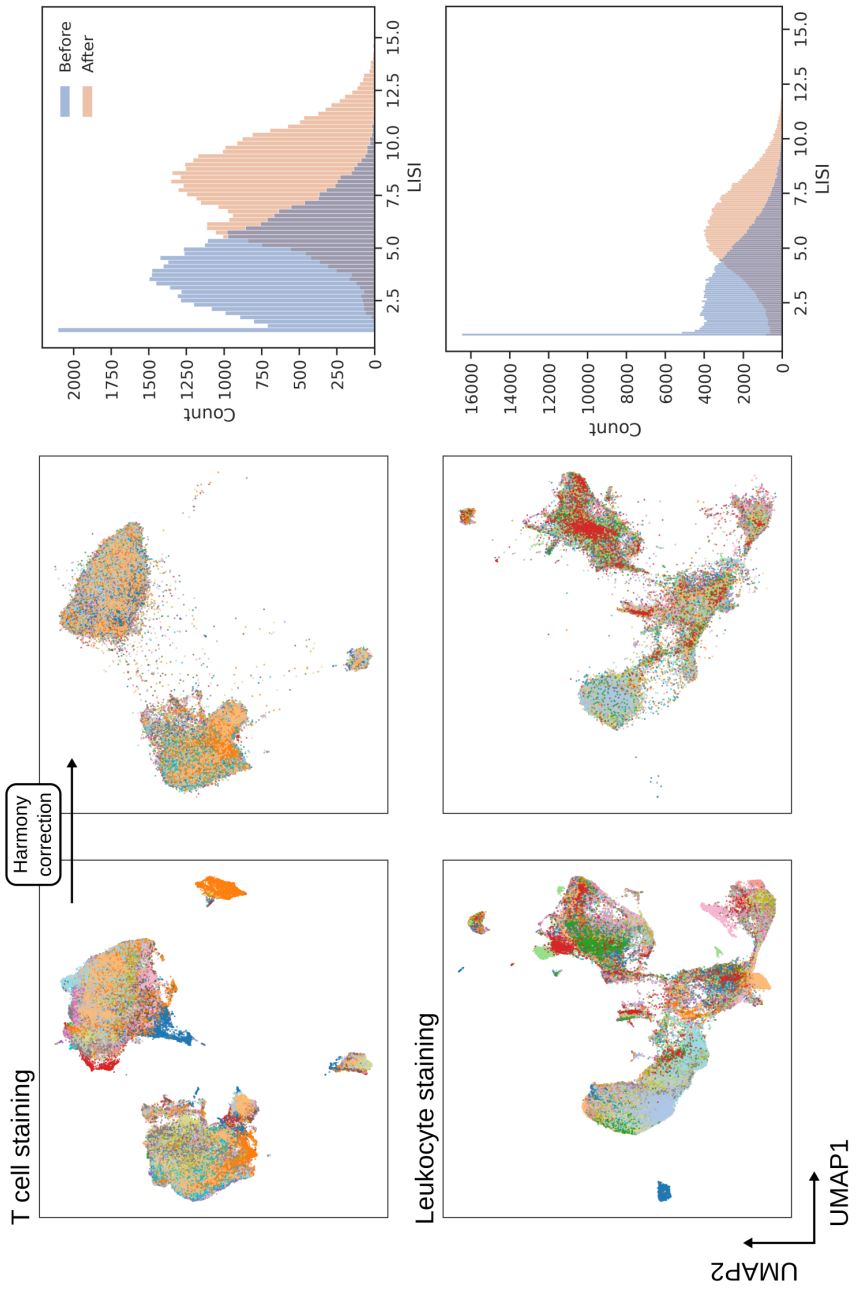

Figure 3.11: Batch correction of T cells (top) and Leukocytes from effluent (bottom) with Harmony. UMAP plots (left) are coloured by cell origin, where each colour represents a unique patient, and show before and after correction. Histograms (right) shows the LISI distribution before and after correction.

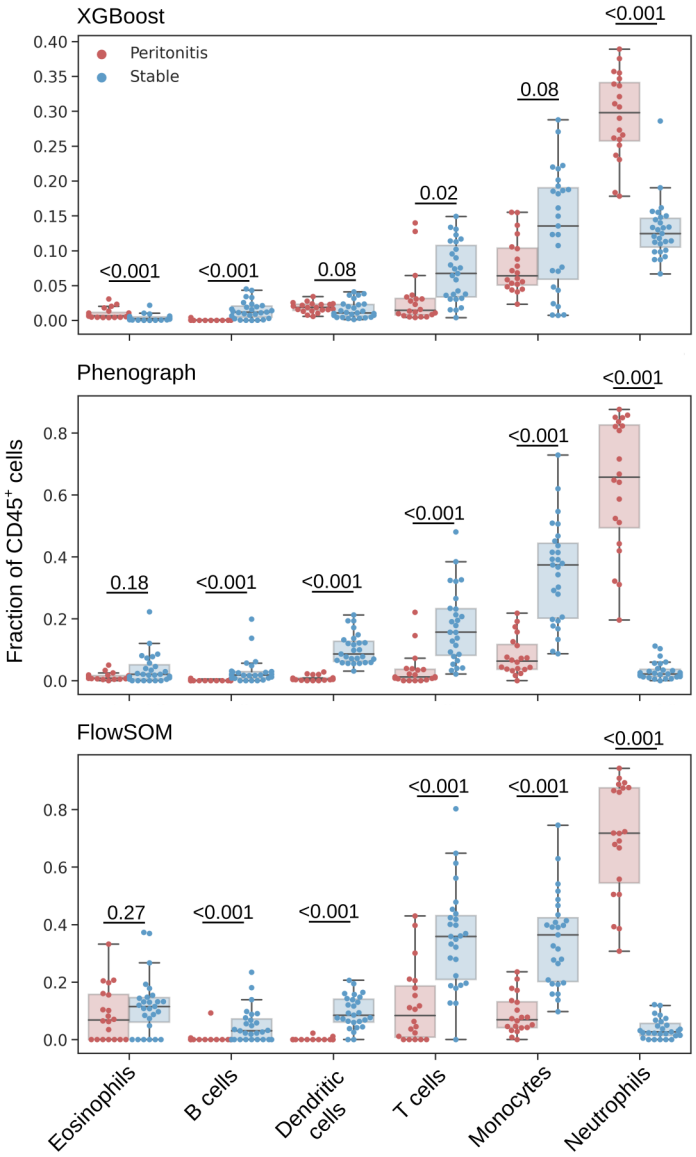

Figure 3.12: Leukocyte subsets in peritoneal effluent as a fraction of $CD45^+$ cells as identified by an XGBoost classifier (top), Phenograph clustering (centre) and FlowSOM clustering (bottom). Mann-Whitney U test were applied for comparisons between patients with acute peritonitis and stable controls, and p-values are reported after correction for multiple comparisons using Holm's method.

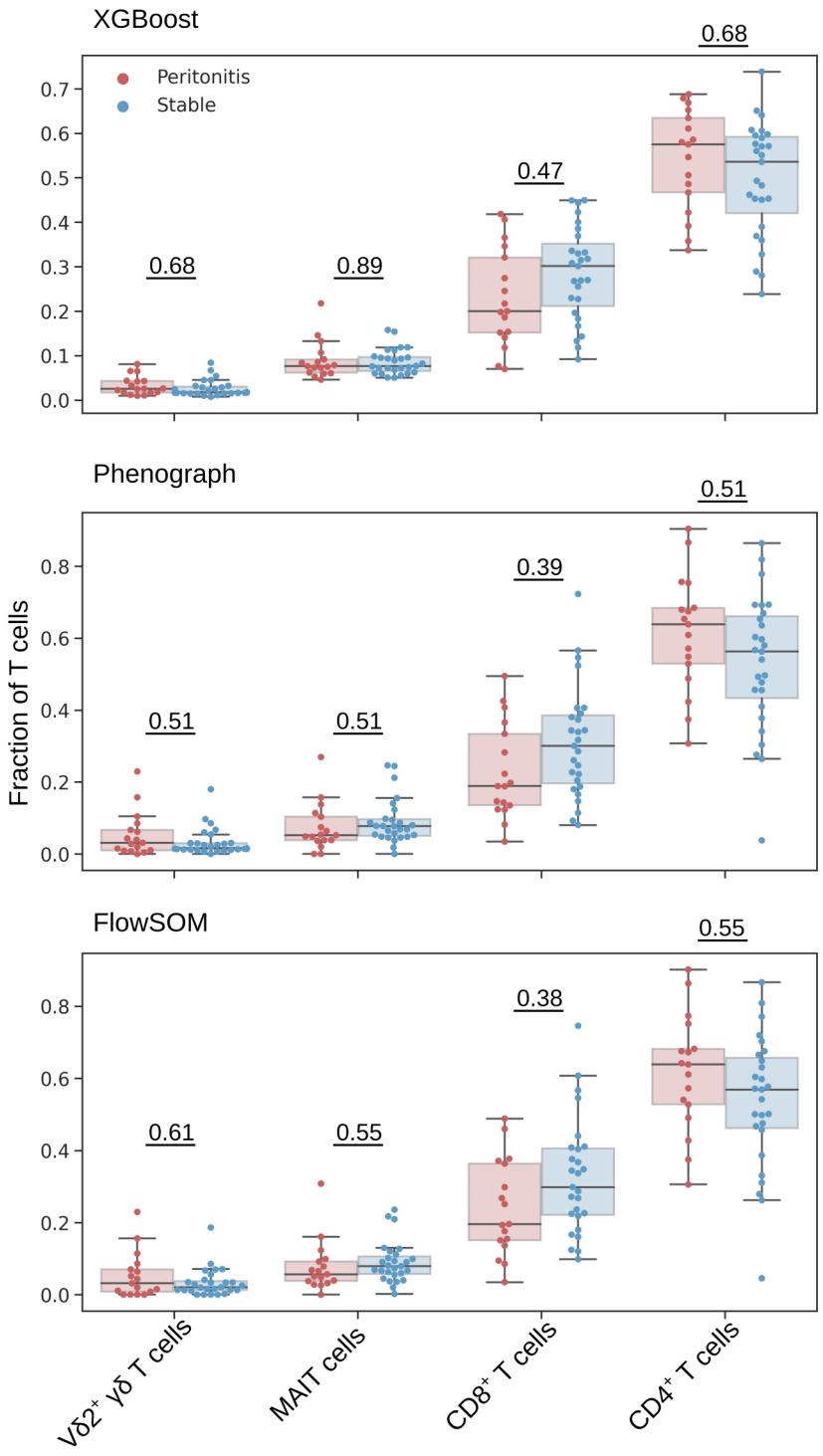

Figure 3.13: T cell subsets as a fraction of $CD3^+$ lymphocytes as identified by an XGBoost classifier (top), Phenograph clustering (centre) and FlowSOM clustering (bottom). Mann-Whitney U test were applied for comparisons between patients with acute peritonitis and stable controls, and p-values are reported after correction for multiple comparisons using Holm's method.

CytoPy's *feature_selection* module contains interpretable models for classification and regression problems, and its *DecisionTree* class can be used to demonstrate how the fraction of neutrophils alone can classify acute peritonitis (Figure 3.14D).

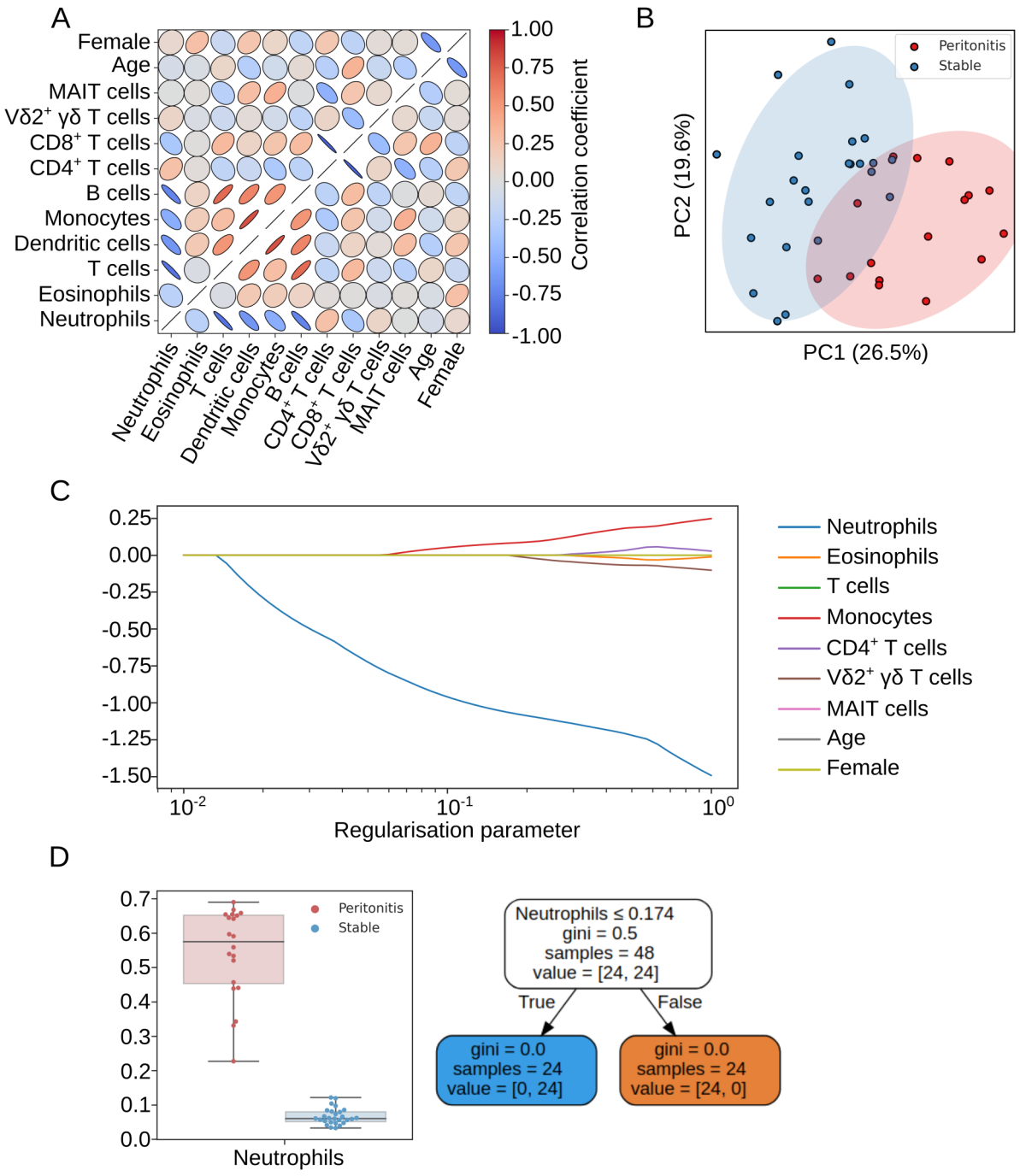

Figure 3.14: (A) Multicollinearity was addressed before generating linear models with redundant features removed prior to further analysis. (B) Principal component analysis shows that patients with acute peritonitis are discernible from stable controls. (C) L1 restricted modelling with a linear support vector machine reveals that neutrophils are the most predictive feature. (D) A simple cutoff applied to neutrophils is predictive of acute peritonitis in this cohort and is demonstrated by a shallow decision tree, where gini index is the chosen criterion for measuring the quality of split.

3.4 Discussion

CytoPy represents a framework for analysing cytometry data that facilities the application of machine learning algorithms whilst introducing robust data management and an iterative analytical environment. In this chapter, the ability of CytoPy to characterise cell populations is demonstrated, and the entire framework was validated by identifying a known immune phenotype that distinguishes patients with acute peritonitis. This dataset was chosen based on our group's extensive experience with this sample type for over a decade. Initially acquiring such samples on a four-colour BD FACSCalibur flow cytometer with two lasers and simple FSC/SSC settings [26], they later utilised an eight-colour BD FACSCanto with three lasers, and FSC/SSC area/height channels [175]. Now in the data presented in this chapter, taking advantage of a 16-colour BD LSR Fortessa with four lasers and FSC/SSC area, height, width and time [199], thus illustrating the technological advancements in the field but also the increasing complexity of the data acquired.

CytoPy exposes multiple techniques for classifying cell populations in cytometry data with a simplistic design and a low-code interface. Autonomous gates provide a familiar interface with cytometry data whilst reducing the labour cost of analysis. Nevertheless, they are biased by the investigator's expectations of the data and sometimes may not generalise well to new data. Following the work in this chapter, I recommend that autonomous gates be employed for pre-processing, generating a clean starting population for downstream analysis, and producing training data for supervised classifiers. Supervised classification offers a more efficient method for guided analysis but requires that batch effects be addressed up-front using methods such as the Harmony algorithm discussed in section 3.3.2. Similarly, unsupervised clustering also requires the attenuation of batch effects. In contrast to supervised classification, it is unbiased. It offers an exploratory analysis that can allude to discovering uncharacteristic cell populations or features that correlate with disease or experimental endpoints. This chapter demonstrates how clustering algorithms such as FlowSOM and Phenograph could not identify rare cell populations for a small fraction of our cohort, highlighting the importance of not relying on a single method when engineering features from cytometry data. A cornerstone of CytoPy's design is to expose multiple methodologies with minimal friction and provide consistent data structures to pool results. This strategy was employed for immune phenotyping peritoneal effluent and confirmed a striking increase

in total neutrophils at the site of infection and a parallel decrease in the proportion of monocytes/macrophages, dendritic cells and T cells, in agreement with previous findings [199, 175], thereby validating the utility of CytoPy. In the chapters that follow, additional methodologies will be introduced, capitalising on the design principles within CytoPy, to diversify analysis and reduce bias.

I have chosen to develop and maintain CytoPy in Python, a programming language with growing popularity in the bioscience domain. To date, Python has been lacking a framework for generalised cytometry data analysis offered by counterparts in R. CytoPy extends cytometry bioinformatics into the Python ecosystem by presenting an object-orientated infrastructure that is algorithm-agnostic and ready for deployment in the cloud. Compared to current frameworks in R [120, 140], CytoPy offers an end-to-end analytical interface that addresses common issues such as data cleaning and batch effect, and its data-centric design promotes iterative analysis for comparing multiple methodologies. Another popular solution for cytometry data analysis is CytoBank, which, whilst supporting many popular algorithms and an accessible graphical user interface, is a propriety product that could limit uptake. In contrast, CytoPy is open-source and, whilst offering popular algorithms, is also designed for expansion by the open-source community; new algorithms can be introduced with straightforward wrapper functions to match existing signatures and expected data types. CytoPy has also been designed with an open infrastructure that recognises the importance of maintainable code for software longevity. The code is accompanied by a documentation website (https://cytopy.readthedocs.io/en/latest/), Jupyter notebooks with examples on how to use the framework for analysis (https://github.com/burtonrj/CytoPyManuscript), and is deployable with Docker, a containerisation solution that can help manage complex dependencies and replicate analytical environments.

This chapter details CytoPy v2.0, which offers the most popular aspects of automated cytometry data analysis, with autonomous gating, high dimensional clustering and supervised learning, whilst also implementing Harmony [196, 197] for batch effect correction. In later chapters, CytoPy v3.0 is applied to the ILTIS study (see Materials & Methods section 2.1) to test the hypothesis that predictive phenotypes of innate immunity can be identified in acute severe sepsis patients.

As high-dimensional cytometry analysis continues to grow in popularity, there is increasing demand for an analytical framework that is friendly for those new to programming, provides a database that directly relates experimental metadata to single-cell data, and scales in a fashion that encourages collaboration and expansion. CytoPy meets all these criteria whilst remaining open-source and freely available on GitHub (https://github.com/burtonrj/CytoPy).

4 | Ensemble clustering of cytometry data

4.1 Introduction

Clustering is an unsupervised method for identifying structure in unlabelled data. In the context of cytometry, the objective here is to categorise events into groups of similar phenotypes. The previous chapter demonstrated clustering analysis as a successful alternative to manual gating, a finding corroborated by others in the field who regard clustering as an acceptable alternative to manual analysis [116, 132, 201]. Despite this increased uptake, the choice of algorithm appears to be driven either by availability in commercial software, the ease of use or is not discussed at all. Clustering algorithms differ in the assumptions made of data, performance tends to be highly data-specific, and results can vary widely depending on the chosen hyperparameters [203, 204, 202]. In this chapter, I propose ensemble clustering as an alternative solution.

Ensemble clustering (also called consensus clustering) aims to combine the partitions of multiple clustering algorithms run on the same data to identify a consensus informed by multiple 'views', thereby reducing the dependence on any individual algorithm. Unlike ensemble methods in supervised classification, ensemble clustering has many challenges: the number of clusters may differ amongst the base partitions, the optimal number of consensus clusters is often unknown, and it is necessary to solve the correspondence issue of matching clusters between individual partitions [203, 205]. A thorough review of the literature has revealed that ensemble clustering methods with specific applications to cytometry data analysis have yet to be proposed. Therefore, this chapter discusses methods from the single-cell RNA sequencing (scRNA-seq) literature and broader computer science and statistics literature in the context of cytometry data analysis.

Ensemble clustering methods can be grouped into co-association methods, feature-based methods, and methods using graph representations [203, 205, 206]. Co-association methods act on the pairwise similarity of clusters sourced from different algorithms. Consensus solutions can be derived from simple techniques such as agglomerative clustering of the binary co-association matrix ($N \times N$ matrix, where N is the number of events, e.g. the number of single cells) [204] or the Cluster-based Similarity Partitioning Algorithm (CSPA), that

forms partitions on the derived similarity graph using the METIS software [207]. Methods that act on co-association are burdened by space complexity and are therefore intractable for large data where such a matrix exceeds the available computer memory [203]. Feature-based methods offer an alternative by presenting the problem as a label-association matrix ($m \times n$ matrix, where m is the number of unique clusters). Consensus solutions can be formulated with iterative voting, finite mixture models, the pairwise agreement between clusters, or agglomerative clustering of this label-association matrix [205].

Another popular consensus clustering approach is using graph based methods, where a weighted graph of the clusters contributing to an ensemble is generated and then partitioned into k parts using a graph partitioning technique [203, 205]. Strehl and Ghosh [207] introduced the Hyper-Graph Partitioning Algorithm (HGPA) and the Meta-CLustering Algorithm (MCLA), both heuristics that represent the clustering ensemble as a hypergraph. Later the Hybrid Bipartite Graph Formulation (HBGF) algorithm was introduced as an alternative approach that models clusters and observations in the same graph, and consensus partitions are constructed from a subsequent bipartite graph [208]. The advantage of the graph methods is their heuristic approach that avoids the need for a co-association matrix, making them applicable to large data.

Ensemble clustering methods have been successfully adopted in the scRNA-seq literature. However, the methodologies adopted are in accordance with the size of the data generated by this technique and do not address the space complexity issues that arise from larger datasets, such as those encountered during cytometry data analysis. Sc-GPE [209] is an example of a solution deploying co-association to the problem of ensemble clustering. Here, a co-association matrix is weighted by contributing clustering methods' similarity (adjusted rand index). Unfortunately, the dependence on a co-association matrix makes this technique intractable for cytometry data. The same limitation applies to SC3 [210], another consensus approach for scRNA-seq employing CSPA for ensemble clustering. SAFE-clustering [211] avoids the need for generating a co-association matrix by applying graph-based methods instead, but the implementation only allows a limited number of contributing algorithms to the consensus and is exclusively designed for scRNA-seq.

In contrast to these advances in scRNA-seq data analysis, ensemble clustering methods have yet to be developed specifically for cytometry data analysis. Weber *et al.* [132] ex-

plored generic techniques from the graph-based ensemble clustering family for cytometry data analysis but failed to find additional benefits over existing algorithms. Aghaeepour *et al.* [176] demonstrated an ensemble methodology that utilised the label-association matrix and showed improved performance compared to individual algorithms. However, that publication did not disclose a readily available implementation of the methodology, making it difficult to reproduce their approach.

The work described in this chapter directly addresses the absence of techniques designed specifically for cytometry data analysis. Expanding on the work of Weber *et al.* [132], the graph ensemble clustering techniques with a strong track record in scRNA-seq data analysis and the capability of scaling to large data will be compared to popular clustering algorithms for cytometry data analysis. A novel ensemble clustering methodology based on geometric median clustering with weighted voting, named GeoWaVe, will also be introduced. Its performance will be compared to the graph ensemble clustering methods. Unlike previous ensemble clustering techniques, GeoWaVe is explicitly designed for cytometry data analysis and offers a computationally inexpensive heuristic approach, permitting the analysis of large data. The performance of GeoWaVe is presented on different sets of high-dimensional data generated using cytometry by time of flight mass spectrometry (CyTOF), spectral flow cytometry, and traditional multicolour flow cytometry.

4.2 Aims

1. Benchmark graph ensemble clustering methods in the application of cytometry data analysis, comparing results to the performance of state-of-the-art clustering algorithms used for cytometry data analysis.

2. Propose alternative methods for ensemble clustering in the context of cytometry data analysis and apply them to external and internal bench-mark data.

3. Demonstrate whether alternative methods for ensemble clustering of cytometry data outperform existing graph ensemble clustering methods.

4.3 Results

4.3.1 GeoWaVe: a novel heuristic ensemble clustering algorithm

Graph ensemble methods address computational complexity issues using a heuristic, deriving the consensus from graph representations of the label-association matrix rather than the unmanageable co-association matrix. Taking inspiration from this approach, I sought to develop a novel alternative heuristic ensemble clustering method that incorporates information about the original feature space: geometric median clustering with weighted voting (GeoWaVe), where the clusters generated by base clustering algorithms contributing to an ensemble are summarised by their geometric median. The geometric median (implemented with the $hdmedians$ package; [212]) was chosen over other measures of central tendency because it is robust to outliers, is not necessarily a point from the original data, can handle negative values, and is defined in any dimension.

A summary of the expression profile of all clusters contributing to the consensus is generated using the geometric median, which can subsequently be clustered into consensus clusters (Figure 4.1 heatmap); a consensus cluster is a collection of clusters of similar phenotypes. Since each cluster is treated as an individual contribution, differences in the number of clusters provided by each input algorithm are not consequential, meaning GeoWaVe can accept the outputs of any combination of clustering algorithms.

The clusters that contribute to a consensus are overlapping sets, given that each base clustering algorithm is exposed to the same data. Therefore, an event can be assigned to more than one consensus cluster. Assignment to multiple consensus clusters will occur more frequently for events on the boundary between clusters. Therefore, where an event is assigned to multiple consensus clusters, a score is calculated for each consensus cluster, and the event is assigned to the consensus with the maximum score.

The consensus cluster score is calculated as follows: given that a consensus cluster can be defined as a set of clusters $c \in C$, and a single cluster c is a finite set of n-dimensional vectors, the geometric median $\hat{u}$ of each cluster c can be calculated according to Equation 4.1 [212]:

$$\hat{u} = \underset{x \in \mathbb{R}}{argmin} \sum_{i=1}^{n} ||x - x_i||_2 \tag{4.1}$$

For each event t assigned to more than one consensus cluster C, the Manhatten distance between the event and the geometric median of each member cluster of C is computed. The sum of these distances normalized by the size of the consensus $|C|$ (i.e. the number of clusters within the consensus) gives a weighting factor p for the consensus cluster C relative to the event t (Equation 4.2):

$$p = \frac{\sum_{c \in C} ||t - \hat{u}(c)||_1}{|C|} \tag{4.2}$$

The consensus cluster score for C relative to an event t is then calculated as the size of the consensus $|C|$ divided by the weighting factor p (Equation 4.3):

$$score = \frac{|C|}{p} \tag{4.3}$$

Not all clusters are equally defined; some may be a poor fit for a given event. Therefore, the majority voting algorithm is weighted by the distance from an event to the centre of each cluster that contributes to a consensus. This method ensures that the consensus an event is assigned to is informed by the number of supporting algorithms and the quality of the clusters in that consensus.

The choice of clustering algorithm applied to the geometric medians of clusters is ambiguous in that any number of existing methods may be suitable to the task. The advantage of geometric medians as a heuristic is that the expression profile can be visualised easily as a heatmap (Figure 4.1), and different clustering methods can be applied and critiqued. The optional visualisation step allows the investigator to introduce prior knowledge, such as known phenotypes expected to occur in the data. The ambiguity of the clustering algorithm applied to the geometric median matrix allows for the use of methods such as the ConsensusClusterPlus method [213], choosing an optimal number of clusters from a given range. Therefore, an investigator can visualise the geometric medians and choose a range of clusters based on an intuition driven by the biological question.

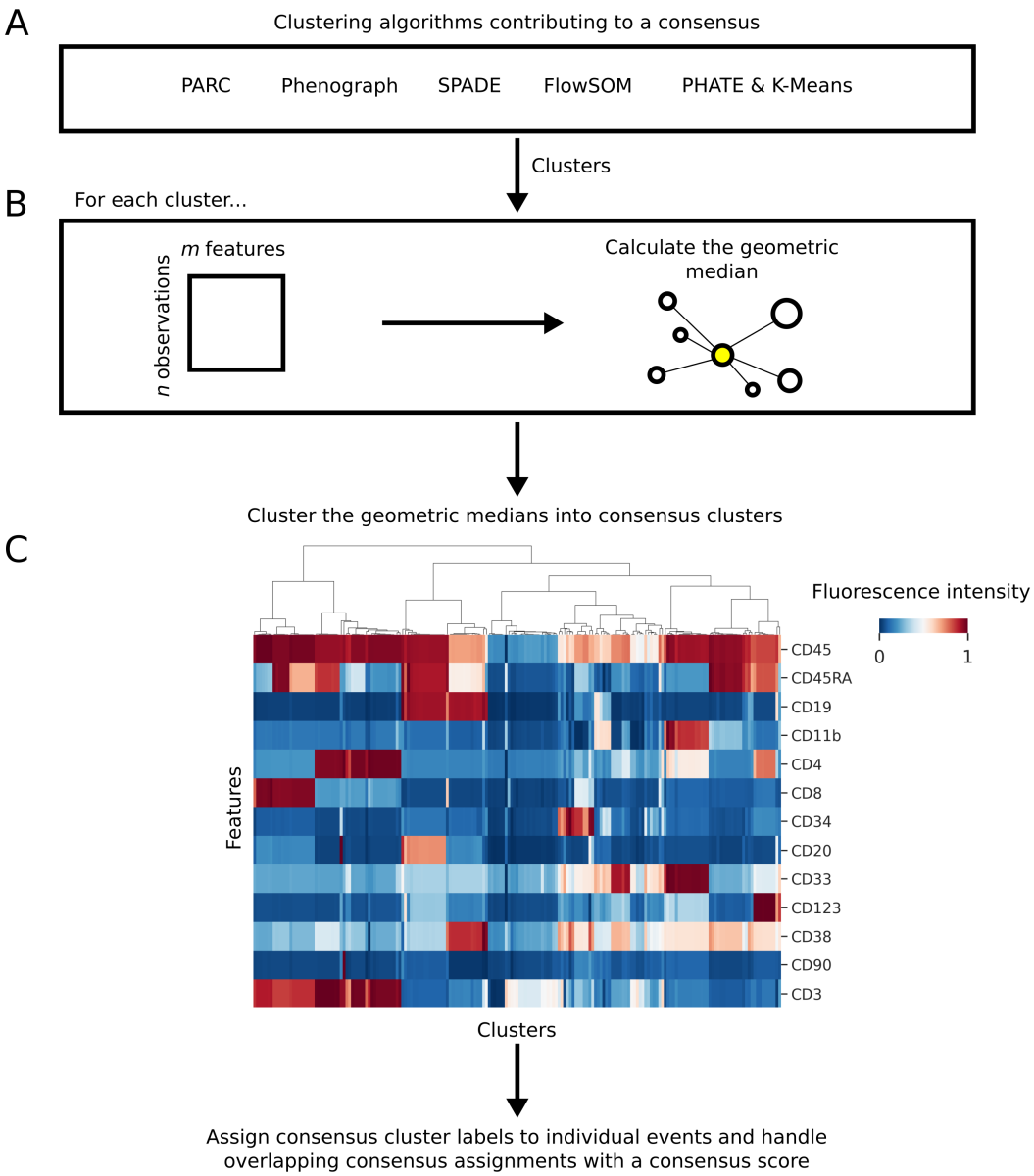

Figure 4.1: Schematic diagram of the GeoWaVe algorithm. (A) Clusters generated by multiple clustering algorithms are pooled, and (B) the geometric median for each cluster is calculated to create a matrix of c clusters. (C) This matrix of cluster geometric medians (example shown using the Levine-13 data introduced in section 4.3.2) is clustered into consensus clusters; groups of clusters within similar expression profiles. Consensus cluster labels are then assigned to individual events and overlapping consensus assignments handled with a score that accounts for the distance of the event to the members of each consensus cluster.

GeoWaVe is available as part of the $CytoCluster$ package, developed for Python version 3.8 or greater. The $CytoCluster$ package is available on the Python Package Index (PyPI). It offers popular cytometry clustering algorithms, graph ensemble clustering, and GeoWaVe ensemble clustering, as well as numerous utilities and plotting tools delivered through a simple object-orientated application programming interface.

4.3.2 Graph ensemble clustering methods fail to outperform individual clustering algorithms for cytometry data analysis

Diversity amongst the members of an ensemble can enhance results [205]. Ensemble clustering solutions should also take input from informative algorithms suited to the analytical task in question. I therefore chose algorithms that have reported good performance for cytometry data analysis, are well understood, have differing underlying methodologies, and are computationally efficient.

The following base clustering algorithms were considered individually, and their outputs served as input to ensemble clustering discussed in this chapter: FlowSOM [130], PHATE [138] with K-means, SPADE [111], Phenograph [131], and PARC [214]. The chosen algorithms have reported good performance for cytometry data analysis and are computationally efficient. Multiple input parameters were tried for each base clustering algorithm to give the best possible performance. The number of clusters generated for each method was determined as a property of the clustering method (as is the case with Phenograph and PARC), selected from a suitable range using the popular ConsensusClusterPlus method [213], or a suitable fixed value was chosen. The choice of the desired number of clusters, either as a range of values or a fixed value, would be driven by an existing biological understanding of the data in general use. For all benchmark data, a range of 5 to 30 was chosen to capture an extensive range of possible clusters. In the case of PHATE combined with K-Means, clustering was performed with k selected using ConsensusClusterPlus and then performed again with a fixed k of 20, a decision to increase the diversity of input clusters to the ensemble algorithms.

Base clustering algorithms and ensemble methods were tasked with clustering six datasets with available ground-truth labels: *Levine-13*, *Levine-32*, *Samusik*, *OMIP*, *Peritoneal Dialysis (PD)*, and *Sepsis* data (Table 4.1). The public CyTOF datasets *Levine-13*, *Levine-32*, and

Samusik, were obtained from open-source repositories [132] and arc-sinh transformed with a standard cofactor of 5. Doublets, debris and dead cells were removed, and ground-truth labels were taken from the original publications, with manual gating performed by the original authors [131, 134]. The *Levine-13* data described bone marrow cells from two healthy human donors and included 13 parameters (Figure 4.2), whilst the *Levine-32* data described bone marrow cells from a single healthy human donor but a higher resolution of 32 parameters (Figure 4.3). Some challenges presented by these data include overlapping monocyte subsets differentiated by CD11b expression in the *Levine-13* data and small subsets of B-cells differentiated on IgM and IgD expression in the *Levine-32* data. The *Samusik* data described bone marrow samples from 10 C57BL/6J mice and identified 24 populations using 39 parameters (Figure 4.4), including many small subsets with similar expression profiles.

In addition to the three CyTOF datasets, the OMIP-44 28-colour spectral flow cytometry dataset for identifying human dendritic cell compartments was included, available from an open-source repository [215]. Data were arc-sinh transformed with a standard cofactor of 150 and manually gated according to the gating strategy described by the original authors. Of the 28 parameters, 15 were retained to identify the manually gated subsets (Figure 4.5).

Two in-house datasets acquired with a 16-colour BD LSR Fortessa were included to examine the performance on traditional flow cytometry data: *Sepsis* and *Peritoneal Dialysis (PD)*. The *Sepsis* data (see Materials & Methods 2.1 for details) was included for the identification of conventional and non-conventional T cell subsets from peripheral blood mononuclear cells (PBMCs) from patients diagnosed with sepsis. Data were arc-sinh transformed (standard cofactor of 150) and batch effect corrected using the Harmony algorithm as discussed in Chapter 3. Each sample was manually gated for single live $CD4^+$ and $CD8^+$ T cells, $V\delta2^+$ $\gamma\delta$ T cells and $CD161^+$ $V\alpha7.2^+$ mucosal-associated invariant T (MAIT) cells (Figure 4.6A). The identified lymphocyte populations then served as ground truth for comparing results from the clustering algorithms. The *PD* data (see Materials & Methods 2.2 for details) were derived from a single adult receiving peritoneal dialysis with no previous infections for at least three months prior to sampling. Data were arc-sinh transformed (standard cofactor of 150), and debris and dead cells were removed prior to analysis. Leukocyte populations in peritoneal effluent were identified as live $CD45^+$ immune cells and manually gated for $CD3^+$ T cells, $CD19^+$ B cells, $CD15^-$ $CD14^+$ monocytes/macrophages, $CD15^+$ neutrophils, $CD15^-$ $CD14^{+/-}$ $CD1c^+$ DCs, and $CD15^-$ SIGLEC-8^+ eosinophils (Figure 4.6B). The

Dataset Name	Number of observations	Number of parameters (No. utilised for clustering)	Technology	Availability
Levine-13	167,044	13	Mass Cytometry (CyTOF)	Flow Repository No. FR-FCM-ZZPH [132]
Levine-32	265,627	32	Mass Cytometry (CyTOF)	Flow Repository No. FR-FCM-ZZPH [132]
Samusik	841,644*	40	Mass Cytometry (CyTOF)	Flow Repository No. FR-FCM-ZZPH [132]
OMIP	2,805,957*	28 (15)	Spectral Flow Cytometry	Flow Repository No. FR-FCM-Z32U [215]
Sepsis	362,361*	10 (6)	Flow Cytometry	Zenodo; doi: 10.5281/zenodo.7134723 [216]
Peritoneal Dialysis (PD)	6,333,084*	11 (9)	Flow Cytometry	Zenodo; doi: 10.5281/zenodo.7134723 [216]

Table 4.1: Description of the benchmark data employed for assessment of ensemble clustering algorithms. * Data were down-sampled before analysis.

identified populations then served as ground truth for comparing results from the clustering algorithms. Both the *Sepsis* and *PD* data offered a unique challenge because of relatively small and ambiguous populations being present amongst a backdrop of more predominant cell types (Figure 4.7). The *OMIP*, *Levine-13*, and *Levine-32* had overlapping populations and the *Samusik* data had a branching topology. The six chosen datasets offered diverse challenges and featured representations from various source technologies.

The output of the base clustering algorithms was used to generate a label-association matrix (m clusters $\times n$ observations) which served as input for three graph ensemble clustering algorithms that have reported successful application in the scRNA-seq literature [211]: HGPA, MCLA, HBGF. The graph ensemble algorithms were implemented using the $ClusterEnsembles$ Python package [217].

For graph ensembles, a required hyperparameter is the number of final partitions in the consensus solution. This problem was addressed in the base clustering algorithms by searching a range of possible clusters and using the ConsensusClusterPlus. This approach requires sub-sampling the feature space and computing the co-association matrix for each value of k (the number of clusters). The cumulative distribution function (CDF) for each co-association matrix is generated, and the optimal k is chosen where the CDF is maximum. Although applicable to methods such as FlowSOM and SPADE that use a heuristic or down-sampled feature space, such an approach is intractable for the graph-based consensus clustering techniques that construct graph representations of a $M \times N$ label-association matrix. Therefore, the optimal number of consensus partitions was chosen using internal metrics (metrics that use internal information from the clustering process to evaluate the quality of a clustering $e.g.$ the variation within clusters or the degree of overlap between clusters). Ensemble clustering was repeated over a range of k, chosen as the smallest and largest number of clusters amongst base clustering algorithms. Four internal metrics, implemented in Scikit-Learn [177], were chosen for their ease of interpretation:

1. **Calinski-Harabasz score** is the ratio of the sum of between-cluster dispersion and the sum of within-cluster dispersion. Higher values correspond to better defined clusters [177].

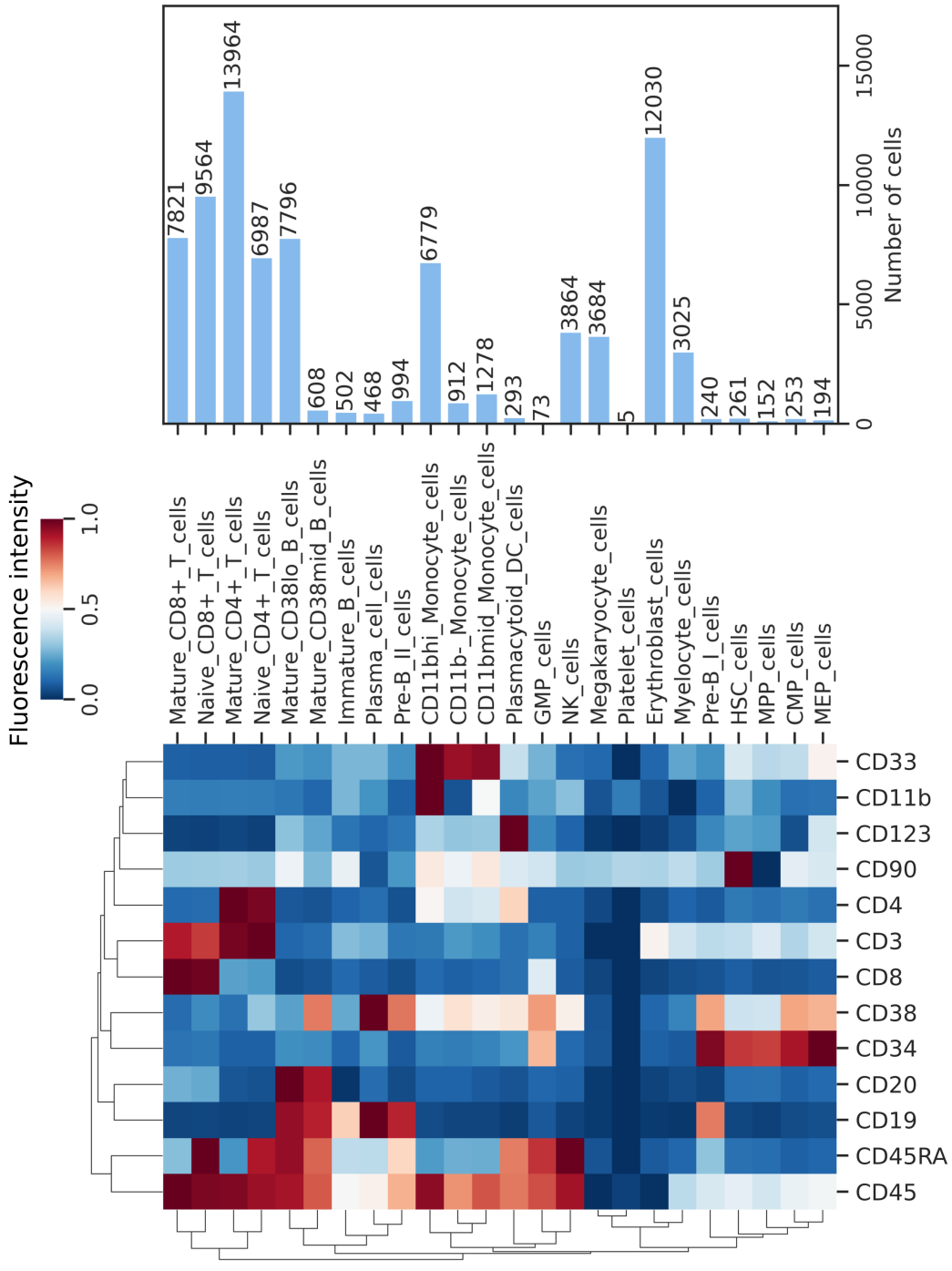

Figure 4.2: Expression profile of the 13 parameter Levine CyTOF data (*Levine-13*) and the total number of observations for each ground-truth population. Heatmap shows expression intensity of cell surface markers and normalised to a range of 0 and 1.

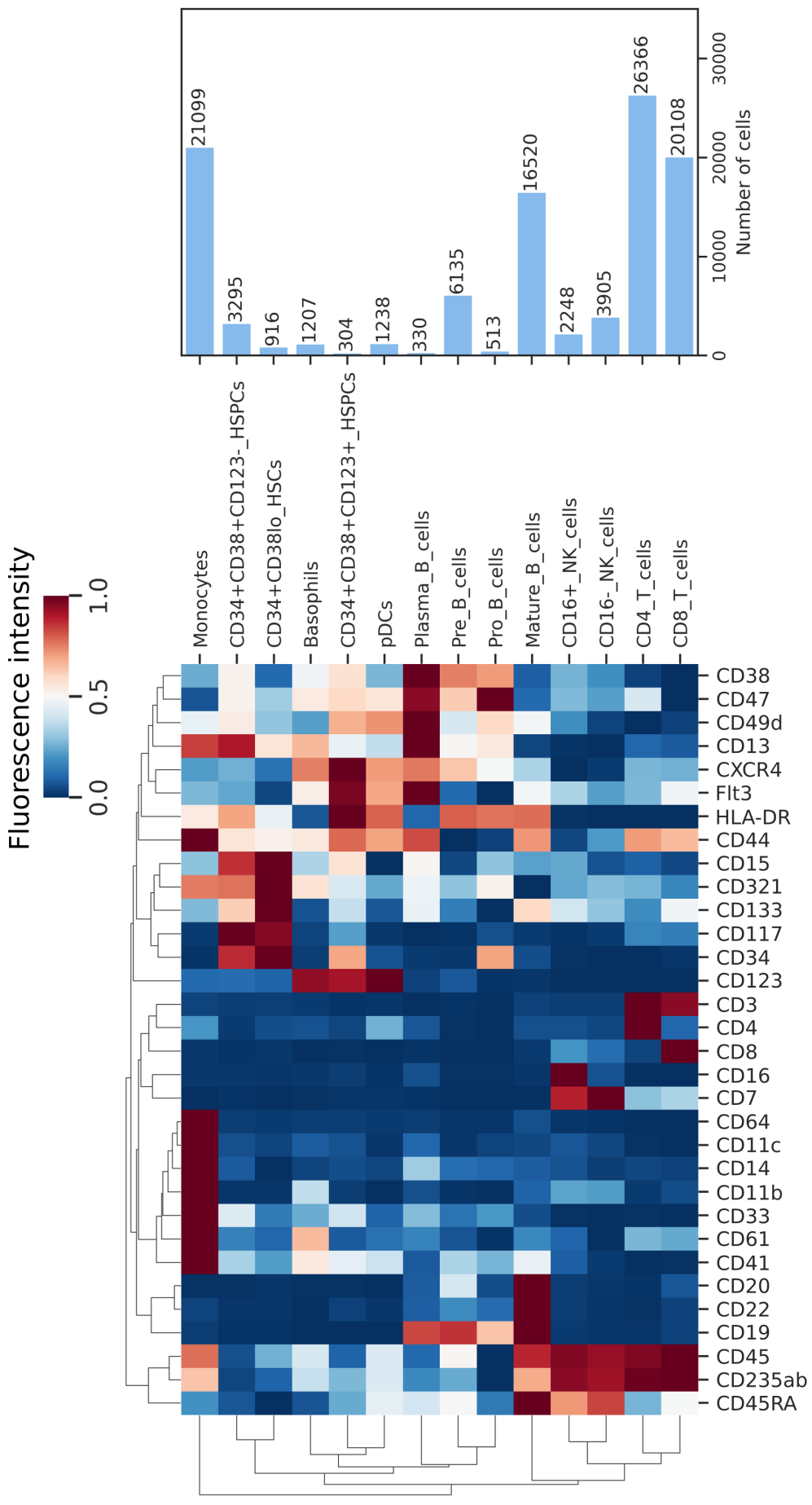

Figure 4.3: Expression profile of the 32 parameter Levine CyTOF data (*Levine-32*) and the total number of observations for each ground-truth population. Heatmap shows expression intensity of cell surface markers and normalised to a range of 0 and 1.

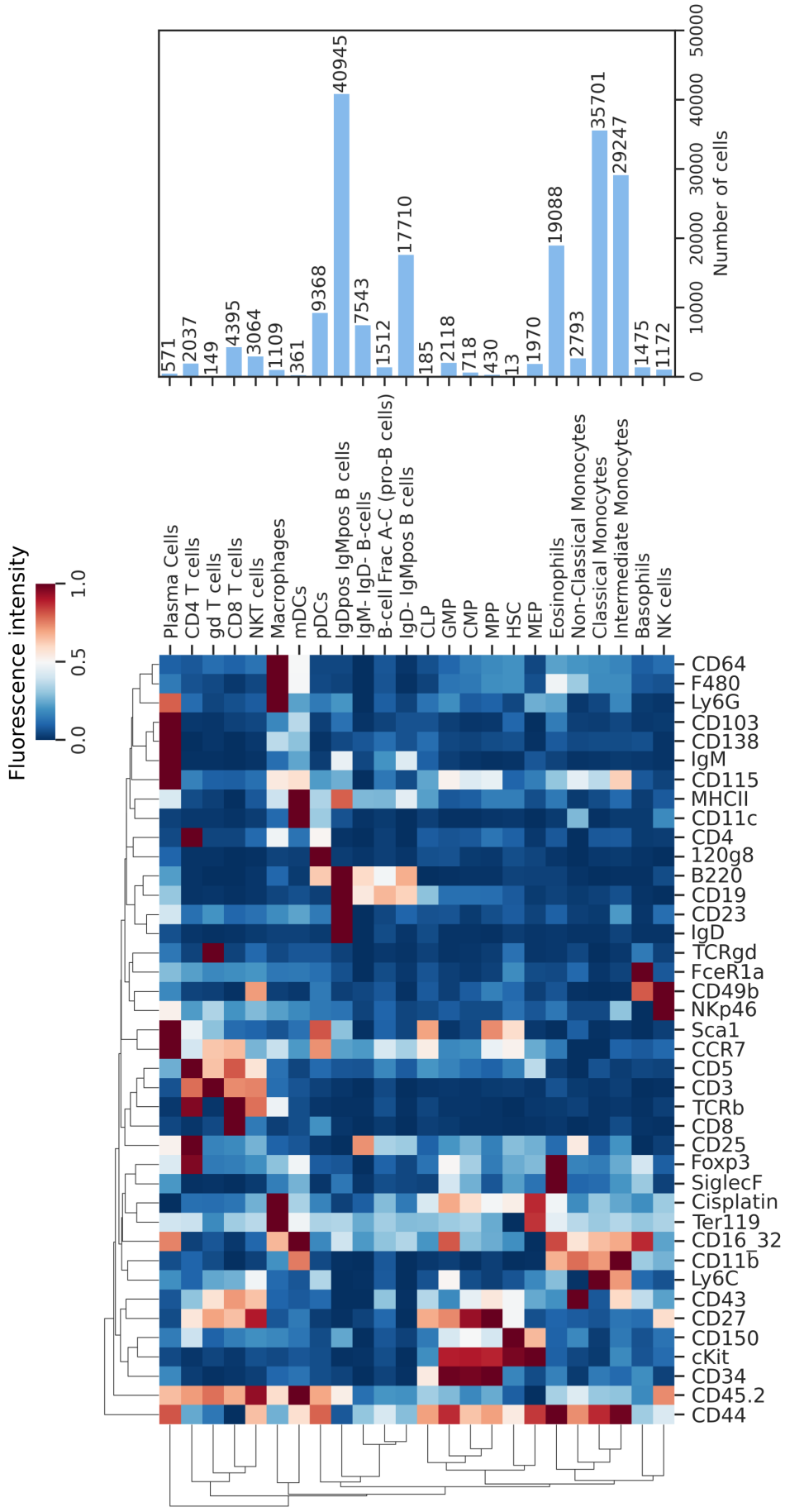

Figure 4.4: Expression profile of the 39 parameter Samusik CyTOF data (*Samusik*) and the total number of observations for each ground-truth population. Heatmap shows expression intensity of cell surface markers and normalised to a range of 0 and 1.

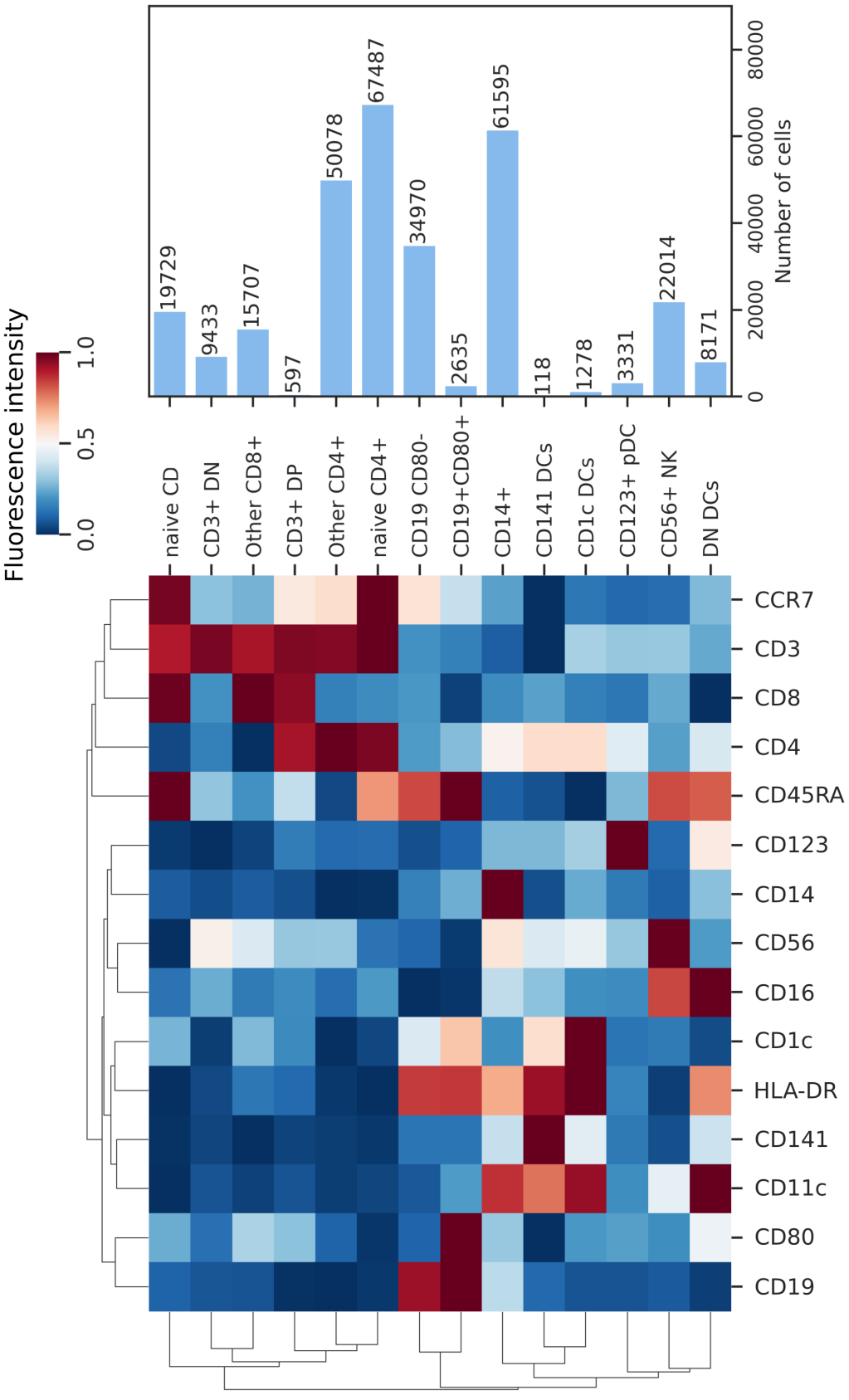

Figure 4.5: Expression profile of the 28 parameter OMIP-44 spectral flow cytometry data (*OMIP*) and the total number of observations for each ground-truth population. Heatmap shows expression intensity of cell surface markers and normalised to a range of 0 and 1.

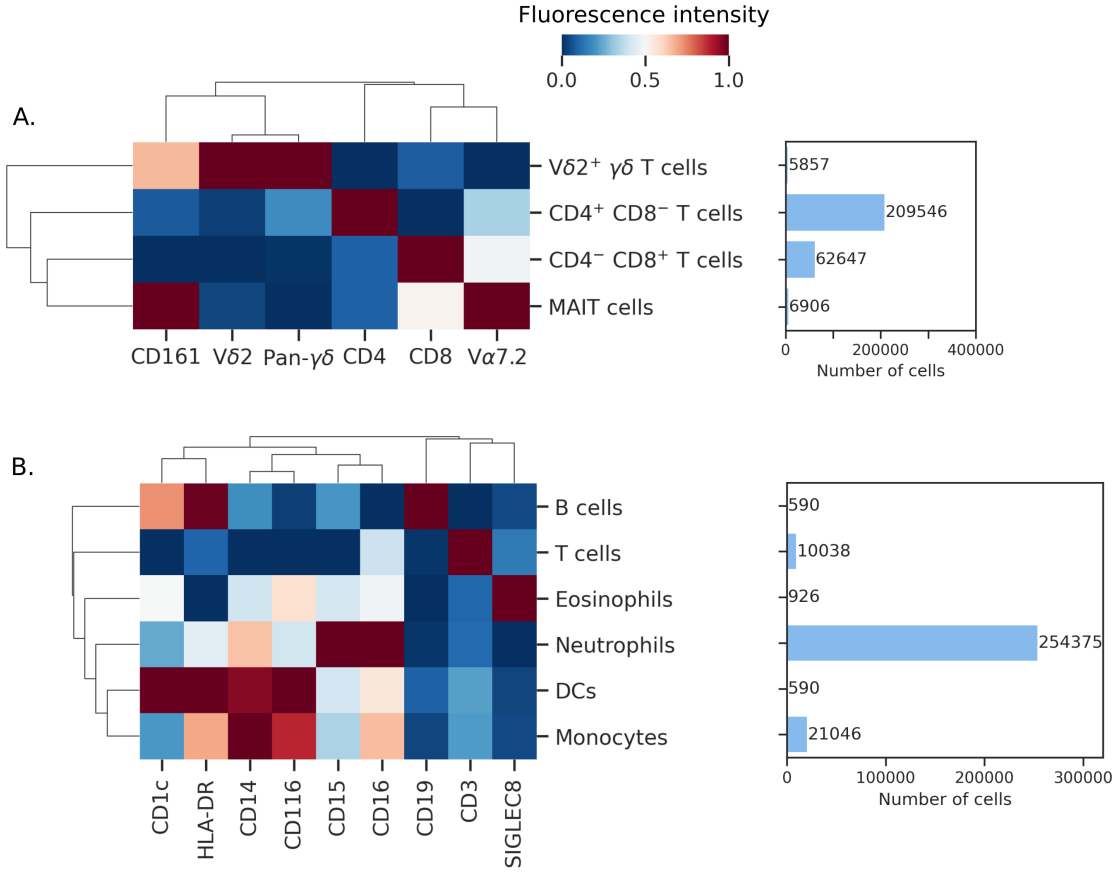

Figure 4.6: Expression profile of conventional flow cytometry data *Sepsis* (A) and *Peritoneal Dialysis* (B) and the total number of observations for each ground-truth population. Heatmap shows expression intensity of cell surface markers and normalised to a range of 0 and 1.

2. **Davies-Bouldin index** compares each cluster to every other cluster measuring similarity as the ratio of within-cluster distances to between-cluster distances. Lower values indicate better define clusters [177].

3. **Distortion score** provides a measure of the compactness of clusters, measured as the average squared distance between each point in a cluster and the cluster centroid [177, 218].

4. **Silhouette coefficient** is measured for each observation and calculated as the distance between the observation and nearest cluster the observation is not a member of (a), minus the mean intra-sample distance (b; the distance between the observation and all other observations in the same cluster), divided by the maximum of a and b. Values are reported between 1 and –1, with values near to 0 indicating overlapping clusters,

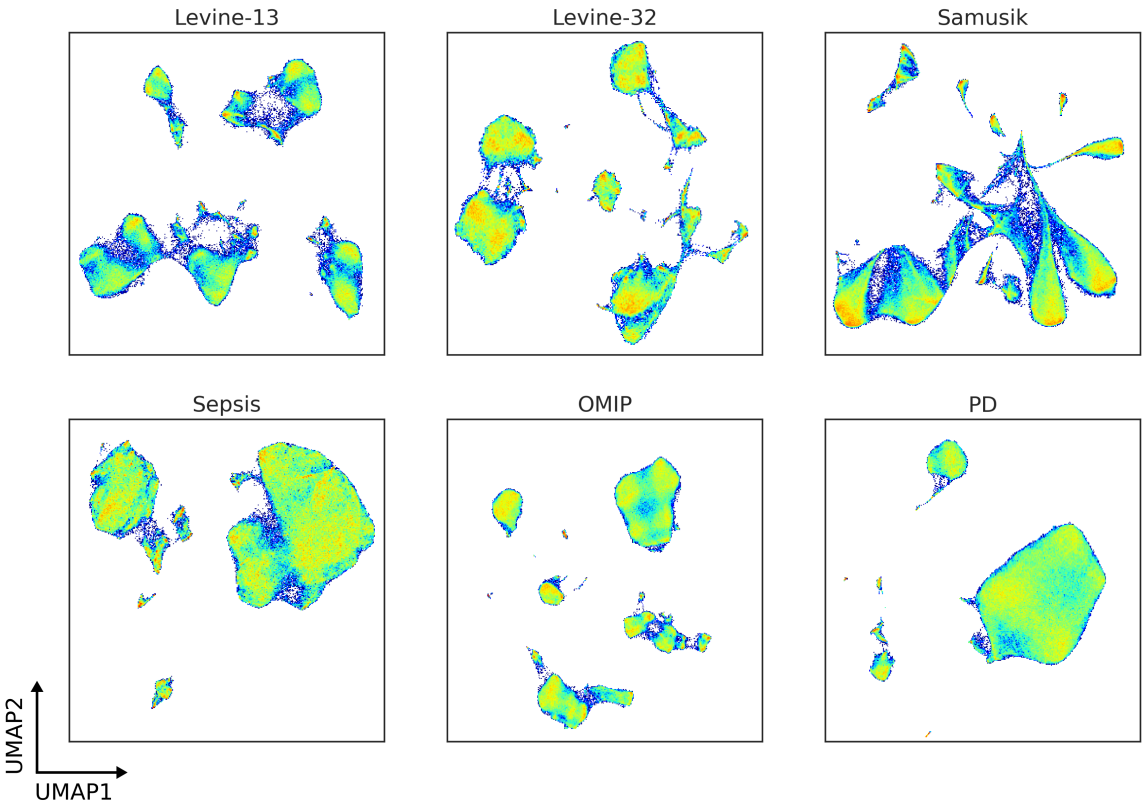

Figure 4.7: UMAP density plots show the topology of the six benchmark datasets for the evaluation of ensemble clustering. Colour intensity corresponds to the density of observations in a region of events.

and negative values generally indicating that an observation was assigned to the wrong cluster [177, 218].

An example is given in Figure 4.8 for HGPA clustering of the Levine-13 data. Internal metrics were measured for 1000 events with 100 re-samples, and the distribution was plotted for each k. The optimal k (chosen as k=6 in the example shown) is visually determined as the value where Calinski-Harabasz score and Silhouette coefficient are maximised, whilst Davis-Bouldin index and distortion scores are minimised.

The performance of the base clustering algorithms and the ensemble methods was evaluated using the following external metrics (metrics that compare cluster results to ground-truth labels) implemented in the Scikit-Learn library [177]:

1. **Adjusted Rand Index (ARI)** provides a measure of similarity between clusters and ground-truth labels by considering all pairs of observations. Pairs assigned to the same or different clusters in the predicted and ground-truth populations are counted

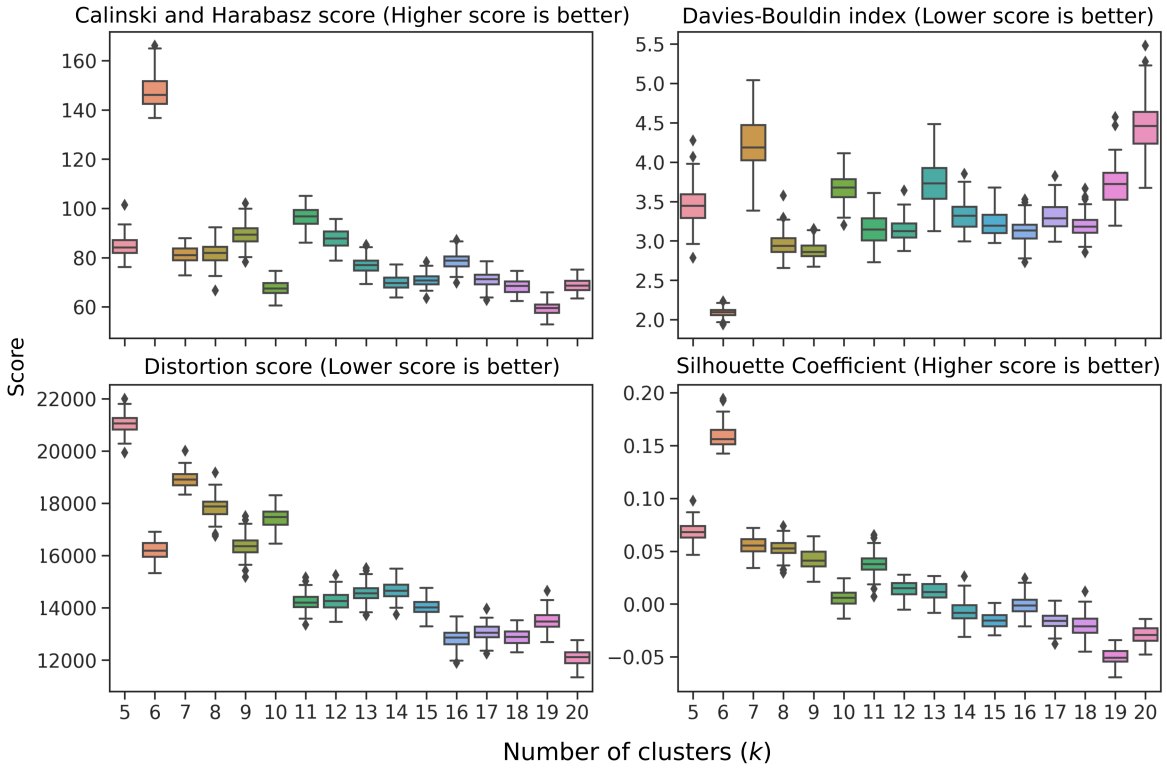

Figure 4.8: Internal metrics for a range of final consensus clusters (k) as generated by HGPA clustering of Levine-13 data.

and contrasted to mismatched pairs. The Rand Index can be described as a measure of the percentage of correct classifications by the clustering algorithm and is adjusted for chance by estimating the expected rand index using a permutation model and then normalising by this expectation. The ARI scores clustering results between -1 and 1, where random label assignment would be negative or close to zero, but perfect clustering would have an ARI close to 1[177, 218].

2. **Adjusted Mutual Information (AMI)**. Mutual Information is derived from information theory and aims to quantify the amount of shared information between the predicted clusters and the ground-truth populations. Mutual Information is not adjusted for chance and will tend to increase as the number of clusters increases, regardless of the quality of additional clusters. To remedy this, AMI first calculates the expected value for mutual information and adjusts for chance similarly to the Adjusted Rand Index. The AMI scores clustering results between 0 and 1, where random label assignments would give a score of 0, but perfect clustering would have a score of 1 [177, 218].

3. **Fowlkes-Mallows Index (FMI)**, calculated as the square root of the product of pairwise precision (for each ground-truth label, the number of true positives over the number of true positives plus the number of false positives) and pairwise recall (for each ground-truth label, the number of true positives over the number of true positives plus the number of false negatives). FMI is, therefore the geometric mean of pairwise precision and recall, and scores clustering results between 0 and 1, with higher values indicating similarity between cluster results and ground-truth labels [177].

Figure 4.9 shows the ARI performance of the base clustering algorithms (the algorithms used to contribute to ensemble clustering) and the graph ensemble clustering algorithms. MCLA offered greater performance than the other graph ensemble methods in most cases, a finding corroborated by FMI and AMI (Figure 4.10). Although in the *Levine-13* and *Levine-32* data graph ensemble methods improved on the performance of algorithms such as SPADE or FlowSOM, in only one of the six datasets (*OMIP*) did any graph ensemble outperform the base clustering algorithms. This evidence makes it difficult to justify using graph ensemble methods for cytometry data.

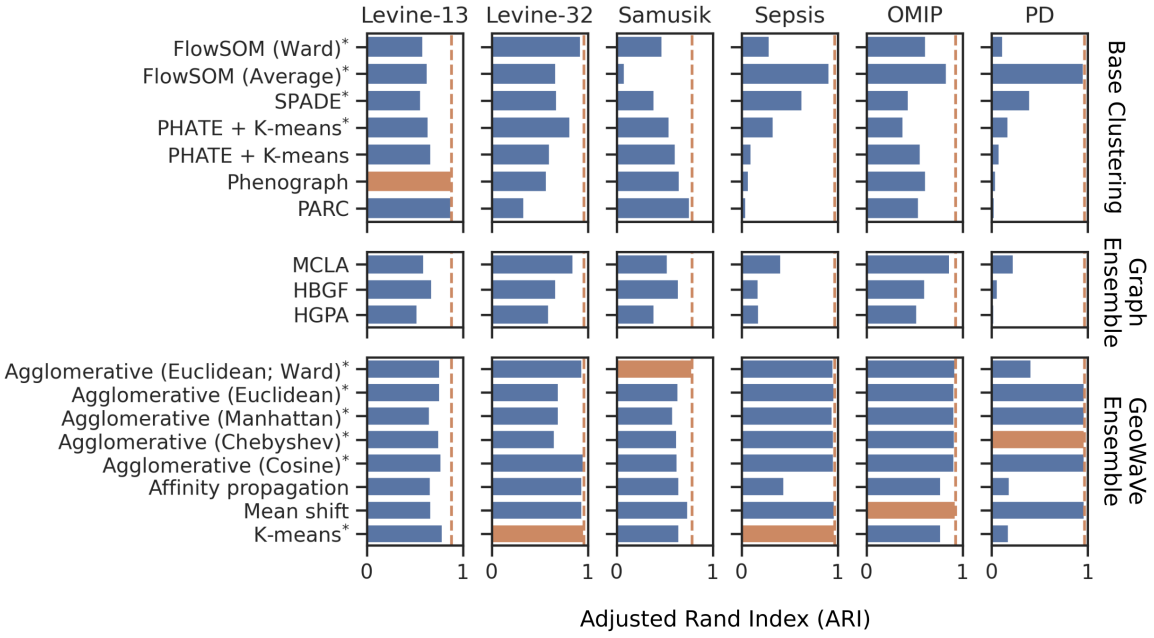

Figure 4.9: Adjusted rand index (ARI) for base clustering algorithms (left), graph ensemble methods (middle) and GeoWaVe ensemble (right) for the six benchmark datasets. The best ARI score for each dataset is shown as a dotted orange line, and the best performing method for those data is coloured in orange. * the optimal number of clusters, k, was chosen using the ConsensusClusterPlus method [213].

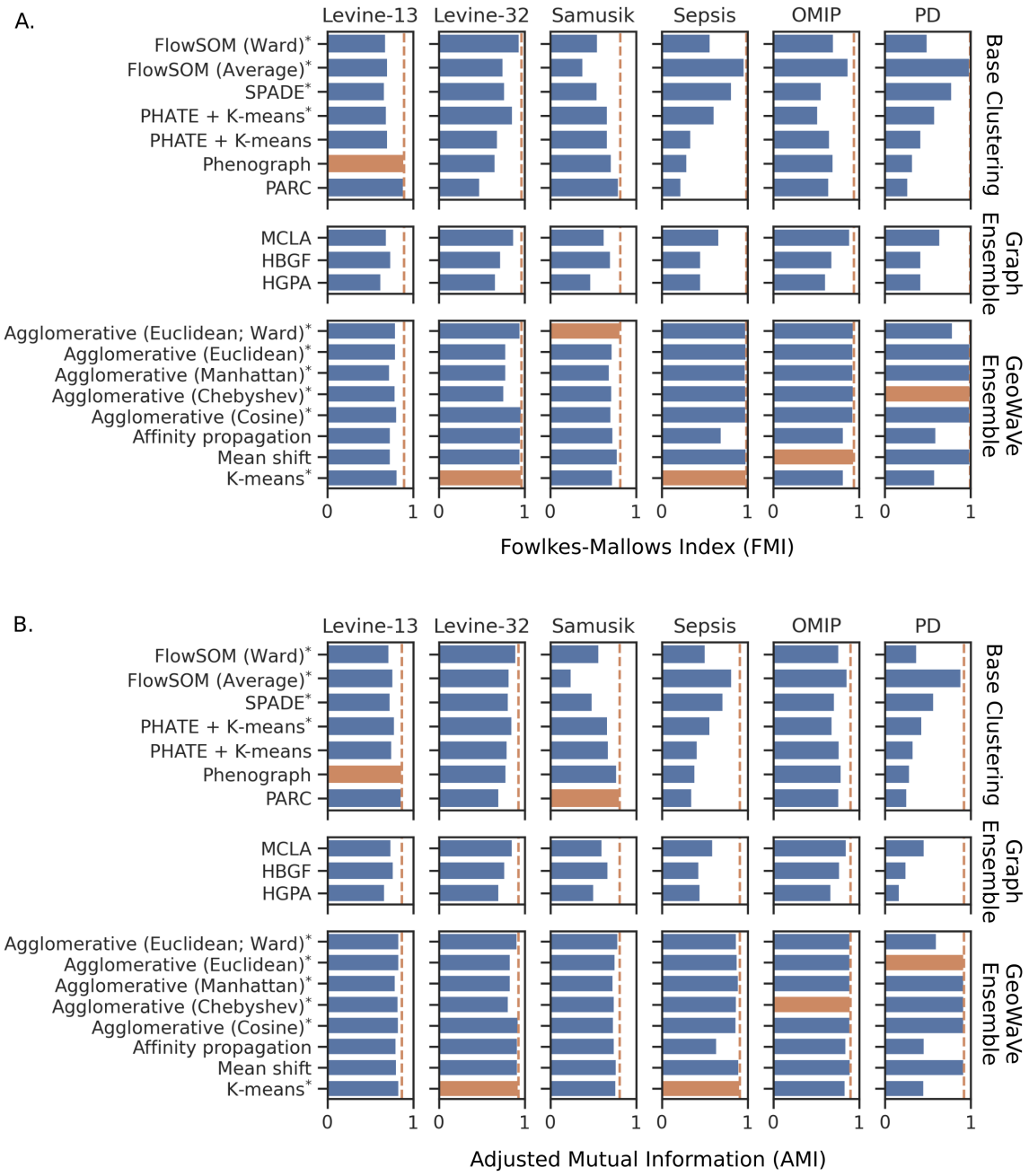

Figure 4.10: Fowlkes-Mallows index (FMI) (A) and Adjusted Mutual Information (AMI) (B) for base clustering algorithms (left), graph ensemble methods (middle) and GeoWaVe ensemble (right) for the six benchmark datasets. The best score for each dataset is shown as a dotted orange line, and the best performing method for those data is coloured in orange. * the optimal number of clusters, k, was chosen using the ConsensusClusterPlus method [213].

It was questioned whether the performance of graph ensemble methods was a direct result of the method employed for selecting k, the number of final consensus clusters ($i.e.$ the use of internal metrics as shown in Figure 4.8). Therefore, the performance of a graph-based clustering algorithm was examined across different values of k using external evaluation metrics. HBGF was chosen for this experiment because it had the best runtime of the three graph ensemble methods. The performance of HBGF for the four datasets is shown in Figure 4.11 across a range of values of k. Performance was optimum for low values of k despite the number of ground-truth populations being much larger for data such as *Levine-13* and *Samusik*. Therefore, the choice of k was assumed not to be a factor in the poor performance of graph ensemble methods in this case.

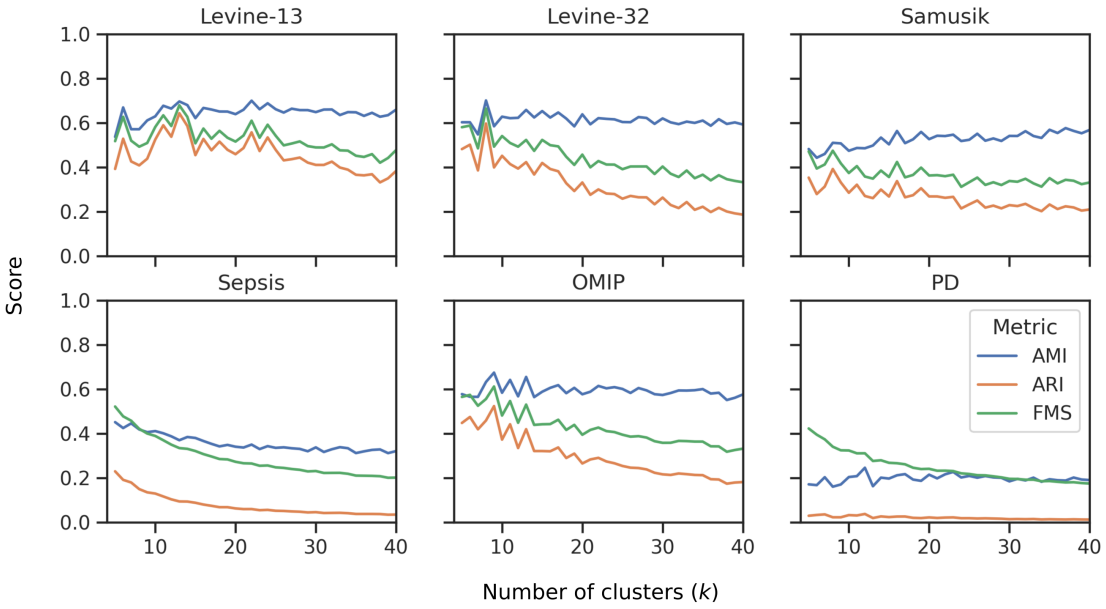

Figure 4.11: Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), and Fowlkes-Mallows Index (FMI) when the number of consensus clusters (k) is varied for HBGF ensemble clustering.

4.3.3 GeoWaVe outperforms graph ensemble methods and improves upon the performance of base clustering algorithms.

The GeoWaVe algorithm was compared to the performance of base clustering algorithms and the graph ensemble clustering algorithms. As discussed in section 4.3.1, GeoWaVe is flexible when clustering the geometric medians of the input clusters and allows the user to choose from many clustering algorithms for this task. Multiple algorithms for clustering the

geometric medians were tried during the validation of GeoWaVe. Affinity propagation and mean-shift were compared because of their ability to select the optimal number of clusters from the characteristics of the data. K-means and agglomerative hierarchical clustering were also tested, with the optimal number of clusters chosen from a range of clusters using the ConsensusClusterPlus method [213]. Agglomerative hierarchical clustering offers an additional advantage to the end user because consensus clusters can be easily visualised as a dendrogram and clustered heatmap, allowing the investigator to choose an appropriate range for the number of consensus clusters driven by their understanding of the underlying biology. For agglomerative hierarchical clustering, various linkage methods and distance metrics were tried.

GeoWaVe performance was compared to base clustering algorithms and graph ensemble methods using the external evaluation metrics discussed in the previous section (ARI, AMI, and FMI). GeoWaVe outperformed all other methods in five of the six datasets when comparing ARI (Figure 4.9) and FMI (Figure 4.10). GeoWaVe also outperformed graph ensemble methods when comparing ARI, FMI and AMI but failed to outperform base clustering methods in terms of AMI in the *Levine-13* and *Samusik* data.

The effect of the choice of clustering algorithm applied in GeoWaVe was data specific. For the *Levine-13*, *Samusik*, and *OMIP* data, the choice of the algorithm was negligible, whereas hierarchical clustering for the Levine-32 data was sensitive to the choice of distance metric. Affinity propagation gave an inferior performance for Sepsis data. Likewise, affinity propagation, along with K-means and Ward clustering, resulted in poor performance for *PD* data.

4.3.4 GeoWaVe outperforms graph ensemble methods for the detection of under-represented populations.

External evaluation metrics used in the prior section offer performance criteria independent of the labels, i.e. they do not require a like-to-like matching of cluster and ground-truth labels. Instead, measures of similarity between the cluster labels and ground-truth labels were used. Aghaeepour *et al.* [116], Samusik *et al.* [134] and Weber & Robinson [132] alternatively framed such problems in the context of a classification task: a one-to-one mapping of ground-truth labels to clusters was achieved using the Hungarian algorithm such that the sum

of F1 scores across ground-truth labels is maximised, and the precision (positive predictive value), recall (sensitivity) and F1 score (harmonic mean of precision and recall) for each ground-truth label are reported.

This procedure was repeated for the clustering algorithms benchmarked in previous sections and the ensemble clustering solutions. Figure 4.12 shows the average F1 score for the base clustering algorithms, graph ensemble methods and GeoWaVe, along with the standard deviation (error bars) showing the variation in F1 score between populations. The F1 score, precision, and recall are reported in Figure 4.13. GeoWaVe continued to outperform graph ensemble methods across the six benchmark datasets but failed to match the F1 score obtained by methods such as PHATE combined with K-means in the *Levine-13* data and Phenograph in the *Samusik* data. While MCLA graph ensemble clustering was more comparable to GeoWaVe in the Sepsis data when observing F1 score, GeoWaVe clustering still outperformed MCLA in terms of precision, recall and F1 score. GeoWaVe clustering offered optimal average F1 scores for *Levine-13*, *Sepsis*, *OMIP*, and *PD* data and outperformed graph ensemble methods across all datasets.

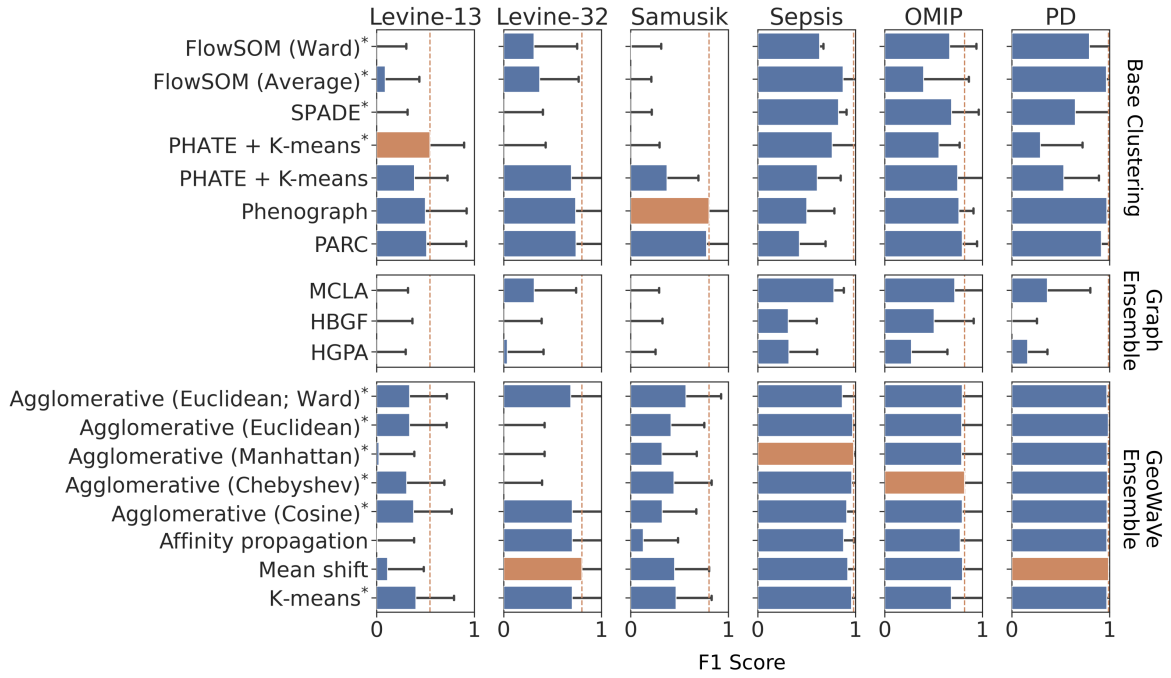

Figure 4.12: Performance of base clustering algorithms, graph ensembles and GeoWaVe ensembles, after matching cluster labels to ground-truth labels using the Hungarian linear assignment algorithm (as described by [132]) and maximising the sum of F1 scores across ground-truth label and cluster label pairings. Median F1 scores are reported with error bars showing the standard deviation either side of the average. * the optimal number of clusters, k, was chosen using the ConsensusClusterPlus method [213]

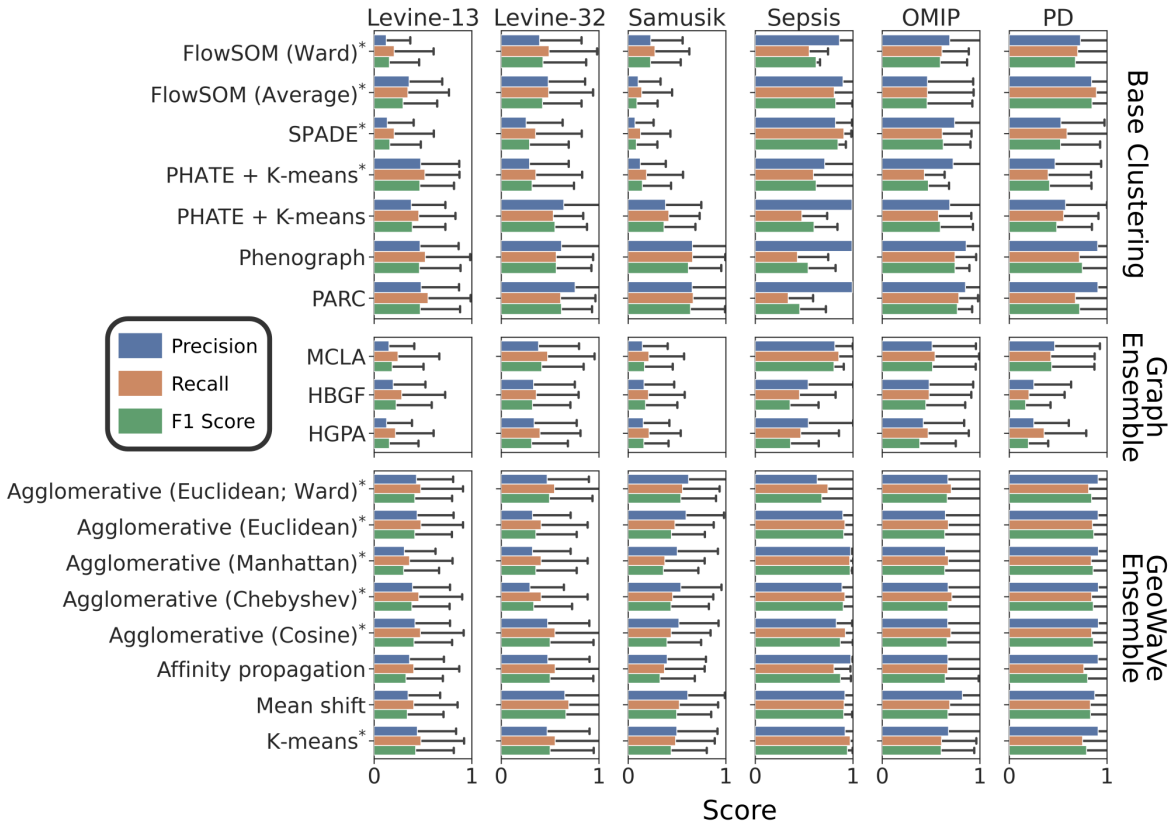

Figure 4.13: F1 score, precision, and recall of base clustering algorithms, graph ensembles and GeoWaVe ensembles, after matching cluster labels to ground-truth labels using the Hungarian linear assignment algorithm (as described by [132]) and maximising the sum of F1 scores across ground-truth label and cluster label pairings. Mean scores are reported with error bars showing the standard deviation either side of the average. * the optimal number of clusters, k, was chosen using the ConsensusClusterPlus method [213]

An advantage to matching clusters to ground-truth populations using the Hungarian algorithm was the ability to compare the performance at the population level. The F1 score for ground-truth populations for the top performing algorithm from the base-clustering, graph ensemble clustering, and GeoWaVe ensemble clustering are shown as heatmaps in Figure 4.14 and 4.15. Each row includes a measure of the population size as an additional heatmap on the y-axis. The heatmaps demonstrate the superior performance of GeoWaVe compared to graph ensemble methods for the identification of under-represented populations such as plasmacytoid dendritic cells (pDCs) in the *Levine-13* dataset, plasma cells, basophils and pro-B cells in *Levine-32*, pDCs in *OMIP* data (Figure 4.14), B cells and dendritic cells (DCs) in the *PD* data, and MAIT cells in *Sepsis* data (Figure 4.15).

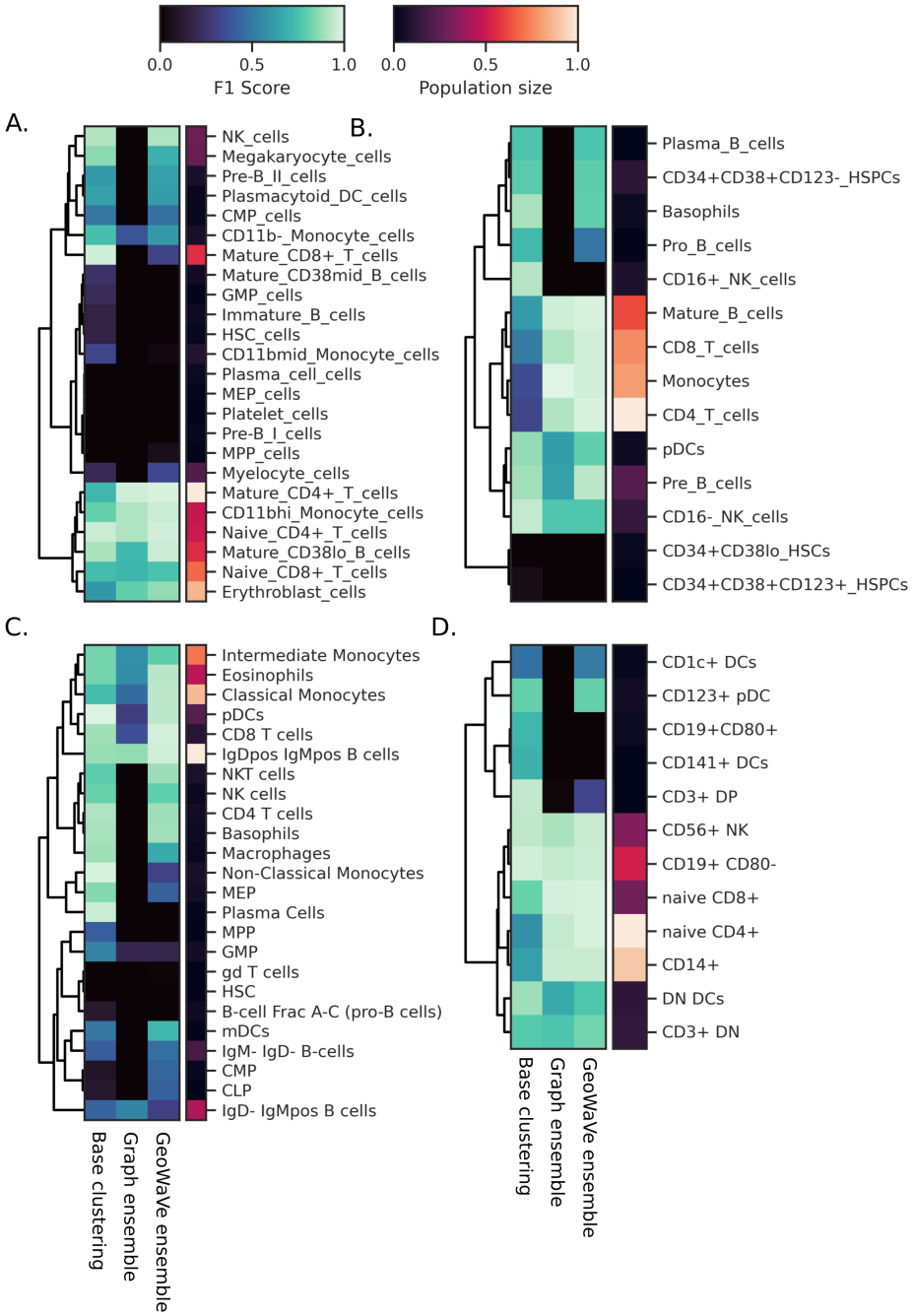

Figure 4.14: Heatmap of population F1 scores for the *Levine-13* (A), *Levine-32* (B), *Samusik* (C), and *OMIP* (D) data. Population level F1 scores are shown for the top performing algorithm amongst base clustering, graph ensemble, and GeoWaVe algorithms. Ground-truth populations (rows) are coloured by F1 score in the central heatmaps, with darker colours indicating a lower F1 score. On the right y-axis each row is labelled with an additional heatmap that describes the normalised size of the population (total number of events) relative to other populations within the same data.

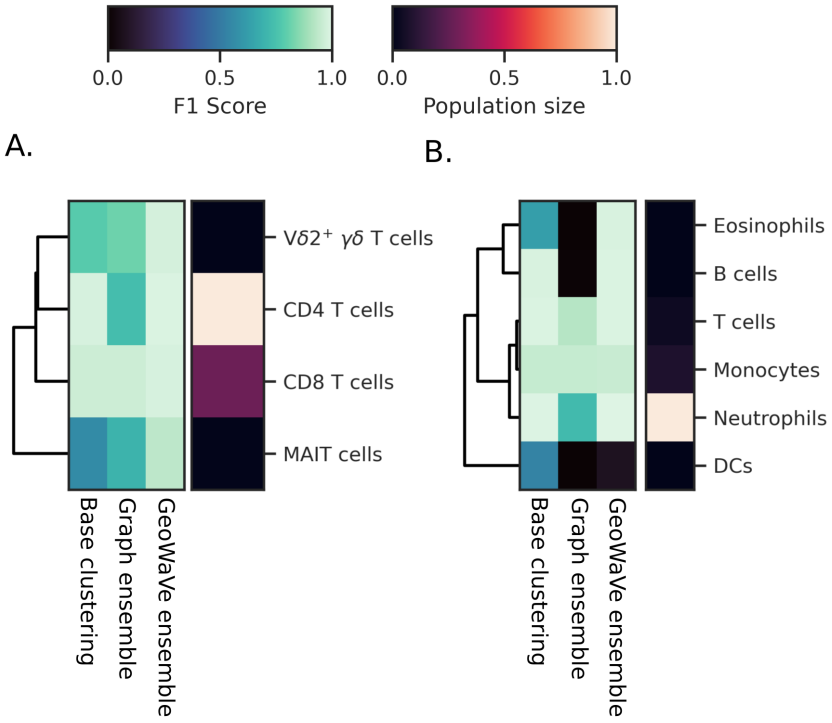

Figure 4.15: Heatmap of population F1 scores for the the *Sepsis* (A) and *Peritoneal Dialysis (PD)* (B) data. Population level F1 scores are shown for the top performing algorithm amongst base clustering, graph ensemble, and GeoWaVe algorithms. Ground-truth populations (rows) are coloured by F1 score in the central heatmaps, with darker colours indicating a lower F1 score. On the right y-axis each row is labelled with an additional heatmap that describes the normalised size of the population (total number of events) relative to other populations within the same data.

GeoWaVe matched the performance of base clustering algorithms for under-represented cell populations, whereas the graph ensemble clustering algorithms failed to do so. GeoWaVe also showed improved performance over base clustering algorithms for identifying populations such as monocytes, and subsets of T cells in the *Levine-32* data, myeloid DCs (mDCs) in the *Samusik* data, MAIT cells in the *Sepsis data*, and eosinophils in the *PD* data. Despite the success of GeoWave in comparison to graph ensemble methods, it still failed to identify some rare subsets completely, such as immature B cells in the *Levine-13* dataset, CD16+ NK cells in the *Levine-32* dataset, and plasma cells in the *Samusik* data. In contrast, base clustering algorithms showed either good performance or identification of at least some of the population.

4.3.5 GeoWaVe is computationally efficient.

Across all variations of the GeoWaVe algorithm run on the four benchmark datasets, the longest recorded runtime was for the 40 parameter Samusik data with 300,000 observations, at a runtime of 2 minutes and 12 seconds. All algorithms were run on an Ubuntu 20.04 operating system with an Intel i7-12700K processor with 12 cores and 32 gigabytes of RAM. The runtimes for all algorithms on benchmark data are reported in Table 4.2, 4.3, 4.4 and 4.5.

The maximum number of observations in the performance comparison experiments discussed so far was limited to 300,000. To assess the ability of GeoWaVe to scale to larger data, it was exposed to synthetic data of increasing size and complexity.

The runtime performance of the GeoWaVe algorithm is affected by two attributes of the data: the total number of observations and the overlap between clusters obtained by base clustering algorithms. Increasing overlap between clusters results in more observations being assigned to multiple consensus clusters, and the consensus cluster score (described in section 4.3.1) must be computed for each event assigned to multiple consensus clusters, therefore increasing the computational burden.

Synthetic data were generated using the $make\_blobs$ function from the Scikit-Learn library [177]. Data were generated with 15 dimensions (features), ranging from 500,000 to 4,000,000 observations in four batches, with increasing cluster standard deviations from one to four. The increasing variation would result in greater overlap between clusters, therefore challenging the performance of GeoWaVe (Figure 4.16). In total, 32 datasets were generated each containing ten Gaussian clusters. The synthetic datasets were clustered using three separate K-means algorithms, each with a different random seed and number of expected clusters (8, 10, and 12, respectively). Mini-batch processing with a batch size of 1024 was used to scale the K-means clustering to large data. The outputs of the K-means clustering algorithms served as input to a GeoWaVe clustering algorithm using Euclidean Ward hierarchical clustering of geometric medians.

The consensus cluster score is a simple calculation, and GeoWaVe employs multiprocessing to distribute these calculations across the available cores of a machine, resulting in an excellent performance, as demonstrated in Figure 4.17. GeoWaVe could generate ensemble

Dataset	No. of obs.	No. of parameters	FlowSOM	Pheno-graph	PARC	SPADE	PHATE + K-Means	PHATE + K-Means*
Levine-13	167,044	13	2.7s	3min 47s	1min 17s	11min 19s	24.7s	12min 3s
Levine-32	265,627	32	5.1s	26min 40s	1min 11s	20min 4s	1min 18s	12min 6s
Samusik	300,000	40	5.8s	34min 4s	1min 35s	13min 8s	1min 9s	12min 6s
OMIP	300,000	15	5.4s	6min 12s	2min 53s	7min 20s	29.1s	10min 20s
Sepsis	300,000	6	2.8s	4min 40s	2min 22s	18min 24s	27.5s	11min 27s
PD	300,000	9	2.9s	7min 13s	1min 17s	1min 17s	2min 57s	10min 46s

Table 4.2: Runtime performance of base clustering algorithms on benchmark data.

Dataset	No. of obs.	No. of parameters	MCLA	HGPA	HBGF
Levine-13	167,044	13	35.3s	14.7s	14.7s
Levine-32	265,627	32	55.8s	22.6s	22.6s
Samusik	300,000	40	11.9s	25.5s	25.5s
OMIP	300,000	15	14s	25.7s	25.7s
Sepsis	300,000	6	15.6s	25.8s	25.8s
PD	300,000	9	8.6s	25.4s	25.4s

Table 4.3: Runtime performance of graph ensemble clustering algorithms on benchmark data.

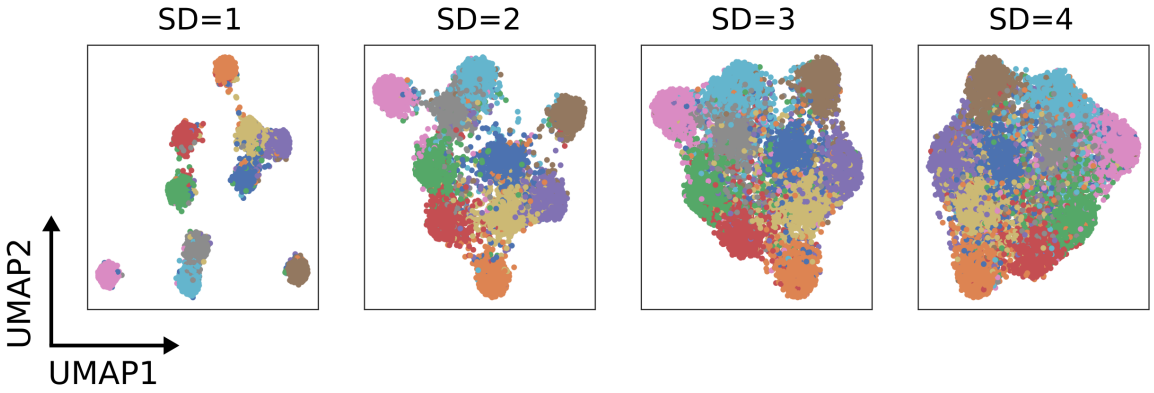

Figure 4.16: UMAP embeddings show the distribution of 10 Gaussian 'clouds' of synthetically generated data points with an increasing standard deviation (from 1 SD to 4 SD) causing increasing overlap.

clusters in less than 10 minutes, even for data scaling to millions of observations. With such reasonable runtimes, the investigator can easily experiment with different hyperparameters.

Dataset	No. of obs.	No. of params.	Ward	Manhattan	Euclidean	Cosine	Chebyshev
Levine-13	167,044	13	57.7s	30.1s	52.7s	42.8s	47.3s
Levine-32	265,627	32	53.6s	31.7s	47.8s	43.3s	47.4s
Samusik	300,000	40	1min 57s	1min 25s	1min 55s	1min 54s	1min 43s
OMIP	300,000	15	44.2s	25.9s	25.7s	15.9s	26.7s
Sepsis	300,000	6	14.3s	13.4s	12.4s	14.6s	14.6s
PD	300,000	9	1min 36s	6.2s	6.39s	8.9s	9.2s

Table 4.4: Runtime performance of GeoWaVe ensemble clustering algorithms on benchmark data, using Agglomerative hierarchical clustering.

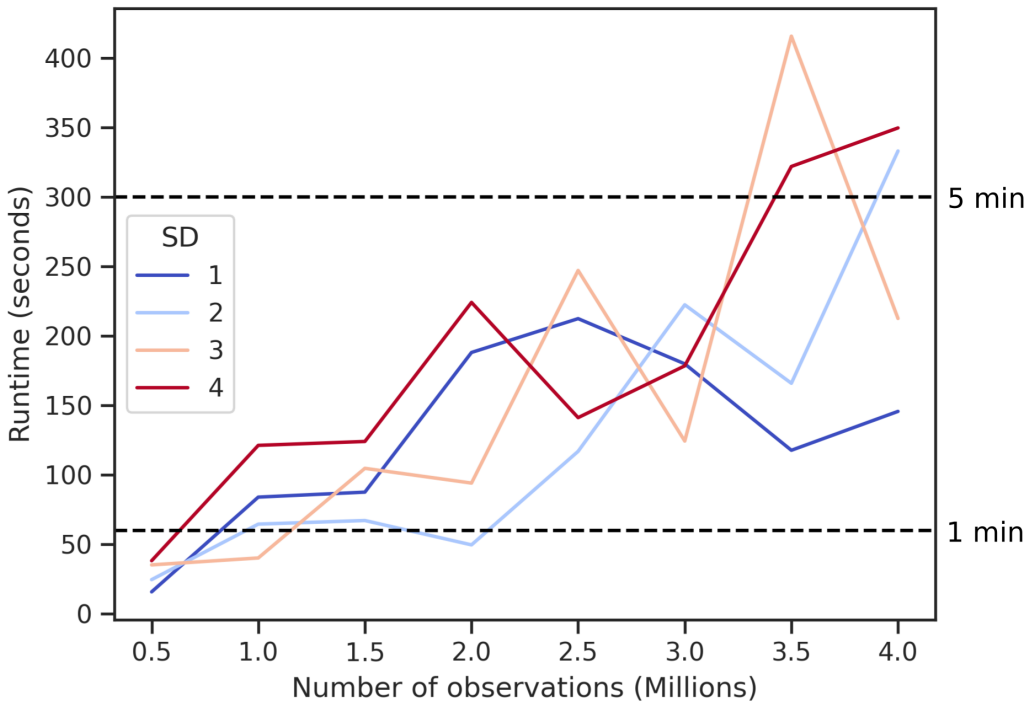

Figure 4.17: Runtime performance of GeoWaVe algorithm on randomly generated synthetic data consisting of ten Gaussian data point clouds with an increased number of observations. Four synthetic datasets are shown each with an increasing standard deviation (SD) used for the generation of Gaussian data point clouds resulting in more overlap between clusters.

Dataset	No. of obs.	No. of parameters	K-Means	Mean shift	Affinity Propagation
Levine-13	167,044	13	84.9s	43.6s	33.2s
Levine-32	265,627	32	47.2s	58.6s	55.3s
Samusik	300,000	40	2min 12s	1min 55s	1min 56s
OMIP	300,000	15	47.3s	37.6s	48.2s
Sepsis	300,000	6	13.1s	10.5s	18.3s
PD	300,000	9	1min 22s	8.2s	1min 13s

Table 4.5: Runtime performance of GeoWaVe ensemble clustering algorithms on benchmark data, using K-means, Mean shift, and Affinity Propagation.

4.4 Discussion

In this chapter, ensemble clustering was developed as a solution to reduce the variance commonly observed amongst clustering methods in the cytometry literature, where results depend upon hyperparameter choice and the particular context in which they are applied.

Presently, there is an absence of a "one size fits all" solution to clustering cytometry data, leaving scientists to rely on exploratory analysis that risks biasing results through data dredging [202]. Ensemble clustering offers an alternative by finding a consensus informed by the results of multiple clustering algorithms exposed to the same data. This multi-view approach theoretically offers robust, consistent, and stable solutions [206, 203] without biasing the analysis with the assumptions of a single algorithm. Employing ensemble clustering also forces the analyst to compare and contrast the results of multiple algorithms, which can be an informative exercise.

Ensemble clustering presents many challenges that come to bear when applied to complex data such as those generated with cytometry. Unlike supervised classification, no defined number of classes are provided by labelled examples. Different algorithms may generate different quantities of clusters, which must be compared and consolidated into consensus clusters. Cytometry data also tends to generate large data that can be difficult to handle with conventional computer resources. A challenge of increasing relevance as studies attempt to phenotype hundreds or thousands of subjects.

An existing ensemble approach that can scale to large data is the graph-based methods, such as HGPA, MCLA, and HBGF. These techniques were benchmarked against four independent datasets but failed to outperform individual clustering algorithms such as FlowSOM, PhenoGraph, or SPADE.

In response, an alternative heuristic ensemble method named GeoWaVe was suggested, suitable to the nature of cytometry data. Given that the dimensions of cytometry data are not beyond the comprehension of the investigator and meaningful phenotypes can be determined by considering sets of features, it is proposed to summarise each cluster contributing to a consensus by its geometric median in the feature space. The geometric medians can be visualised in a heat map as was shown in Figure 4.1. It was demonstrated in this chapter that clustering the matrix of these geometric medians can generate informative consensus clusters. GeoWaVe is novel in its computational efficiency, ability to handle millions of observations and communication of the consensus clusters to the investigator in a familiar manner that reflects the underlying biology.

GeoWaVe outperformed the graph methods of HGPA, MCLA, and HBGF. Using geometric medians also provides a helpful visual aid when choosing the number of consensus clusters

to be formed. One can estimate a suitable number of partitions by visualising the heat map of geometric medians in combination with UMAP, tSNE or PHATE embeddings. The visualisation allows the investigator to introduce informative priors and select clusters based on knowledge of the underlying biology. If uncertain, a range of partitions can be searched using the ConsensusClusterPlus method [213].

The use of geometric medians as a heuristic is not without limitations. Summarising a cluster using the geometric median tells nothing of the topology. A significant loss of information could result in misinformed consensus clusters that are not representative of the data. Additionally, the optimal choice of clustering method applied to the matrix of geometric medians is not immediately apparent and performance can vary depending on the data. It should be noted, however, that the use of a heuristic means that the run-time of GeoWaVe is fast enough to accommodate hyperparameter tuning. The investigator should therefore experiment with different clustering algorithms and hyperparameters and inspect the partitions on the geometric median heat maps and embeddings generated from a suitable dimension reduction technique. Although this fails to remove the exploratory approach to clustering cytometry data, it introduces the multi-view consensus necessary for robust results.

Weber & Robinson [204] performed a similar assessment of clustering algorithms without the focus on consensus methods and framed their assessment as a classification problem, inspired by the work of Samusik *et al.* [134]. They chose to use F1 score by first mapping clusters to ground-truth labels using the Hungarian algorithm and maximising F1 score across reference populations. This methodology was repeated in the present chapter and supported the conclusion that GeoWaVe ensemble methods outperform the graph ensemble methods of HGPA, MCLA, and HBGF. Closer inspection of individual population F1 scores revealed that graph ensemble methods often did not identify rare cell populations. Although identifying these subsets was improved in GeoWaVe, performance was worse than individual clustering algorithms in certain cases, and some populations, such as Platelet cells in the *Levine-13* data, remained unidentified. The performance of the base clustering algorithms for many rare cell populations was also poor, possibly impacting the performance of ensemble outputs. Further work is needed to generate clustering methodologies that directly address this limitation.

There is a significant flaw in assessing clustering performance through F1 score. Mapping clusters to ground-truth labels in such a way implies that a one-to-one relationship must exist between the clusters generated and the reference populations. Clustering analysis can be complicated by sub-structures in data captured as clusters but absent in the ground-truth labels. If the purpose of clustering cytometry data is to identify a precise number of clusters, then this form of evaluation seems justified, but one could argue in such a scenario that a supervised classification approach is more suitable. Clustering analysis tends to be applied in the interest of discovery when the number of clusters is unknown. Despite this flaw, it was deemed necessary to replicate the methods of Webber *et al.* [204], identifying the role population size plays. It showed that although the consensus clustering of geometric medians outperforms graph-based methods, there is still work to ensure rare cell populations do not go undetected. It would be advisable that if rare cell populations are suspected to be present, the consensus is formed by methods with high resolution, such as those formed on nearest-neighbour graphs [131, 134, 214].

Future work should focus on more diverse ensemble clustering. In this work, four algorithm classes were chosen based on their popularity in the cytometry literature and their available implementations. However, a wide variety of clustering algorithms could be explored for inclusion in ensemble clustering. The use of ensemble clustering is more prevalent in the scRNA-seq literature with examples such as SC3 [210], SAFEClustering [219], and SCENA [220]. As discussed, some of these methods might not scale to the size of data encountered in cytometry data analysis, which can be hundreds of times greater than what is encountered in scRNAseq analysis. There are efforts to address the computational complexity, such as improvements to SC3 that currently exist as a pre-print publication [221]. Other solutions to the computational complexity may come from advances in the statistical and computational literature, such as consensus formed on heuristics of cluster similarity using metrics $e.g.$ Jaccard index [222]. In the meantime, clustering on geometric medians could be a viable solution for cytometry data analysis and has been implemented in a manner that is compatible with the CytoPy software described in Chapter 3.1 of this thesis.

5 | Phenotypes of severe sepsis patients and their relationship with mortality and causative pathogen.

5.1 Introduction

Sepsis is a life-threatening syndrome characterised by organ failure caused by a dysregulated host response to infection requiring complex patient management and care. Timely diagnosis of sepsis is vital, but identifying those at risk of higher mortality is also imperative for triaging intensive care. Early prediction of patient outcomes has implications for resource allocation and personalised care. Historically, it has been the role of severity scores to direct care, which can be broadly categorised into those that indicate the risk of in-hospital mortality *e.g.* Acute Physiology and Chronic Health Evaluation (APACHE) or Simplified Acute Physiology Score (SAPS), and those that also define the degree of organ failure as well as indicating the likelihood of mortality *e.g.* Sequential Organ Failure Assessment (SOFA) and Multiple Organ Dysfunction score (MODS) [223]. These tools rely on routinely collected clinical data and observations, yet their performance for predicting in-hospital mortality is relatively poor [225, 224].

Novel prognostic biomarkers derived from the pathophysiology of sepsis could be more informative for clinicians and help guide the treatment and monitoring of the disease. Several biomarkers have been proposed, the most well-studied being C-reactive protein (CRP) and Procalcitonin (PCT). Elevated plasma level of CRP on admission has failed to present itself as a reliable predictor of mortality [59, 227, 226], and although meta-analysis has shown that early levels of PCT plasma levels significantly differ between survivors and non-survivors, high heterogeneity between study populations puts the general applicability of these findings into question [60].

Many biomarkers have been proposed for prognostic use in sepsis, covering multiple biological systems. Biomarkers of cardiovascular function and circulation have shown potential; for example, pro-adrenomedullin levels in the blood (a marker of vascular permeability, inflammation, endothelial barrier regulation, and stabilisation of micro-circulation [228]) significantly predicted mortality with an AUC of 0.87 [154, 71]. An active research area with potential is the study of microRNAs (miRNAs), known to regulate the pathophysiology of

sepsis, such as pro-inflammatory cytokine pathways, with several studies reporting AUC scores >0.8 [71]. Biomarkers of the innate immune response have shown promise as predictors of mortality. CD64 expression on neutrophils has a reported AUC score for predicting mortality of 0.75 [65, 229], although definitions of outcome differ between studies. IL-6 potentially shows high specificity for predicting early mortality [50], but its ability to predict mortality at later time points is poor [230, 51]. Lymphopenia is a hallmark of sepsis, is associated with bacteremia, and is inversely correlated with outcome [231]. The neutrophil to lymphocyte ratio is also increased in sepsis, possibly due to lymphocyte apoptosis, and has been implicated as a potential diagnostic biomarker [232].

Most biomarker research in sepsis focuses on diagnosis rather than prognosis, given the importance of early interventions, such as anti-microbials, on survival [233, 78, 76]. The reality of sepsis is that the causative pathogen and appropriate antimicrobials are unknown at the time of diagnosis, and therefore broad-spectrum antibiotics are often administered. Broad-spectrum antibiotic use is a controversial topic with arguments for [234] and against [235] the rapid use of empirical broad-spectrum antibiotics in sepsis. Whilst some argue that the risk to a patient by withholding antibiotics in suspected sepsis whilst awaiting confirmation of the causative pathogen is substantially greater than the risk of using empirical broad-spectrum antibiotics, others highlight the risk of creating the conditions for multidrug-resistant organisms to thrive and the adverse drug effects of overly broad treatment regimens. What remains clear, however, is the need for more specific identification of causative pathogens without requiring lengthy bacterial culture that could take up to 72 hours to yield results. Earlier recognition of the causative pathogen could lead to more targeted therapy and contribute to improved antibiotic stewardship [236].

As highlighted by a recent technology review, multiple molecular diagnostic methods for pathogen identification have come to market, potentially reducing the time needed to identify the causative pathogen by up to 30 hours [73]. Many technologies require a positive blood culture, yet the sensitivity of blood cultures is negatively impacted by antibiotic use, and the incidence of culture-negative sepsis is reported as anywhere between 28 and 80% [77].

Several biomarkers have been investigated for their ability to distinguish Gram-positive and Gram-negative infections, the most notable being PCT. In one study, plasma levels of CRP and PCT were reported to have AUC scores of 0.79 and 0.68, respectively [237], and other

studies have reported plasma concentrations of PCT being higher amongst those with Gram-negative infections compared to Gram-positive [85, 87, 83, 238]. Other biomarkers that have shown promise are soluble CD14 (sCD14 or 'presepsin'), which may be increased in Gram-negative bacteremia [69], and the cytokines IL-1β, IL-6, and IL-18, with concentrations significantly higher in patients with Gram-positive infection [82].

Another potential biomarker for identifying aetiology in sepsis is unconventional T cells such as Mucosal Associated Invariant T cells (MAIT) and $\gamma\delta$ T cells, which are capable of microbial pattern recognition and bridging the innate and adaptive immune system by orchestrating acute inflammatory responses. MAIT cells are an abundant population of T cells characterised by a semi-invariant T cell antigen receptor (TCR) with specificity for microbial riboflavin-derivative antigens presented by HLA-1b major histocompatibility complex (MHC)-related protein 1 (MR1) [28, 29]. The cell specificity towards microbial vitamin B metabolites makes this population an interesting candidate for predictive signatures of infectious disease [30, 28, 239]. $\gamma\delta$ T cells are invariant T cells with a TCR composed of a γ and δ chain and are capable of antigen recognition independent of MHC presentation. Vγ9/Vδ2 $\gamma\delta$ T cells are highly responsive to the microbial isoprenoid precursor (E)-4-hydroxy-3-methyl-but-2-enyl pyrophosphate (HMB-PP), a molecule produced by the majority of Gram-negative pathogens, some Gram-positive pathogens, but notably absent from staphylococci, streptococci and fungi [25]. These unconventional T cells' innate functionality and specificity could potentially contribute to the pathogen-specific signatures that have already been shown to successfully characterise patients with acute peritonitis [175, 240].

In this chapter, I provide a descriptive overview of the comprehensive data captured from patients diagnosed with sepsis and sampled within 36 hours of diagnosis. Comparisons are made between survivors and non-survivors, culture-positive and culture-negative sepsis, and amongst those with confirmed infection, the differences observed between Gram-positive and Gram-negative causative pathogens. The routine clinical data available for these patients are described first, summarising the cohort and giving perspective on the value of existing data in the clinic. Then soluble analytes and the immunophenotype of those patients are characterised, focusing on monocytes, neutrophils, conventional $CD4^+$ and $CD8^+$ T cells, $\gamma\delta$ T cells, and MAIT cells. The data described within this chapter provides the basis for input variables for multivariant modelling discussed in Chapter 6.

5.2 Aims

1. Evaluate the performance of routinely available clinical parameters and biomarkers for predicting mortality, positive bacterial culture, and/or the causative pathogen.
2. Define the difference in soluble components of immunity, such as cytokines and chemokines, during early sepsis comparing survivors and non-survivors, and Gram-negative and Gram-positive causative pathogen.
3. Define the phenotype of classical and unconventional T cells, monocytes, and neutrophils in early sepsis.
4. Identify cellular phenotypes that are correlated with either mortality, positive bacterial culture, or causative pathogen in early sepsis.

5.3 Results

5.3.1 Characterising a cohort of acute severe sepsis patients.

A total of 77 severe sepsis patients in the intensive care unit (ICU) in Cardiff, UK, were enrolled in the ILTIS study (see Material & Methods section 2.1) between 2018 and 2021. Figure 5.1 provides an overview of the sampling and phenotyping of these patients. Whole blood was obtained within the first 36 hours of sepsis (defined as a SOFA score greater than 2 with suspected infection) from each patient, and flow cytometry was employed to capture the phenotype of monocytes, neutrophils, classical T cells, and unconventional T cells. Cell-free plasma was obtained and frozen within the first hour of sample collection and later analysed with Luminex™ multi-plex assays and standard plate-based ELISA to quantify soluble biomarkers.

Patients were categorised based on mortality and underlying causative pathogen (Figure 5.2). Within each category, the number of patients with available data for activated T cell subset staining, memory T cell subset staining, monocyte and neutrophil staining, cytokine/chemokine Luminex™ multiplex assay/ELISA, clinical parameters, and lipid data are shown by coloured boxes beneath each category. Cytometry staining data are absent where a technical error occurred during sample processing or the sample integrity/volume prevented acquisition. Luminex™ multiplex assay/ELISA results are absent for seven patients because of insufficient sample volume. Mortality at 30 and 90 days captured short- and medium-term outcomes. The causative pathogen was determined from the patient's discharge letter and cross-checked with their microbiology results. Twenty five patients were excluded from comparing the causative pathogen because the cause of infection could not be ascertained.

Table 5.1 and 5.2 show a comparison of patient demographics, admission criteria, and clinical observations between survivors and non-survivors at 30 and 90 days after sepsis diagnosis. A mortality rate of 22.1% and 27.3% was observed at 30 and 90 days after sepsis diagnosis, respectively. Mortality was lower than the international average of 24.4% and 32.2% reported by a recent meta-analysis [2]. Compared to national statistics, the intensive care national audit and research centre reported critical care unit mortality as 27.6% for England, Wales, and Northern Ireland in 2012 [241]. A multi-centre prevalence study of sepsis on general wards and emergency departments in Wales in 2016 reported a 22% mortality rate at

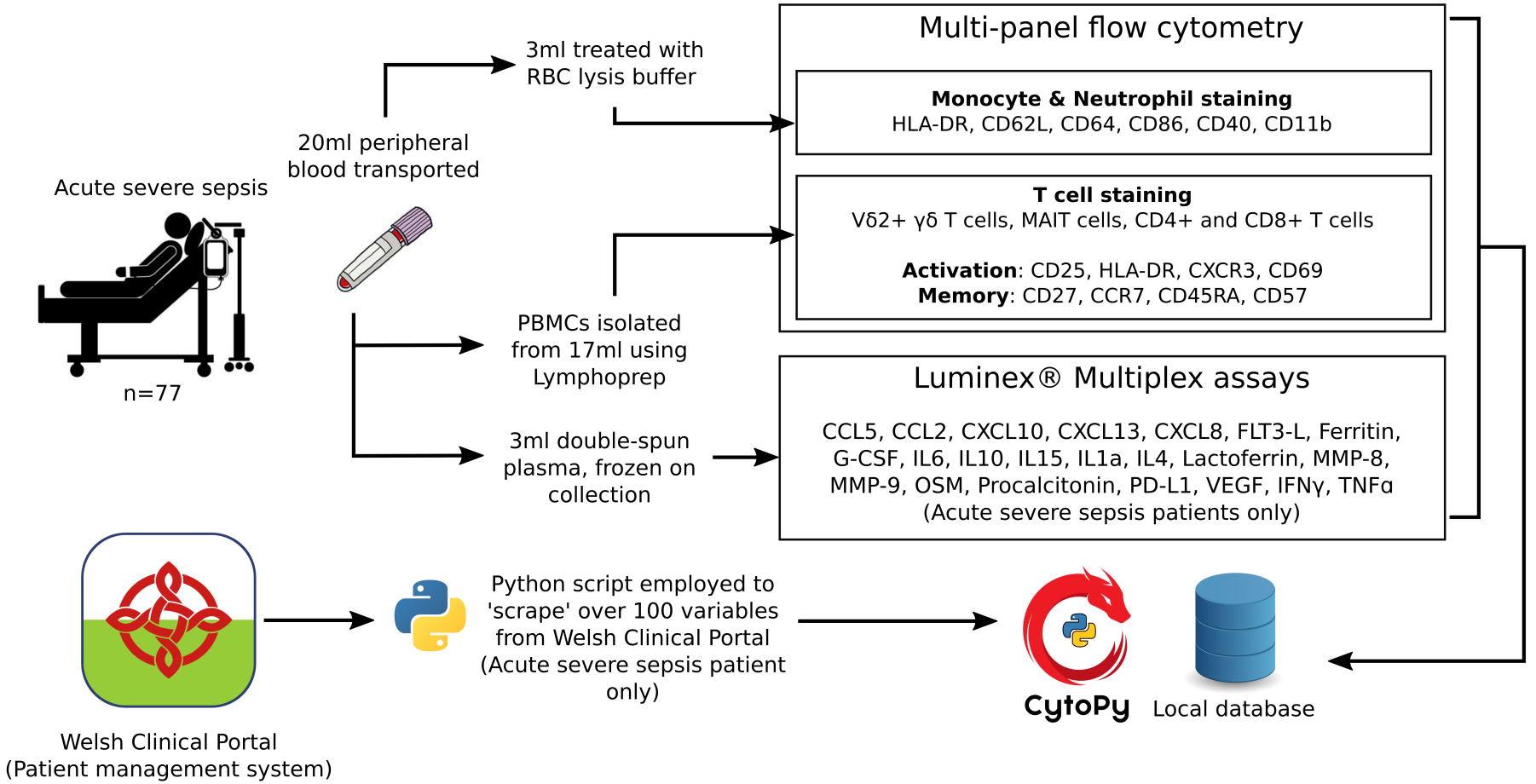

Figure 5.1: Overview of ILTIS study data capture. Total leukocyte fraction and PBMCs from whole blood in EDTA were phenotyped by flow cytometry to characterise monocytes, neutrophils, and T cells. Cell-free plasma was simultaneously isolated and frozen at -80°C and later analysed for soluble biomarkers using Luminex™ Multiplex assays. Data were combined with extensive clinical information obtained from the patient management system and stored in a local CytoPy database.

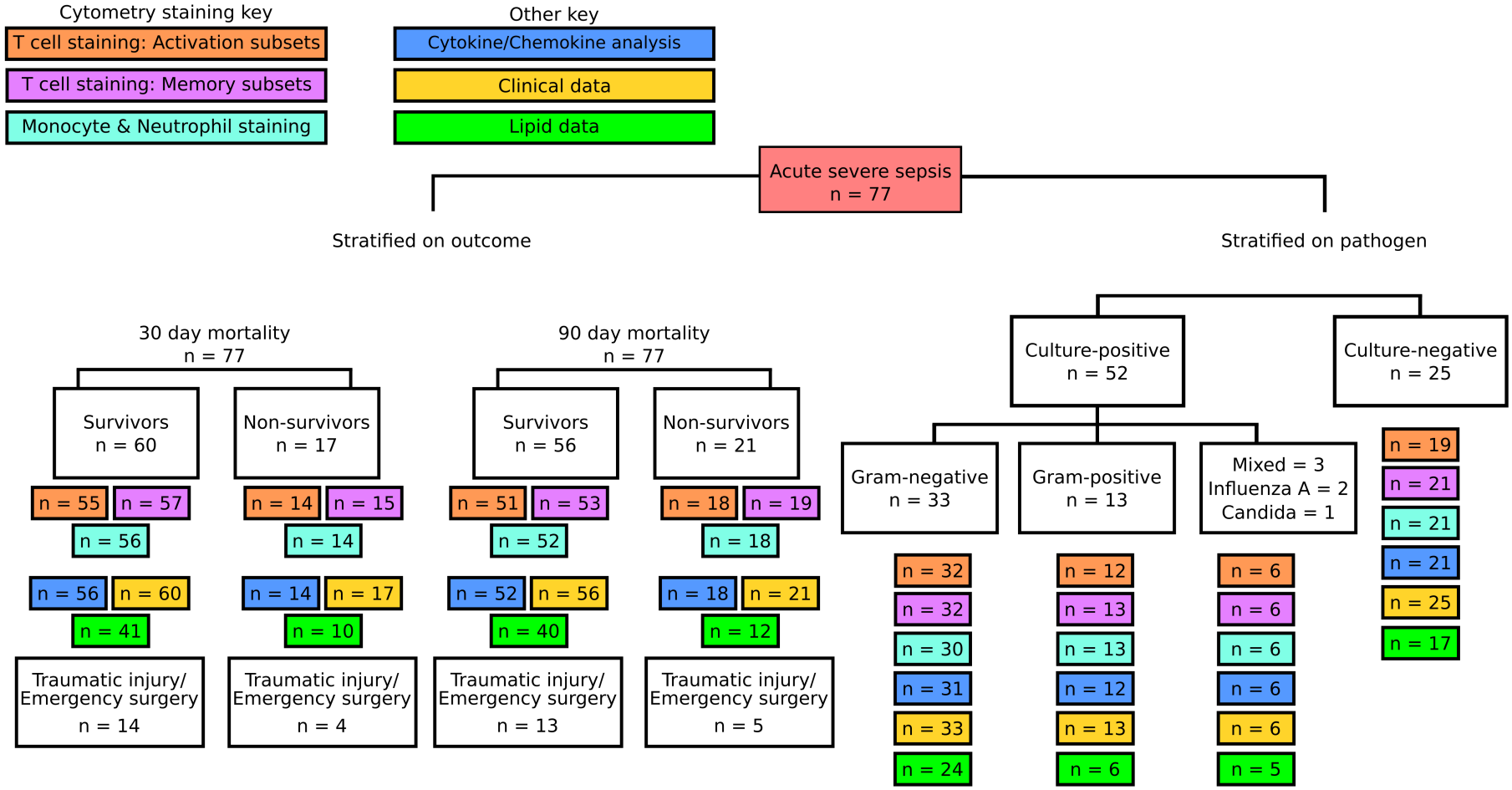

Figure 5.2: Stratification of acute severe sepsis patients and available data within each subcategory. Patients were divided into survivors and non-survivors at 30 and 90 days after sepsis diagnosis. The total number of patients with available data sources is shown by colour boxes within each category. Patients were also categorised into those with and without microbiologically confirmed infection, and amongst those with a positive culture, on the Gram-status of the causative pathogen.

	Survivors (n=60)	Non-survivor (n=17)	P-value
Age (Years)	65.5 [18 - 86]	71.0 [47 - 84]	0.149
Male (%)	51.7%	70.6%	0.268
BMI	28.7 [17.3 - 51.6]	29.4 [22.4 - 52.2]	0.751
APACHE II Score	17.0 [0 - 30.0]	19.0 [0 - 33]	0.294
Days in critical care	9.0 [0.8 - 64.8]	9.4 [1.5 - 34.4]	0.980
Mechanically ventilated (%)	53.3%	76.4%	0.103
Renal Rt (%)	35.0%	47.1%	0.404
Trauma/Emergency surgery	23.3%	23.5%	1.000
Microbiology confirmed	68.3%	64.7%	0.776

Table 5.1: Comparison of survivors and non-survivors at 30 days after diagnosis of sepsis. Continuous values are reported as the median [range] and P-values reported using two-tailed Mann-Whitney U test. P-values for proportions were generated using Fishers Exact test.

	Survivors (n=56)	Non-survivor (n=21)	P-value
Age (Years)	65.0 [18 - 86]	71.0 [47 - 84]	0.067
Male (%)	51.8%	66.7%	0.307
BMI	28.7 [17.3 - 51.6]	29.4 [22.4 - 52.2]	0.496
APACHE II Score	17.0 [0 - 30.0]	19.0 [0 - 33.0]	0.132
Days in critical care	7.83 [0.8 - 64.8]	9.9 [1.5 - 55.1]	0.403
Mechanically ventilated (%)	51.8%	76.2%	0.070
Renal Rt (%)	32.1%	47.6%	0.120
Trauma/Emergency surgery	23.2%	23.8%	1.000
Microbiology confirmed	67.9%	66.7%	1.000

Table 5.2: Comparison of survivors and non-survivors at 90 days after diagnosis of sepsis. Continuous values are reported as the median [range] and P-values reported using two-tailed Mann-Whitney U test. P-values for proportions were generated using Fishers Exact test.

30 days and 31.5% at 90 days [242]. Survivors appeared slightly younger than non-survivors when comparing both 30- and 90-day mortality, although not significantly. Statistics on medical interventions during the patient's ICU stay but after their sepsis diagnosis were collated. The proportion of patients undergoing renal replacement therapy or mechanical ventilation was less among survivors, but those differences were not statistically significant.

A total of 52 patients (67.5% of the cohort) had a microbiologically confirmed infection, which was slightly higher than previous descriptions of 30 to 40% of sepsis diagnoses yielding a positive bacterial culture [243]. Three patients had a mixed culture result with an undefined causative pathogen, two patients had an Influenza A infection with no bacterial isolates identified, and one had candidiasis (Figure 5.2). The remaining cohort could be divided into Gram-negative, Gram-positive, and culture-negative sepsis (Table 5.3). Pa-

tients with a microbiologically confirmed infection were of greater age, and the majority were Gram-negative (63.5%). Fewer patients with a Gram-positive or Gram-negative cause had undergone emergency surgery or were admitted due to traumatic injuries compared to culture-negative sepsis patients. No significant difference in patient demographics, severity score, therapeutic intervention, or mortality was observed between Gram-negative, Gram-positive, and culture-negative sepsis.

5.3.2 Insights from routine clinical data in sepsis patients

Before extensive immunophenotyping, routine clinical data such as full blood count, liver profile, and blood gas analyser data were explored. Individual biomarkers were compared between the patient subsets described in Figure 5.2. Variables captured for less than five unique patients were removed, leaving 63 routinely collected variables. Average and interquartile range for each variable are detailed in the appendix, Table A.1.

Routine clinical data collected retrospectively are not driven by a study design that would ensure conformity between patients but rather by the patient's clinical condition at the time of sampling. The consequence is exclusive variables for some patients resulting in missing data for others and varying time points for sampling. The complication of different sampling time points drove the decision to summarise routine clinical data as follows:

- The sample closest to the study enrolment time was chosen and all other time points were ignored, or
- Samples taken 48 hours prior to enrolment or 8 hours after enrolment were averaged.

The time window for averaging was chosen to capture measurements in the hours prior to diagnosis of sepsis, with an 8-hour delay after enrolment to account for sampling delay.

Routine data readily available for the majority of subjects in this study (i.e. less than 10% missing data) included extensively characterised biomarkers such as CRP and arterial lactate concentration. CRP is used clinically as a marker of inflammation and is recognised as a diagnostic marker in sepsis [155]. As shown in Figure 5.3, all patients had an elevated blood CRP level (above the local hospital laboratory reference value of 5 mg/L). While there was no significant difference between survivors and non-survivors or between culture-positive and culture-negative patients, CRP was significantly increased in Gram-positive infections

	Gram-negative (n=33)	Gram-positive (n=13)	Culture-negative (n=25)	P-value
Age (Years)	71.0 [36.0 - 86.0]	69.0 [31.0 - 84.0]	60.0 [18.0 - 80.0]	0.108
Male (%)	57.6%	76.9%	48.0%	0.259
BMI	29.4 [20.2 - 52.2]	27.8 [17.3 - 34.2]	29.9 [20.2 - 42.7]	0.393
APACHE II Score	19.0 [4.0 - 33.0]	19.0 [8.0 - 25.0]	17.0 [0.0 - 30.0]	0.482
Days in critical care	7.8 [0.8 - 48.0]	12.2 [1.0 - 64.8]	8.8 [1.1 - 19.9]	0.362
Mechanically ventilated (%)	54.5%	53.8%	60.0%	0.904
Renal Rt (%)	36.4%	61.5%	28.0%	0.120
Trauma/Emergency surgery	15.2%	23.1%	32.0	0.320
Mortality (at 30 days)	24.2%	15.4%	24.0 %	0.720
Mortality (at 90 days)	27.3%	23.1%	28.0	0.945
Infection site (%)				
Abdominal	18.2%	23.1%		0.698
Respiratory	33.3%	38.5%		0.744
Urinary	39.4%	0%		
Soft tissue	3.0%	23.1%		0.062
Cardiovascular	3.0%	0%		
Unknown	3.0%	15.4%		0.189

Table 5.3: Comparison of Gram-negative, Gram-positive, and unknown causative pathogen in sepsis patients. Continuous values are reported as the median [range] and p-values generated using Kruskal-Wallis H-test for independent samples. P-values for comparison of categorical variables were generated using Fishers Exact test.

compared to Gram-negative infections. Arterial lactate concentration (captured by blood gas analysers) is a marker of tissue hypoxia and is clinically used as a measure of severity [244]. Despite this, lactate was only significantly increased in those who died 90 days after enrolment and was only significantly different when considering the sample closest to enrolment time and not the average within the 48-hour window (Figure 5.4).

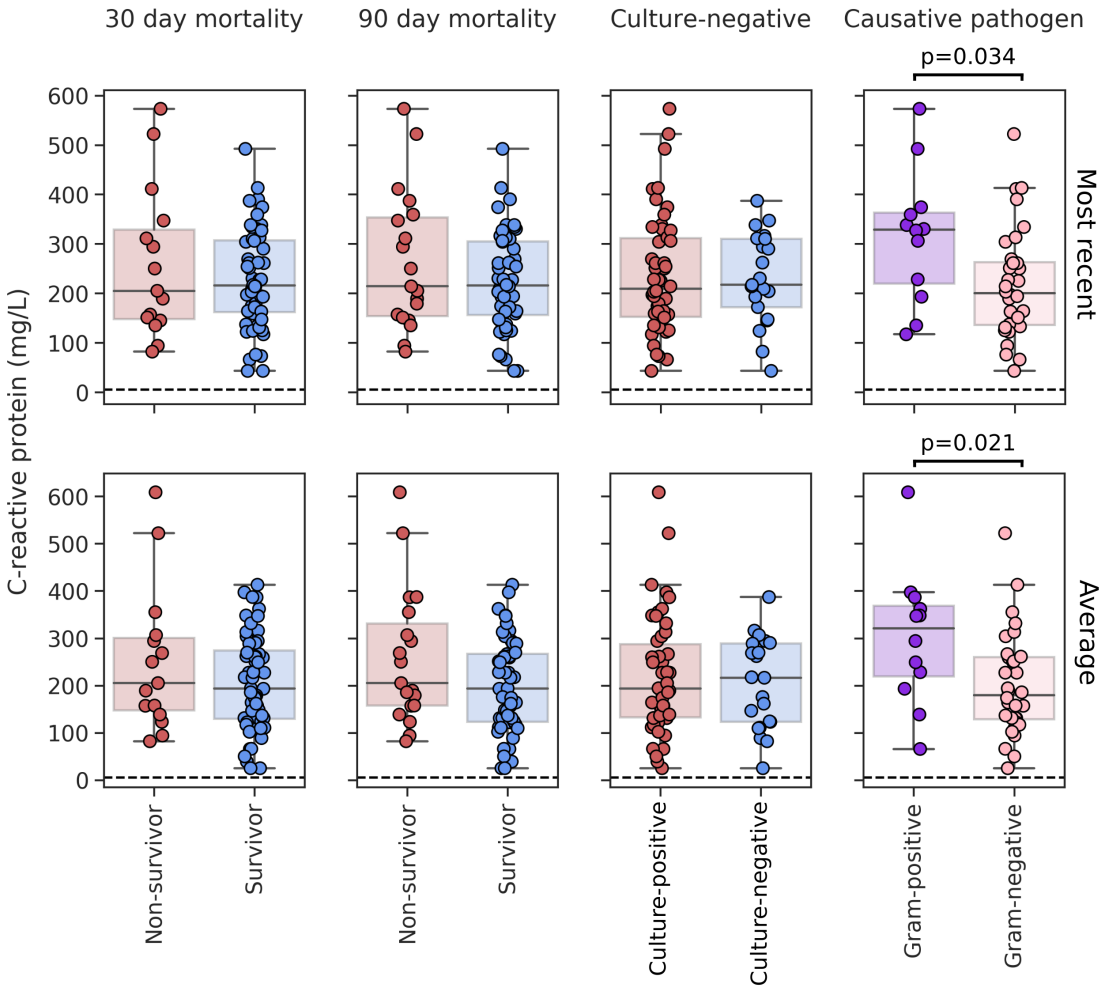

Figure 5.3: C-reactive-protein (CRP) concentration in blood taken from patients diagnosed with sepsis and enrolled into the ILTIS study. Values are shown for samples taken closest to enrolment time (top) and the average concentration within a window of 48 hours prior enrolment up until 8 hours after enrolment (bottom). P-values report comparison using two-tailed Mann-Whitney U test. Dotted line represents the reference range used for CRP by Cardiff and Vale Health Board and values above this line are considered 'raised levels of CRP'.

Additionally, cell counts for major immune cell populations were available for most subjects. No significant difference was observed amongst these subsets when comparing the most recent sample relative to the enrolment time (Figure 5.5).

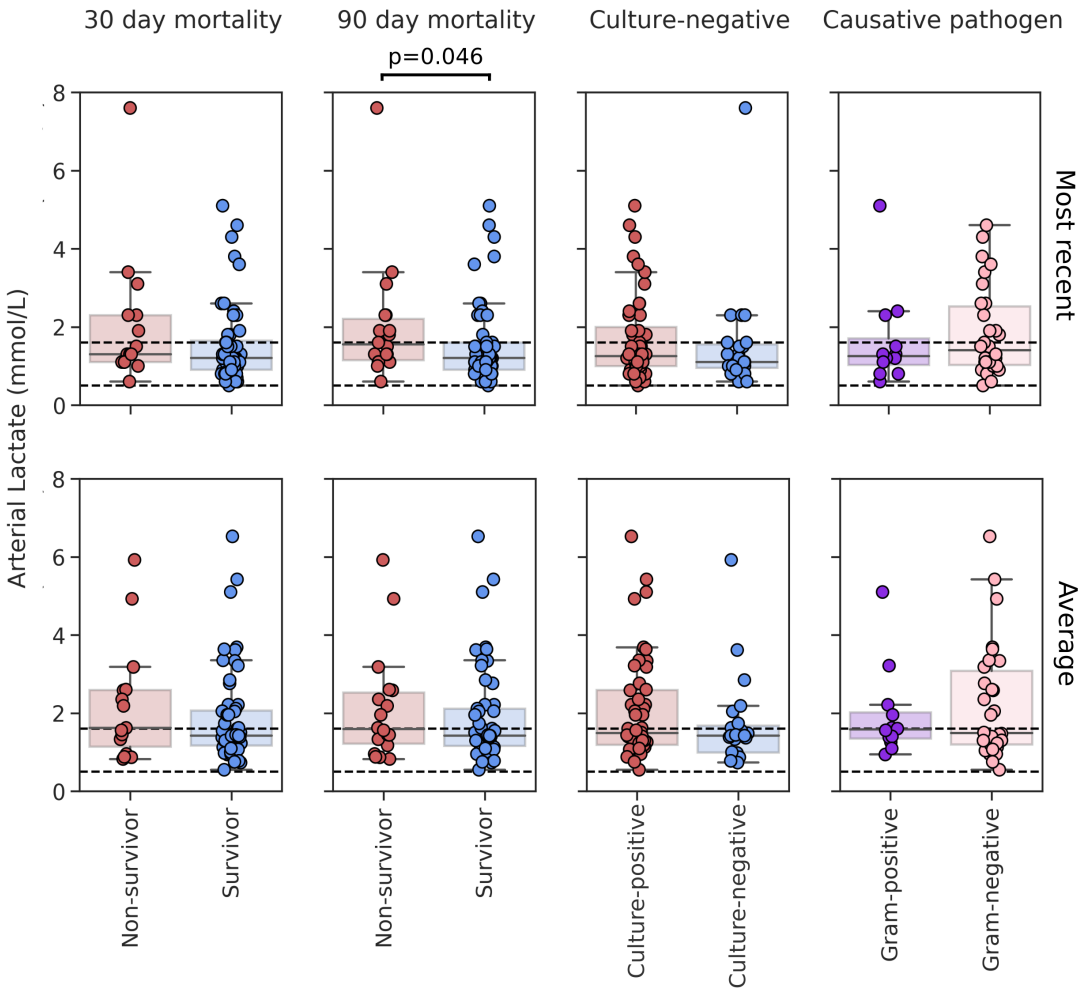

Figure 5.4: Lactate concentration in blood taken from patients diagnosed with sepsis and enrolled into the ILTIS study. Values are shown for samples taken closest to enrolment time (top) and the average concentration within a window of 48 hours prior enrolment up until 8 hours after enrolment (bottom). P-values report comparison using two-tailed Mann-Whitney U test. Dotted line represents the reference range for blood lactate used by Cardiff and Vale Health Board and values outside this range are considered 'abnormal'.

The approach taken in this study of broad data mining of all available routine clinical data provides an overwhelming number of possible biomarkers. This situation is best addressed with feature selection and multivariant modelling that will be explored in Chapter 6. Before this, however, it was valuable to ascertain if any particular biomarker successfully differentiated the subgroups of interest amongst sepsis patients. All biomarkers were considered individually and compared between patient subsets by a two-tailed Mann-Whitney U test, with correction for multiple comparisons made using the Benjamini–Hochberg procedure to control false discovery rate at an α of 0.05. Biomarkers were excluded if less than five observations were available for any patient subset.

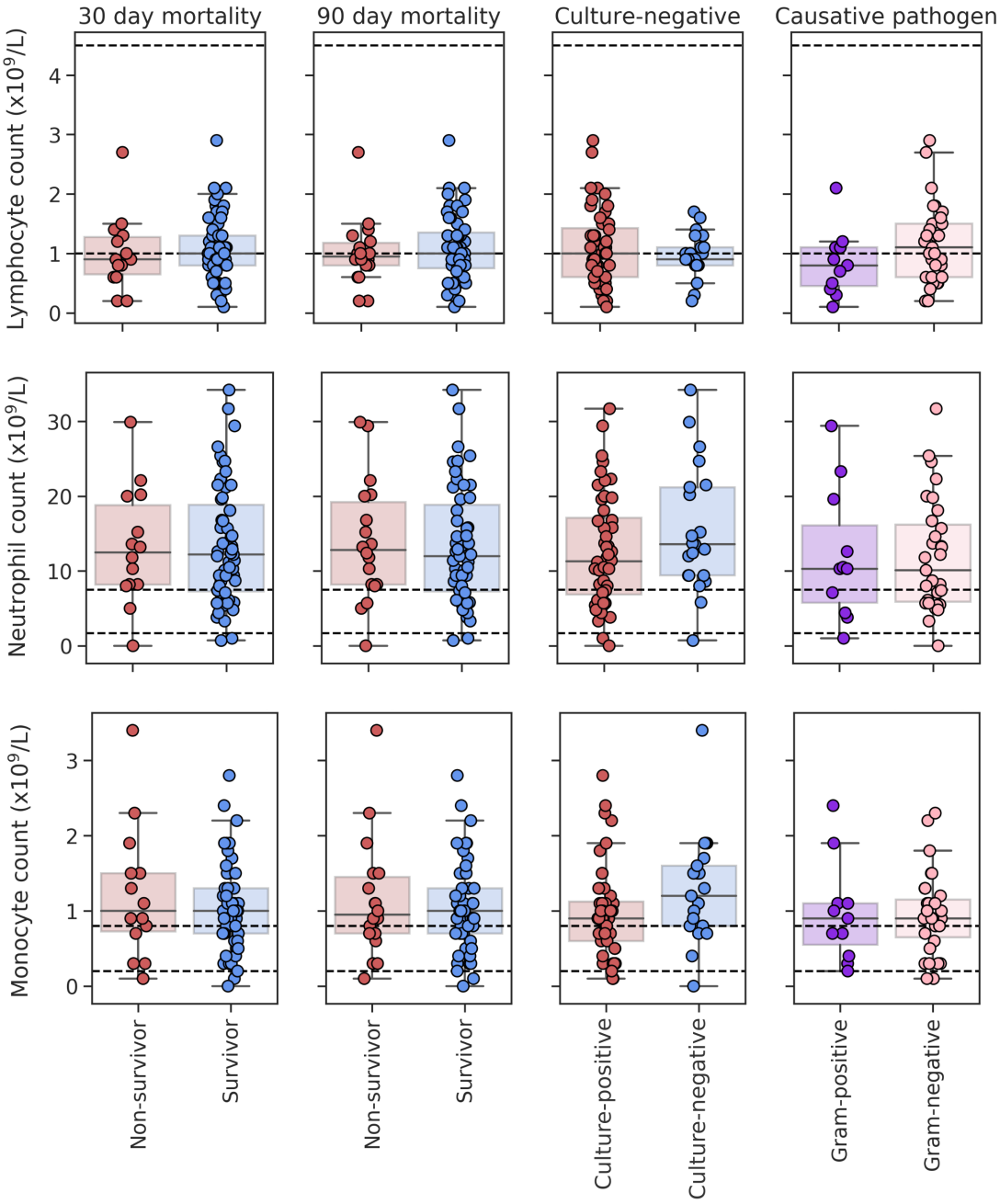

Figure 5.5: Concentration of lymphocytes, neutrophils, and monocytes in blood from patients diagnosed with sepsis and enrolled into the ILTIS study. Values are shown for samples taken closest to enrolment time. Dotted line represents the reference range for cell counts used by Cardiff and Vale Health Board and values outside this range are considered 'abnormal'.

Figure 5.6A shows the common language effect size (CLES) vs the corrected p-value for all biomarkers routinely collected under the null hypothesis that values are similar amongst patient sub-groups. The CLES gives the probability that a random observation from the distribution of non-survivors will be higher than a random observation from the distribution

of survivors. The p-value for each variable was generated using a Mann-Whitney U Test with the Benjamini-Hochberg procedure applied to control the false discovery rate. The only routinely collected variable that showed a significant difference between survivors and non-survivors was the venous fraction of inspired oxygen (FiO2) value taken closest to the diagnosis of sepsis, with increased levels amongst non-survivors compared to survivors, corroborating findings by Dahl *et al.* [245]. No other biomarkers demonstrated a significant difference relative to mortality. Figure 5.6B demonstrates the same is true for identifying those without identification of causative pathogens or when comparing Gram-positive and Gram-negative infections.

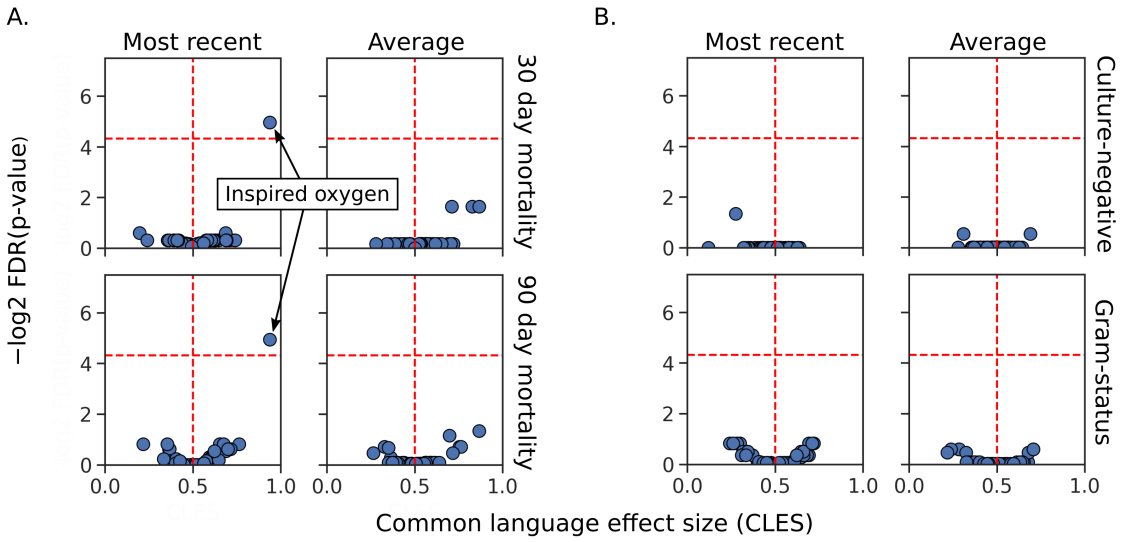

Figure 5.6: Comparisons of variables captured in routine clinical data and their ability to differentiate mortality, culture-positivity, and the Gram status of the causative pathogen in sepsis. Data are summarised as either the most recent value relative to enrolment time or the average value within a window of 48 hours prior to enrolment and 8 hours after enrolment. On the left panel (A), survivors at 30 and 90 days after enrolment are compared to non-survivors, whilst on the right (B), those with confirmed microbiology are compared to those without. Where microbiological confirmation is present, Gram-positive and Gram-negative pathogens are also compared (B). P-values are reported using a two-tailed non-parametric Mann-Whitney U test with Benjamini–Hochberg procedure to control the false discovery rate at an α of 0.05. Values below the horizontal red line have a p-value greater than 0.05. The vertical red line represents a CLES of 50%.

5.3.3 Quantifying soluble biomarkers from plasma of sepsis patients.

Routine clinical practice does not currently quantify the various immunophenotypes and signalling cascades present in acute severe infectious diseases such as sepsis. As described in Figure 5.1, cell-free plasma was obtained from whole blood samples and analysed for cy-

tokines and chemokines to identify potentially valuable biomarkers. Samples were analysed using multi-plex Luminex^{TM} assays (see 2.1.7 for complete methodology) in two batches. Batch effects were minimised by converting to log base two followed by computing the z-score (by subtracting the mean from each value and dividing by the standard deviation) as previously described by Tomic *et al.* [173]. Figure 5.7 shows the improved overlap in the distribution of biomarker concentration before and after applying the aforementioned batch correction technique.

The differences amongst analytes measured by Luminex multiplex assays and ELISAs for survivors and non-survivors 30 days after sepsis diagnosis are shown in Figure 5.8. It should be noted that for many analytes data crowd at the bottom or top of the analytical range (most obvious for TNFα, IL-10, and IL-1α), this reflects the number of samples where concentrations of analytes were outside the detectable range of the assay. The issue was later addressed by converting measurements to discrete variables above and below detection limits.

CXCL10 levels in plasma were significantly decreased in non-survivors when observing 30-day mortality (Figure 5.8). CXCL10 is a chemokine produced by T cells, a ligand for CXCR3, and is important for the recruitment of lymphocytes to the sites of infection [17]. IL-15, a 'bi-directional' cytokine with both pro-inflammatory and immunoregulatory effects [246], was significantly increased in non-survivors. Over 80% of patients had levels below the detection limit, so the significance of this finding should be treated with caution. The statistical significance of the trends seen for CXCL10 and IL-15 was diminished when observing 90-day mortality (Figure 5.9). Although this could suggest that these analytes are more informative for early mortality, class imbalance (the ratio between survivors and non-survivors) is less severe for the 90-day mortality endpoint, and therefore a larger sample size would be needed to clarify that the relationship is different between 30- and 90-day mortality. No other analytes were significantly different between survivors and non-survivors.

The same analytes were compared amongst patients with and without microbiologically confirmed infections (Figure 5.10). Flt3L (FMS-related tyrosine kinase 3 ligand) and it's corresponding tyrosine kinase receptor (Flt3) regulates dendritic cell (DCs) development in steady state, and is required for the generation of non-migratory, lymphoid-tissue-resident conventional DCs and interferon-producing plasmacytoid DCs [247]. Flt3L levels were moderately increased in culture-positive sepsis compare to those without a confirmed infection. No other

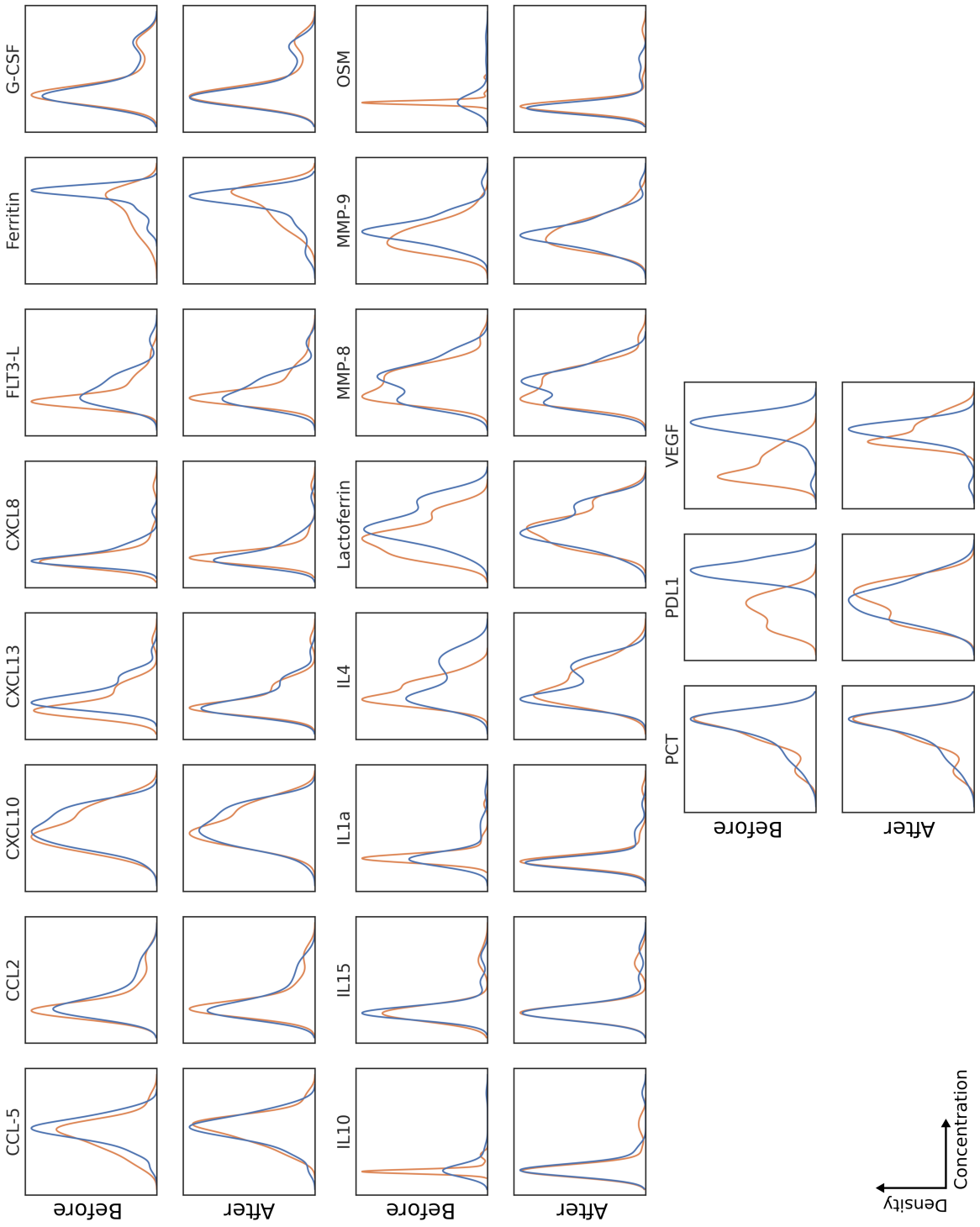

Figure 5.7: Distribution of biomarkers captured by Luminex multiplex assays before and after batch correction by $\log_2$ and z-score normalisation.

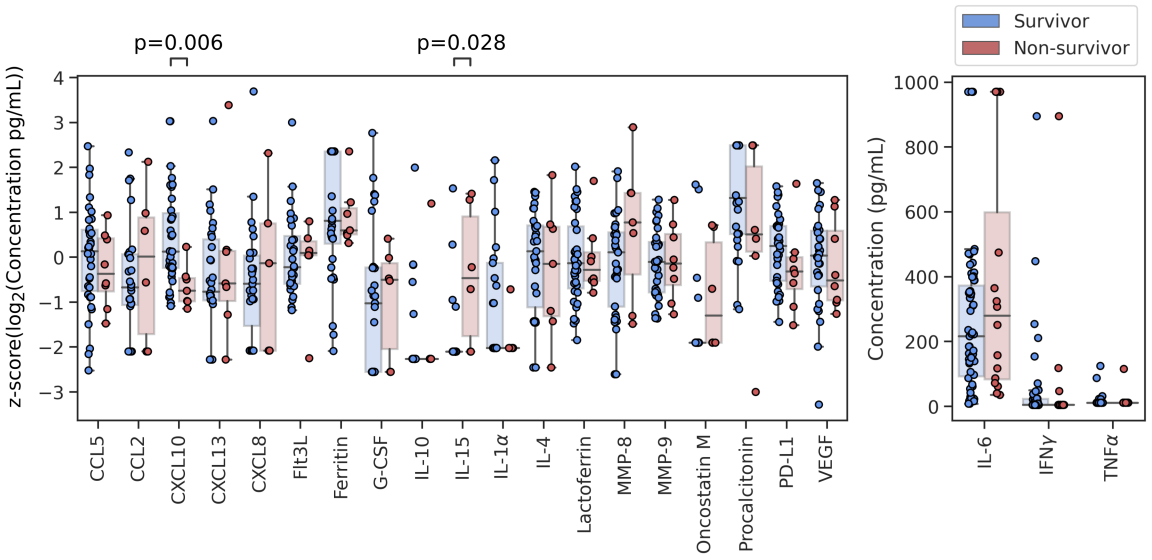

Figure 5.8: Concentration of soluble analytes in cell-free plasma isolated from sepsis patients, comparing survivors and non-survivors 30 days after sepsis diagnosis. Batch corrected concentration of analytes measured by Luminex™ multiplex assays (left panel) and concentrations measured by ELISA (right panel) are shown. Significance testing was performed using a two-tailed Mann-Whitney U test with correction for multiple comparisons made using the Benjamini–Hochberg procedure to control false discovery rate at an α of 0.05.

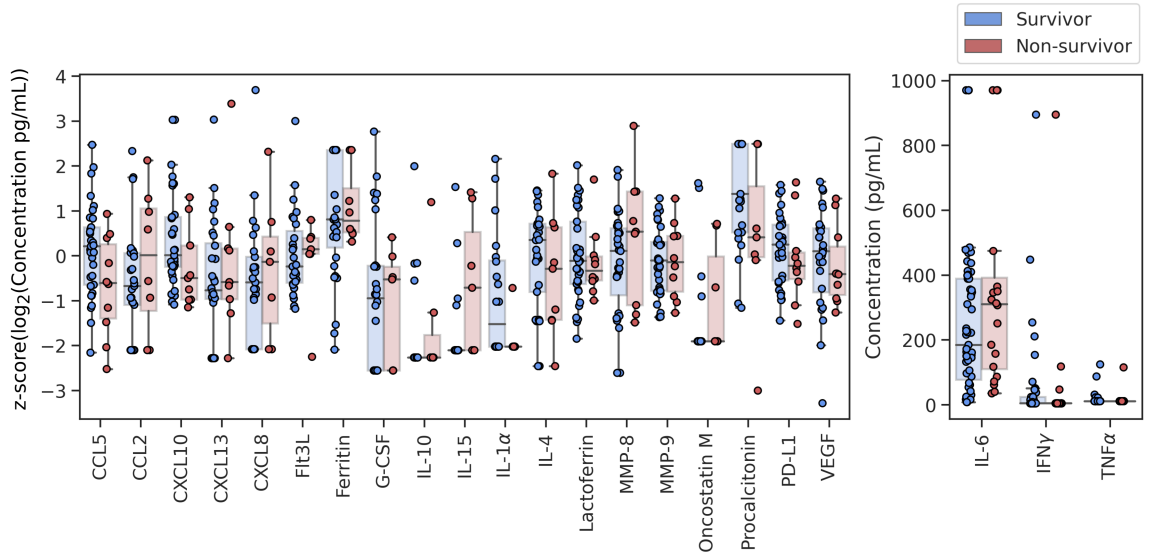

Figure 5.9: Concentration of soluble analytes in cell-free plasma isolated from sepsis patients, comparing survivors and non-survivors 90 days after sepsis diagnosis. Batch corrected concentration of analytes measured by Luminex™ multiplex assays (left panel) and concentrations measured by ELISA (right panel) are shown. Significance testing was performed using a two-tailed Mann-Whitney U test with corrections for multiple comparisons, with correction for multiple comparisons made using the Benjamini–Hochberg procedure to control false discovery rate at an α of 0.05.

analytes significantly differed between those with and without microbiologically confirmed infections.

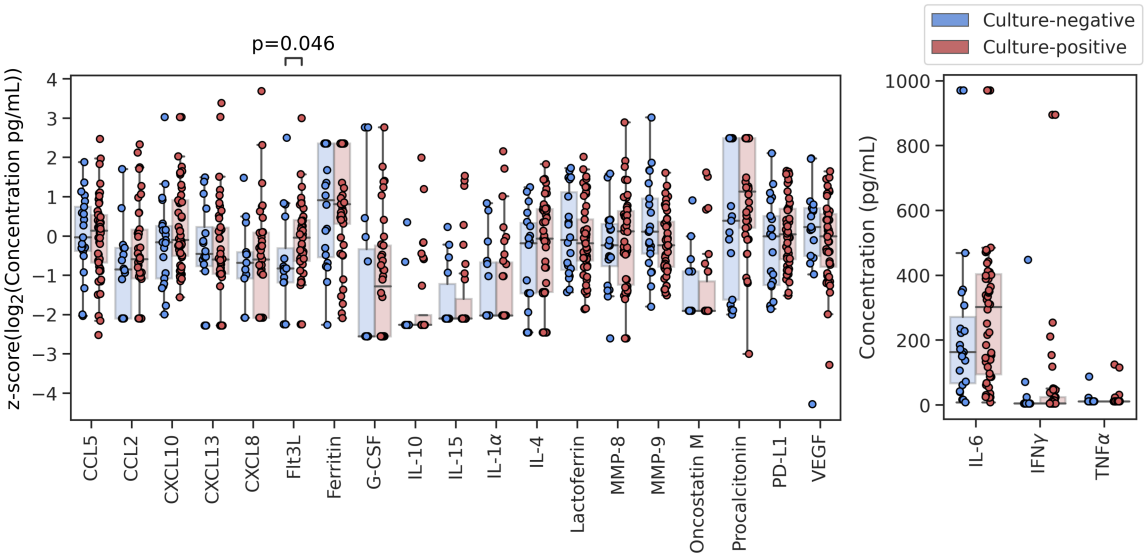

Figure 5.10: Concentration of soluble analytes in cell-free plasma isolated from sepsis patients, comparing those with and without a microbiologically confirmed infection. Batch corrected concentration of analytes measured by Luminex^{TM} multiplex assays (left panel) and concentrations measured by ELISA (right panel) are shown. Significance testing was performed using a two-tailed Mann-Whitney U test with corrections for multiple comparisons, with correction for multiple comparisons made using the Benjamini–Hochberg procedure to control false discovery rate at an α of 0.05.

When further sub-setting those with microbiologically confirmed infections into Gram-positive and Gram-negative causative pathogens (Figure 5.11), Ferritin was significantly decreased in Gram-negative compared to Gram-positive infections, Ferritin is produced by macrophages during infection in response to IL-1, and TNF-α nuclear factor kappa B (NFκB) activation [248]. Other trends included higher oncostatin M (OSM) levels and lower PCT levels in Gram-negative infections. However, neither observation was statistically significant (p-values were greater than 0.1 for both).

Unfortunately, the detection limit of the assays introduced 'floor and ceiling effects' that severely limited the analysis and made it difficult to draw conclusions about the analytes measured. Figure 5.12 shows the proportion of samples below, within, and above the detection limit of their respective assay. In an attempt to obtain the maximum information from the available data, an additional analysis was performed using the detection limits as thresholds to create binary variables. To answer the question as to whether the proportion of patients above or below these thresholds were significant, IL-6, ferritin, and PCT were

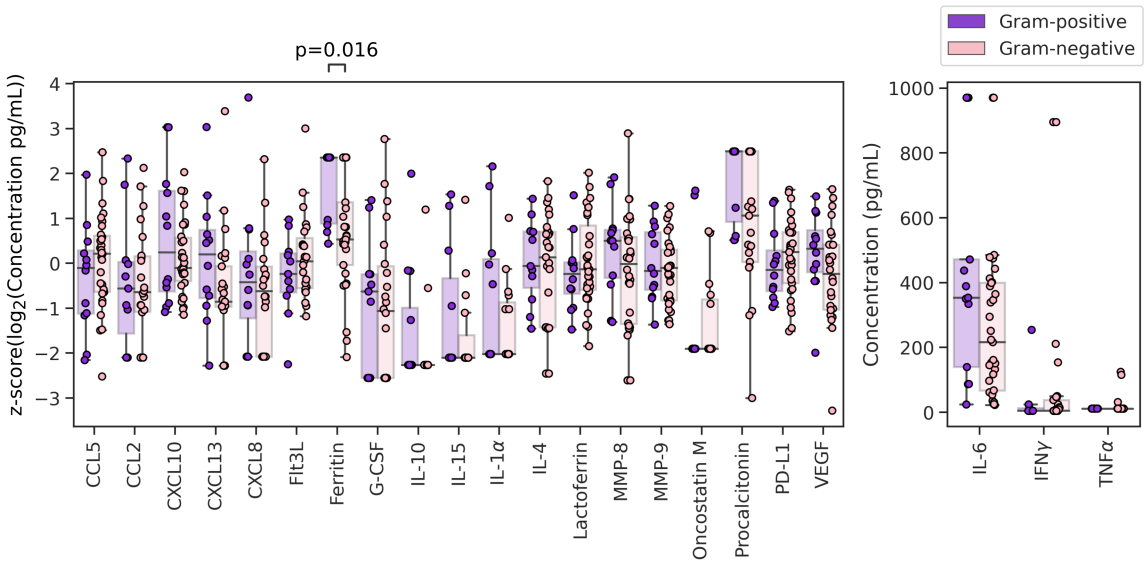

Figure 5.11: Concentration of soluble analytes in cell-free plasma isolated from sepsis patients, comparing those with a Gram-positive and Gram-negative infection, amongst those with a positive bacterial culture. Batch corrected concentration of analytes measured by Luminex™ multiplex assays (left panel) and concentrations measured by ELISA (right panel) are shown. Significance testing was performed using a two-tailed Mann-Whitney U test with corrections for multiple comparisons, with correction for multiple comparisons made using the Benjamini-Hochberg procedure to control false discovery rate at an α of 0.05.

separated into those below and above the upper bound of the detection limit, and IL-15, OSM, VEGF, IL-10, IL-1α, MMP-8, CXCL8, G-CSF, IL-4, CXCL13, CCL2, TNFα, Flt3L and IFNγ were separated into those below and above the lower bound of the detection limit. CXCL10, CCL5, Lactoferrin, MMP-9, and PD-L1 were excluded because more than 90% of samples were within the detectable range.

Figure 5.13 shows the odds ratio for mortality at 30 days after sepsis diagnosis (top left), mortality at 90 days after sepsis diagnosis (top right), odds of culture-negative sepsis (bottom left), and odds of a Gram-negative causative pathogen amongst those with a microbiologically confirmed infection (bottom right). Categories were compared for significance using Fisher's exact test and confidence intervals for odds ratios were calculated as described by Tenny and Hoffman [249]. Although an IL-6 concentration greater than 500 pg/ml, an IL-15 concentration greater than 6.3 pg/ml, and an OSM concentration greater than 369.7 pg/ml appeared to show a greater odds of 30-day mortality, once accounting for multiple comparisons with the Benjamini–Hochberg procedure at an α of 0.05, none of the analytes had a statistically significant odds ratio.

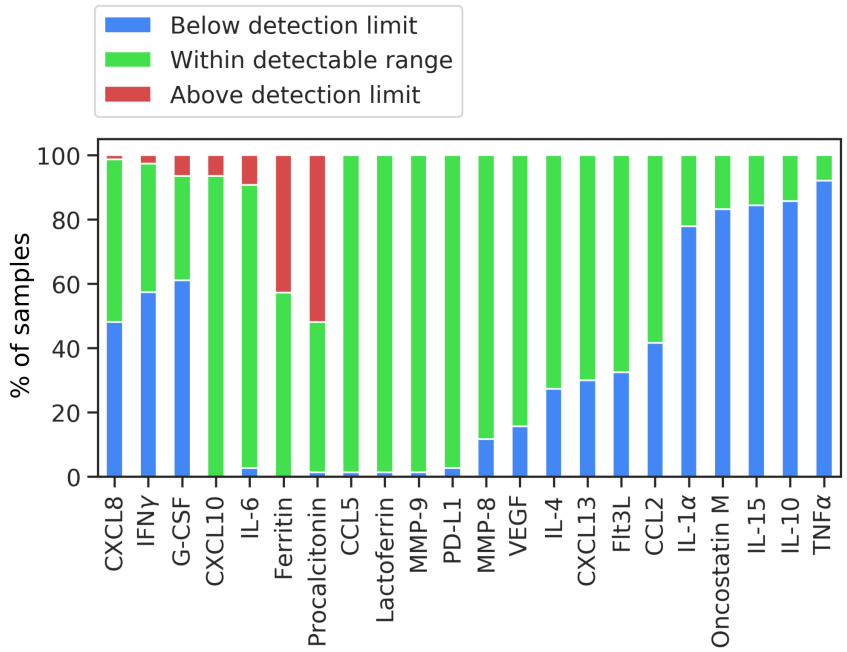

Figure 5.12: Proportion of samples above, below, or within the detectable range of the assay used for the measurement of an analyte.

5.3.4 Immune cell profiling in acute severe sepsis patients demonstrates phenotypes that correlate with mortality and causative pathogen.

The PBMC and red cell-free fraction of whole blood were analysed by flow cytometry as described in Materials & Methods section 2.1.5 and 2.1.6. Figures 2.2 and 2.3 show the gating strategies applied using autonomous gates to identify starting populations for downstream clustering analysis. Autonomous gates identified the major cell subsets of monocytes, neutrophils, and T lymphocytes, as well as the main subsets of T cells of interest in this study: $CD4^+$ and $CD8^+$ conventional T cell subsets, mucosal-associated invariant T cells (MAIT cells), and $V\delta2^+$ $\gamma\delta$ T cells.

The first observation was a significant reduction in T cells as a percentage of PBMCs in non-survivors compared to survivors at 30 days following enrolment (Figure 5.14) and 90 days following enrolment (Figure 5.15). There was no significant difference in monocytes and neutrophils as a percentage of leukocytes between survivors and non-survivors, although some extreme values were observed; for example, in two non-survivors, more than 20% of leukocytes were monocytes. Upon further investigation, these patients appeared to be neutropenic, with less than 30% of their leukocytes consisting of neutrophils. Amongst the

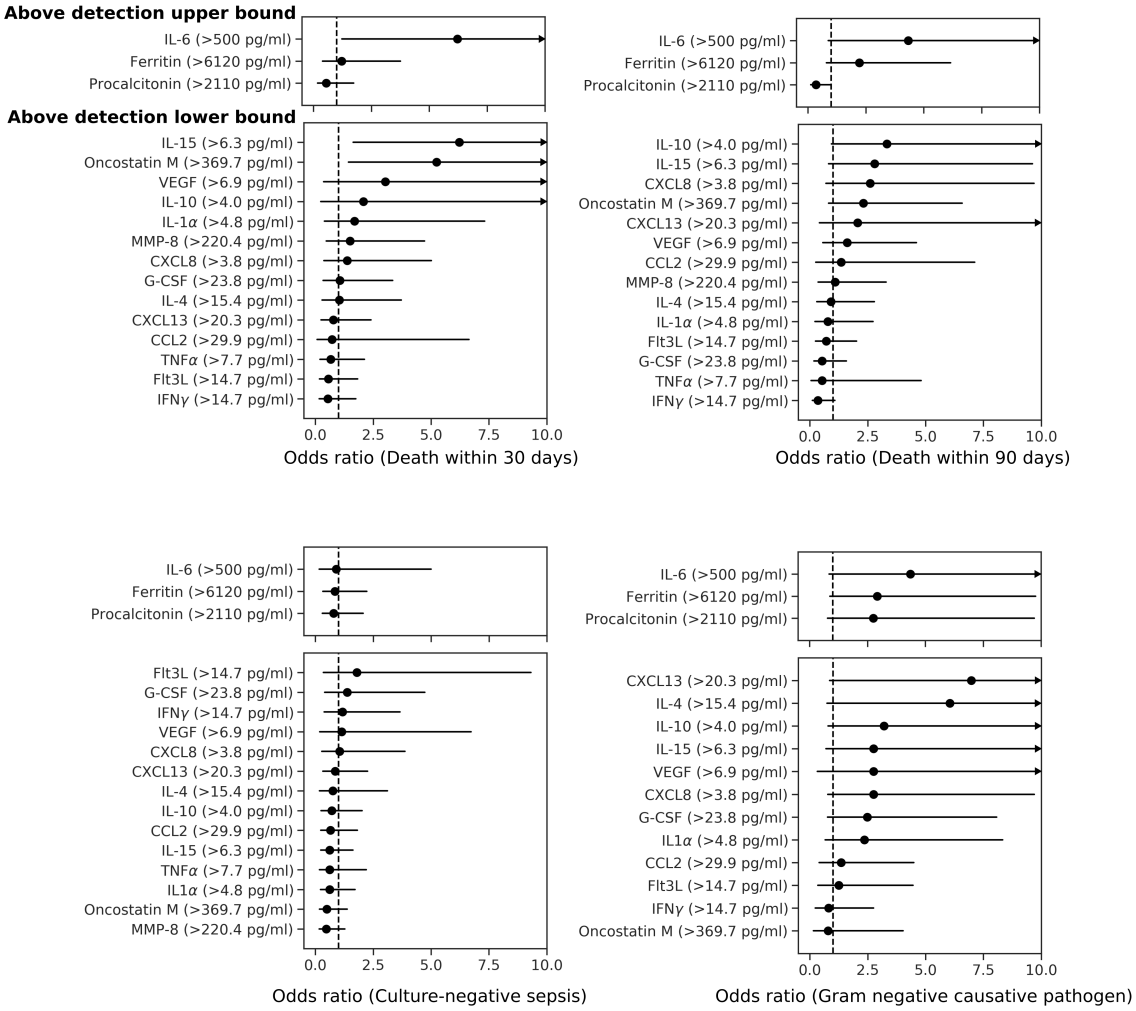

Figure 5.13: Odds ratios for death within 30 days of enrolement date (top left), death within 90 days of enrolement date (top right), culture-negative sepsis (bottom left), and Gram-negative causative pathogen (bottom right). For the analytes IL-6, ferritin, and procalcitonin, patients were grouped into those above and below the upper detection limit of the assay. Whereas IL-15, oncostatin M, VEGF, IL-10, IL-1α, MMP-8, CXCL8, G-CSF, IL-4, CXCL13, CCL2, TNFα, Flt3L and IFNγ were separated into those below and above the lower bound of the detection limit. Comparisons between groups were tested for significance using Fisher's exact test and corrected for multiple comparisons with Benjamini–Hochberg procedure at an α of 0.05. 95% confidence intervals for odds ratios were approximated as previously described by Tenny and Hoffman [249]

T cell subsets, a trend in the reduction of $CD8^+$ T cells was visible amongst non-survivors, but to a greater extent when observing 30 days post enrolment compared to 90 days post enrolment. The observations could suggest that a reduction of the proportion of $CD8^+$ T cells is an indicator of early mortality, but it could also be an effect of a more balanced ratio of survivors and non-survivors for the 90-day mortality endpoint. No significant difference was observed amongst subsets of unconventional T cells.

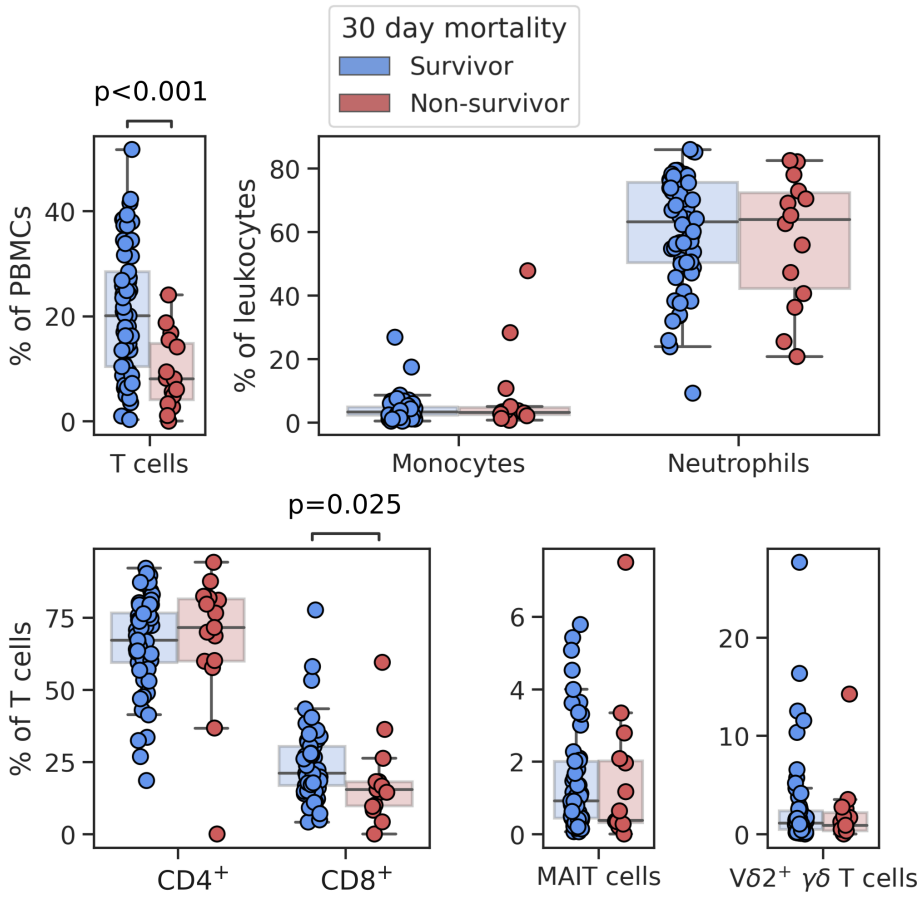

Figure 5.14: Comparison of the proportion of T cells, monocytes and neutrophils, and conventional and unconventional T cell subsets in survivors and non-survivors of sepsis 30 days after sepsis diagnosis. P-values were generated using a two-tailed Mann-Whitney U test with Bonferroni-Holm correction for multiple comparisons.

The primary immune compartments of T cells, monocytes, and neutrophils, as well as the subsets of T cells, were comparable between those with and without a microbiologically confirmed infection (Figure 5.16) with no significant differences seen. In contrast, when comparing Gram-positive and Gram-negative causative pathogens in those with confirmed infection, there was a significant increase in the proportion of neutrophils as a percentage of leukocytes in Gram-negative infections and a significant decrease in MAIT and V$\delta 2^+$ $\gamma\delta$ T cells as a percentage of T cells in Gram-negative infections (Figure 5.17).

After identifying the main subsets by autonomous gating, the discovery of differentiating subsets was driven by unsupervised clustering. However, recruitment for the ILTIS study spanned years and practical limitations around the desire to capture the phenotype of monocytes and neutrophils required that each patient sample be measured independently, intro-

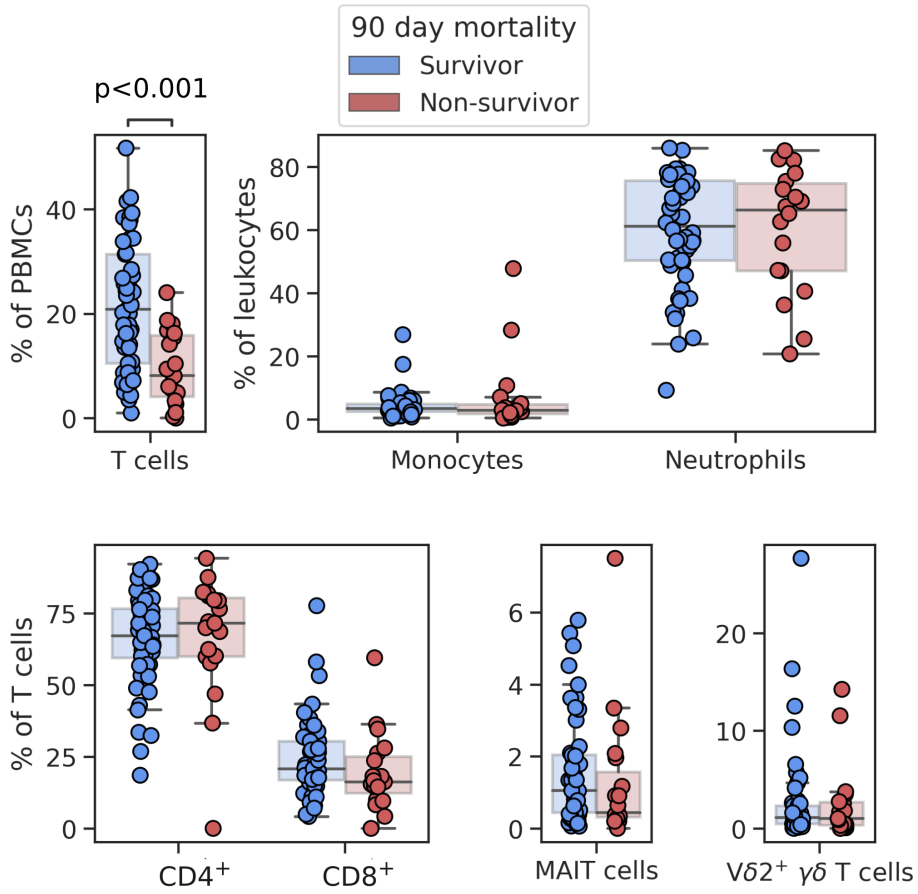

Figure 5.15: Comparison of the proportion of T cells, monocytes and neutrophils, and conventional and unconventional T cell subsets in survivors and non-survivors of sepsis 90 days after sepsis diagnosis. P-values were generated using a two-tailed Mann-Whitney U test with Bonferroni-Holm correction for multiple comparisons.

ducing batch effects. As with the PERIT-PD study discussed in Chapter 3.1, the Harmony algorithm was applied to ILTIS data using CytoPy version 3.0 to align samples whilst reducing the risk of losing biological information. A suitable starting population was chosen depending on the staining panel (*i.e.* T cells, monocytes, or neutrophils) and a sample of 30,000 cells taken from each patient. A sample size of 30,000 was chosen to limit the risk of undersampling rare cell populations whilst reducing the computational burden of subsequent procedures.

Figure 5.18 shows the outcome of batch correction using Harmony for T cell staining for markers of activation (top row) and memory subsets (second from top), monocyte staining (second from bottom) and staining of neutrophils (bottom row). Of note is the reduction of regions in UMAP embedded space dominated by individual patient samples. Figure 5.19

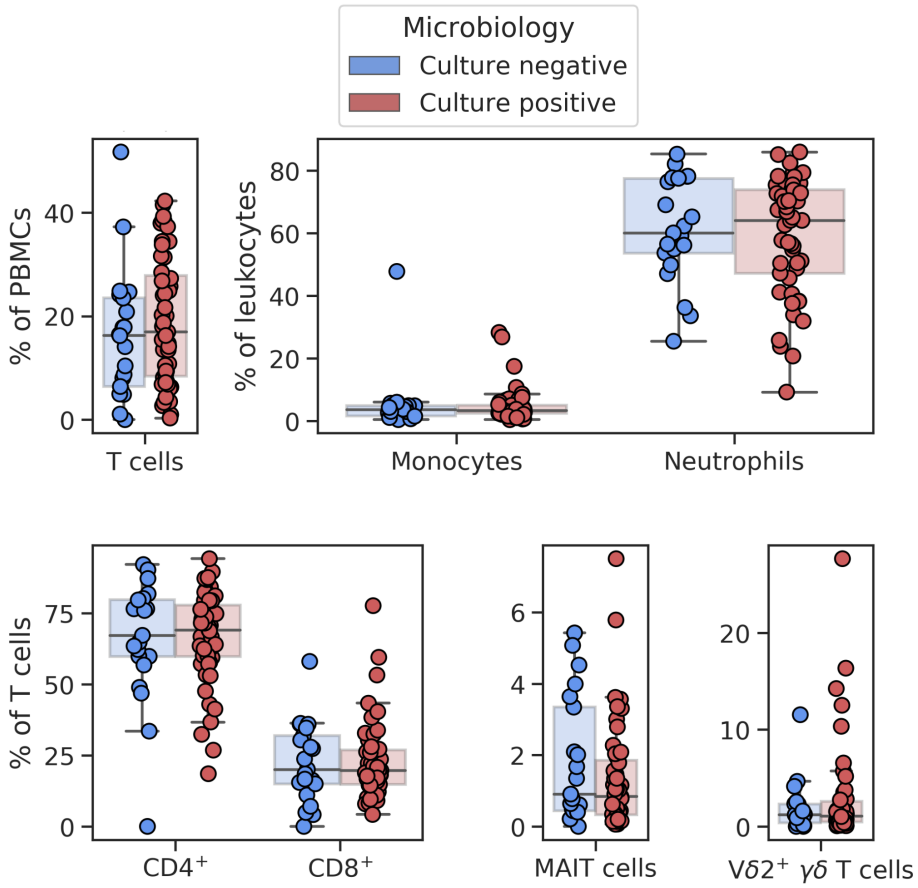

Figure 5.16: Comparison of the proportion of T cells, monocytes and neutrophils, and conventional and unconventional T cell subsets in sepsis patients with and without a microbiologically confirmed infection. P-values were generated using a two-tailed Mann-Whitney U test with Bonferroni-Holm correction for multiple comparisons.

shows the shift in the distribution of the local inverse Simpson index, a measure of the diversity of batches seen in the neighbourhood surrounding an individual cell (a detailed explanation was given in Chapter 3.3.2 where the Harmony algorithm was introduced). Successful batch correction should demonstrate a shift in this distribution as cell neighbourhoods become more diverse, evidenced by a greater representation of batches within each neighbourhood. Additionally, Figure 5.20 demonstrates that major linage markers in T cells remain distinct following batch correction.

As an additional validation to verify that Harmony data integration was retaining biologically distinct cell populations within the sepsis T cell subsets, manually gated $CD4^+$, $CD8^+$, $V\delta2^+$ $\gamma\delta$ T cell, and MAIT cell populations were compared before and after Harmony correction (Figure 5.21). The proportion of cell populations across the 77 sepsis patients was consistent

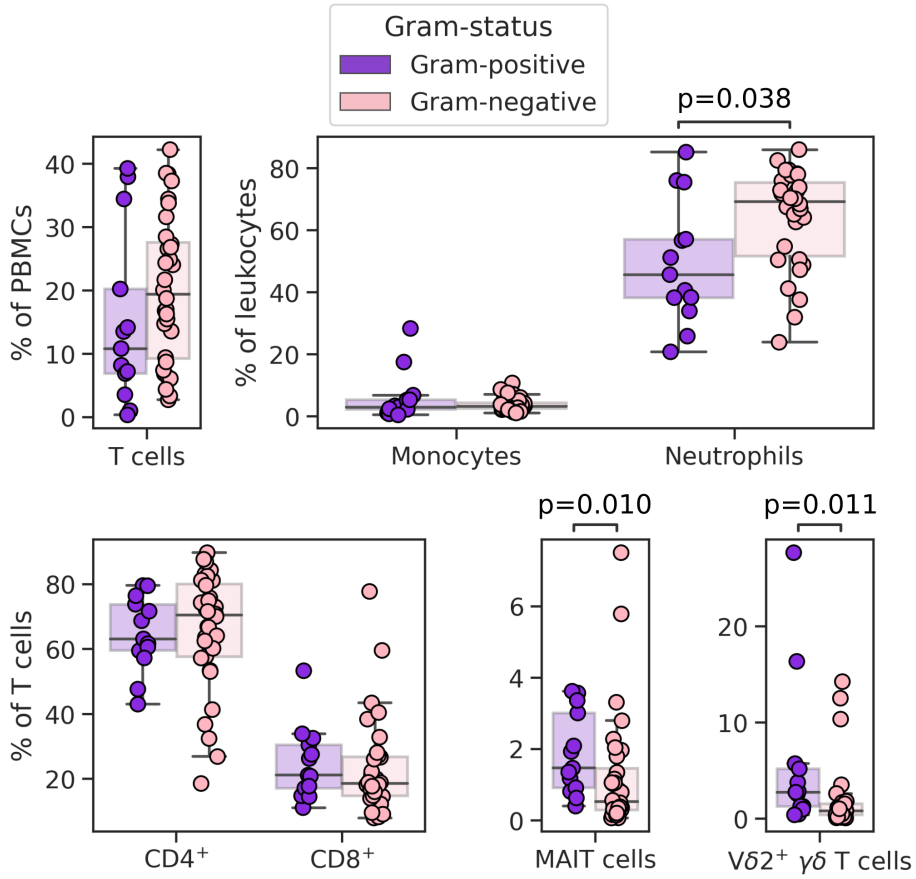

Figure 5.17: Comparison of the proportion of T cells, monocytes and neutrophils, and conventional and unconventional T cell subsets in sepsis patients with a Gram-positive or Gram-negative infection, where sepsis was microbiologically confirmed. P-values were generated using a two-tailed Mann-Whitney U test with Bonferroni-Holm correction for multiple comparisons.

before and after batch correction, indicating that Harmony batch correction was a reliable methodology that did not distort the data through over-integration.

MAIT cells and V$\delta 2^+$ $\gamma\delta$ T cells were of particular interest given their ability to recognise bacterial metabolites, and therefore their potential use as a biomarker of infection [28, 250], therefore, MAIT cells and V$\delta 2^+$ $\gamma\delta$ T cells were studied individually for activation and memory subsets. Down-sampling was unnecessary, given their respective population size, and all available events from each subject were included during batch correction and subsequent downstream analysis. Figure 5.22 shows the outcome of batch effect correction using the Harmony algorithm for both V$\delta 2^+$ $\gamma\delta$ T cells (Figure 5.22; first and second row) and MAIT cells (Figure 5.22; third and fourth row).

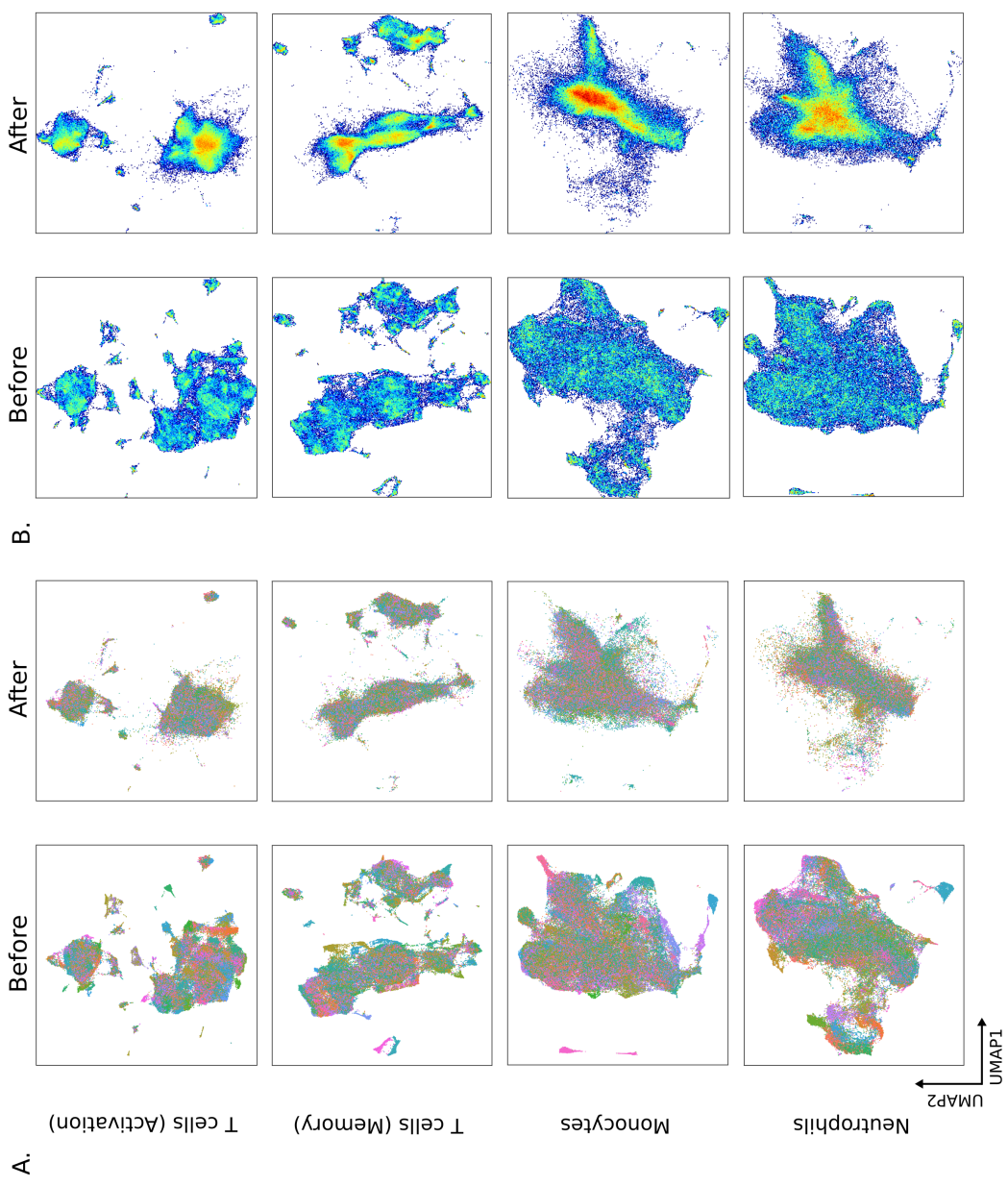

Figure 5.18: Before and after correction of batch effects with the Harmony algorithm as applied to single cell flow cytometry data from the ILTIS study. Batch correction was applied to T cells stained for activated subsets, T cells stained for memory subsets, monocytes, and neutrophils. The left column (A) shows UMAP scatterplots with each batch coloured differently, whereas the right hand column (B) shows a density plot where greater intensity of colour represents a more densely populated neighbourhood of cells.

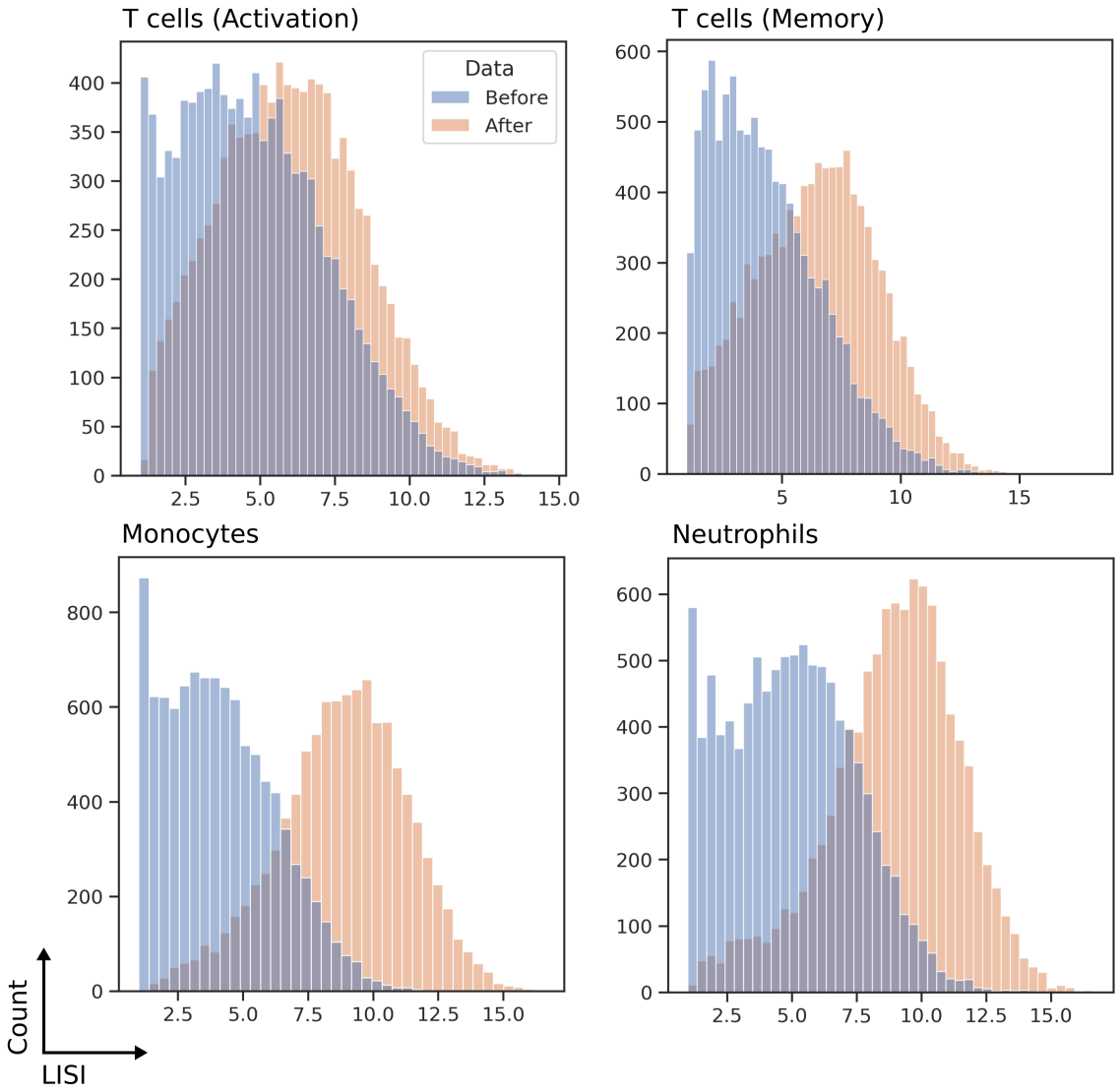

Figure 5.19: The distribution of Local Inverse Simpson Index (LISI) for a sample of 10000 events before (blue) and after (orange) the Harmony algorithm was applied to correct batch effect.

GeoWaVe ensemble clustering (introduced in Chapter 4) was performed on batch-corrected data for T cells, monocytes, and neutrophils. Ensembles were informed using multiple clustering algorithms popular for analysing cytometry data, providing diverse input for ensembles and preventing biased analysis driven by a single method. The FlowSOM [130], Phenograph [131], and SPADE [111] algorithms were chosen as input for ensembles due to their popularity in the cytometry literature. Additionally, K-Means and FlowSOM clustering of PHATE embeddings were included to offer the opportunity for improved clustering performance gained from a dimension reduction technique.

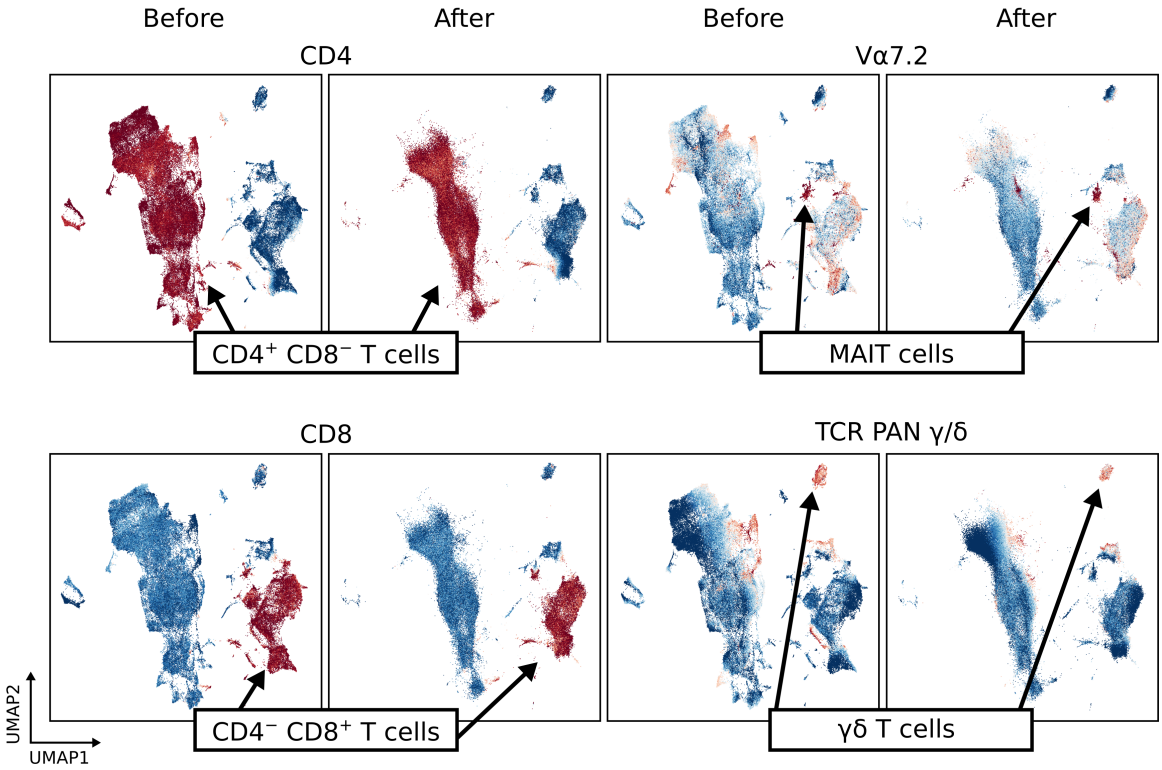

Figure 5.20: UMAP scatterplots show the preservation of population structure before and after the application of the Harmony algorithm to T cells acquired in many batches. Each pair shows the fluorescence intensity (red for high expression) of CD4 (top left), Vα7.2 (top right), CD8 (bottom left), and TCR Pan $\gamma\delta$ (bottom right).

Monocytes are one of the main effectors of innate immunity, with the capacity to ingest microbes, present antigens to prime T cells, and produce inflammatory mediators [17]. Figure 5.23 shows GeoWaVe consensus clusters identified amongst monocytes. The heatmap shows hyperbolic arcsine transformed fluorescence intensity for markers of activation and adhesion for each of the consensus clusters. The same clusters are shown in embedded UMAP space in the accompanying scatterplot.

Clusters 0 and 3 could be differentiated from clusters 2, 1, and 4 based on HLA-DR expression, with the former exhibiting higher expression of HLA-DR. Cluster 3 was the smaller of the two HLA-DR^{hi} clusters and was distinct from cluster 0 with a higher expression of CD62L and lower expression of CD64. The largest cluster was cluster 1, and when compared with the other HLA-DR^{lo} clusters (2 and 4) showed a higher expression of CD40 and a lower expression of CD62L. Most monocytes exhibited subdued expression of HLA-DR and the co-stimulatory molecule CD86, a phenotype known to be prevalent in sepsis [251, 253,

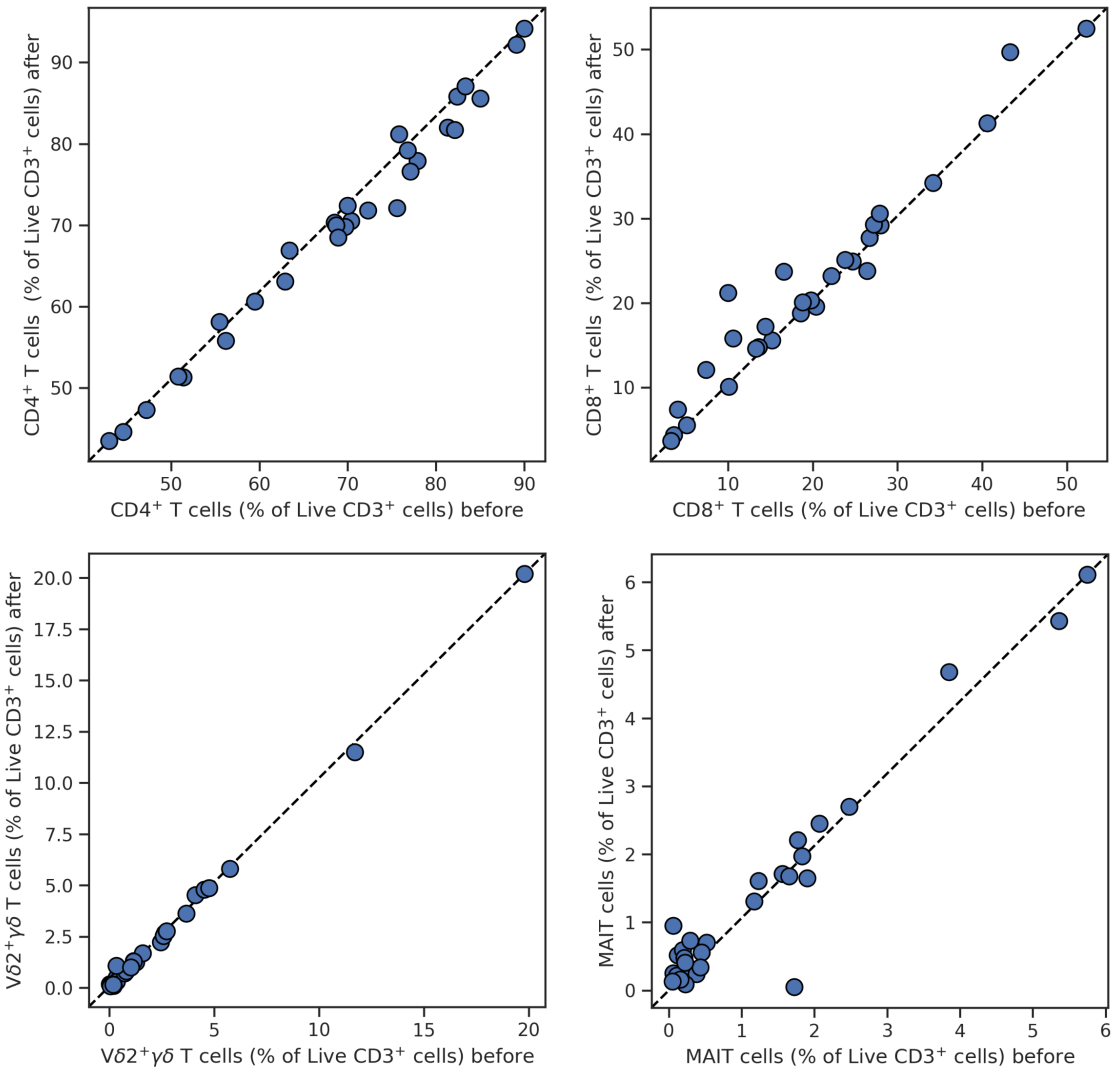

Figure 5.21: Proportion of manually gated T cell populations (as a percentage of $CD3^+$ cells) before (x-axis) and after (y-axis) Harmony batch correction. $CD4^+$ (top left), $CD8^+$ (top right), $V\delta2^+$ $\gamma\delta$ T cells (bottom left), and MAIT cell (bottom right) populations are shown.

36, 252, 70]. Comparisons between patient subsets (Figure 5.23, left panel) failed to demonstrate any significant difference in monocyte clusters between survivors and non-survivors, microbiologically confirmed infections or Gram-negative vs Gram-positive infections.

It was not apparent that survivors and non-survivors could be differentiated on monocyte clusters, even though some clusters could be grouped into those with an activated profile of higher CD40, HLA-DR, and CD86 expression (clusters 0 and 3) versus those with lower expression of these markers (clusters 1, 2, and 4). Despite this, the mean fluorescence intensity of batch-corrected monocytes showed a decrease in HLA-DR expression in non-survivors

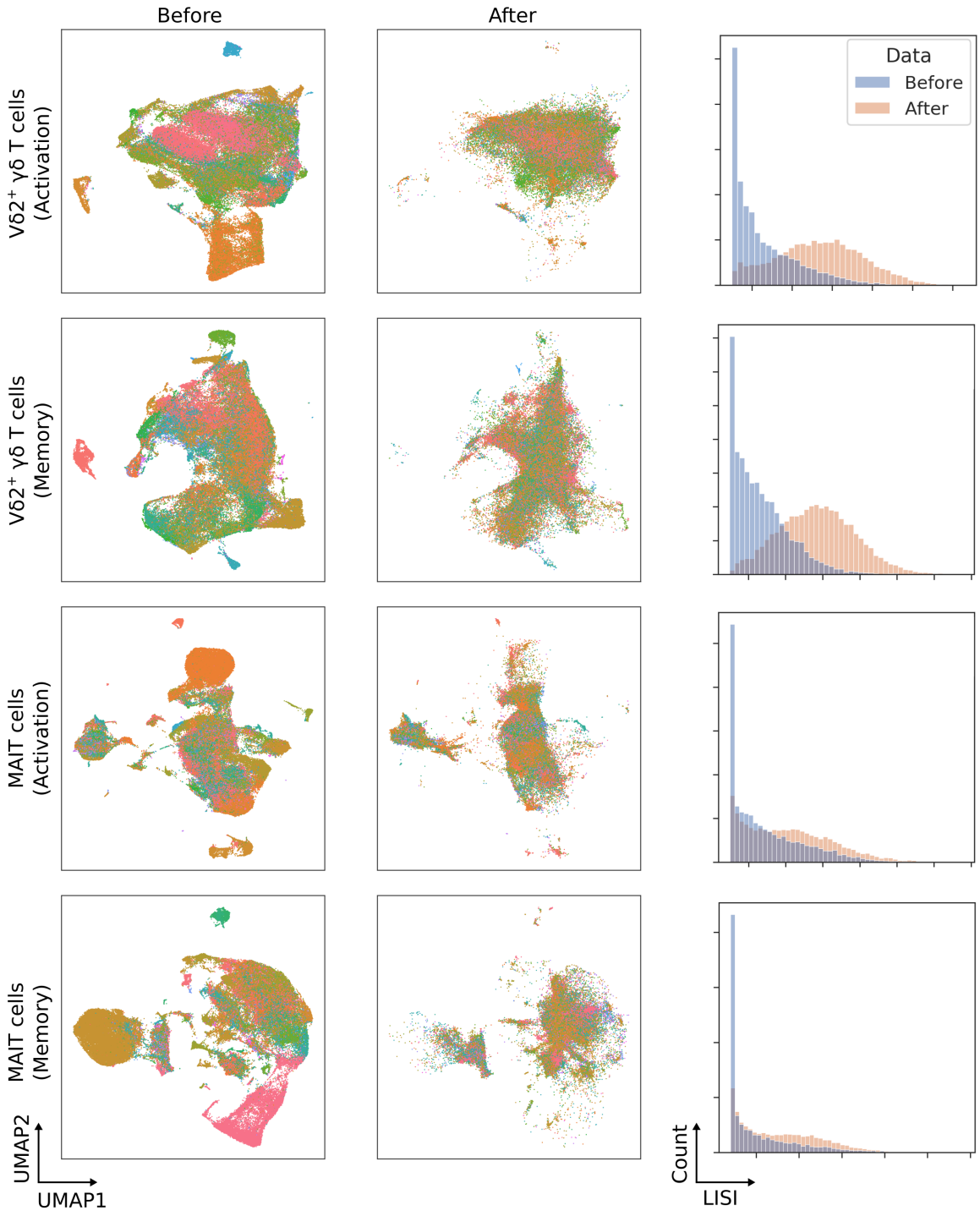

Figure 5.22: Before and after correction of batch effects with the Harmony algorithm, applied to V$\delta 2^+$ $\gamma\delta$ T cells stained for activated subsets and stained for memory subsets, and MAIT cells stained for activation subsets and memory subsets. The left UMAP scatterplots show individual batches coloured separately and the distribution before and after correction. The right histograms show the distribution of Local Inverse Simpson Index (LISI) for a sample of 10000 events before (blue) and after (orange) the Harmony algorithm was applied.

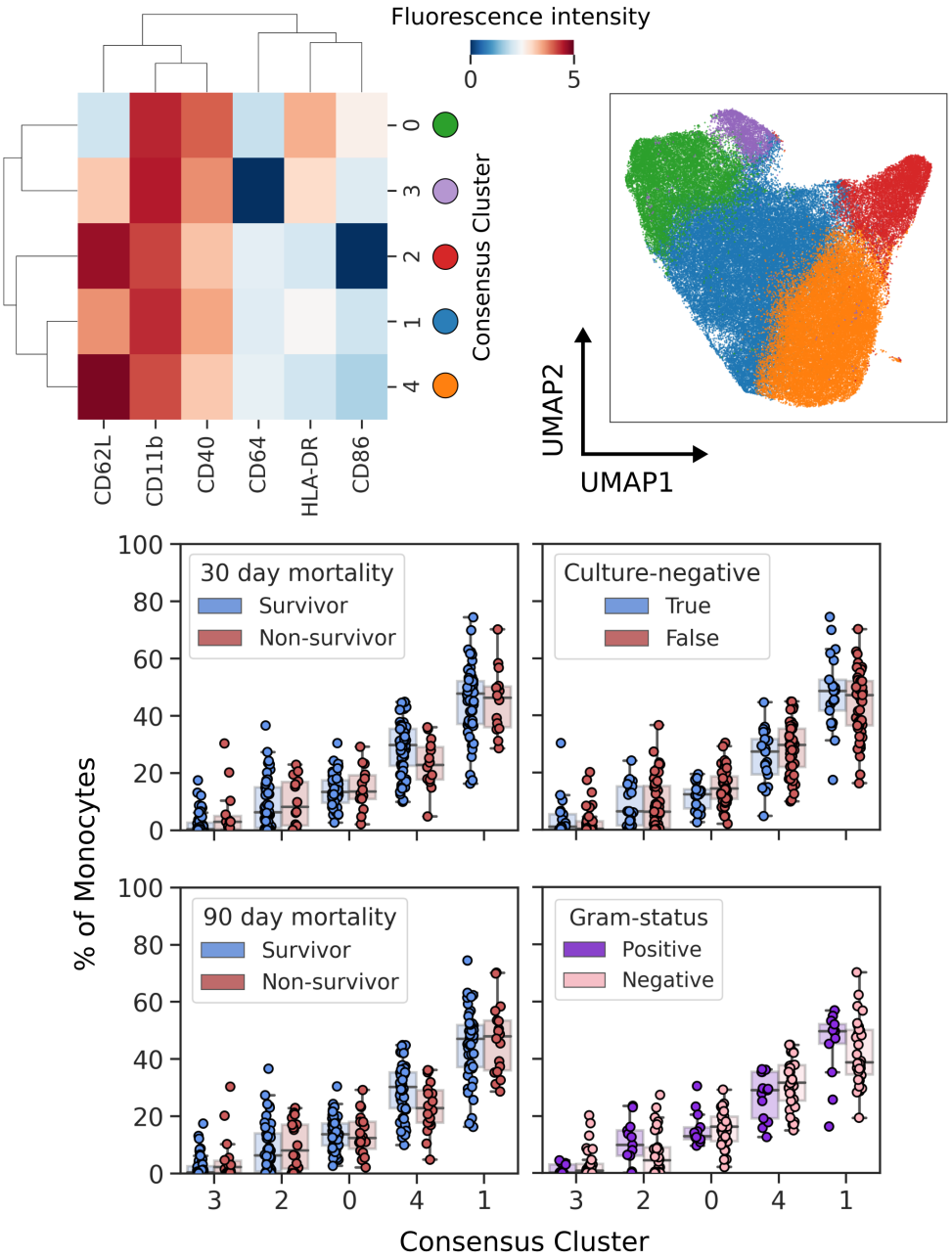

Figure 5.23: GeoWaVe ensemble clustering of monocytes. The heatmap and accompanying UMAP scatterplot (right) show the identified consensus clusters and their expression profile. Fluorescence intensity is shown as hyperbolic arcsine with a cofactor of 150. The proportion of each cluster as a percentage of total number of monocytes (left) is shown, with comparisons between survivors and non-survivors at 30 and 90 days after sepsis diagnosis. Additionally, comparisons are shown for those with and without microbiologically confirmed infection (culture negative) and, for those with a confirmed infection, the difference between those with a Gram-positive versus a Gram-negative causative pathogen.

compared to survivors at both 30 and 90 days post sepsis diagnosis (Figure 5.24). Interestingly, a trend was visible for the cell adhesion molecule CD62L (L-selectin) with increasing expression of this marker in non-survivors. However, this was not reflected in cluster 0 despite CD62L being a defining feature of this cluster.

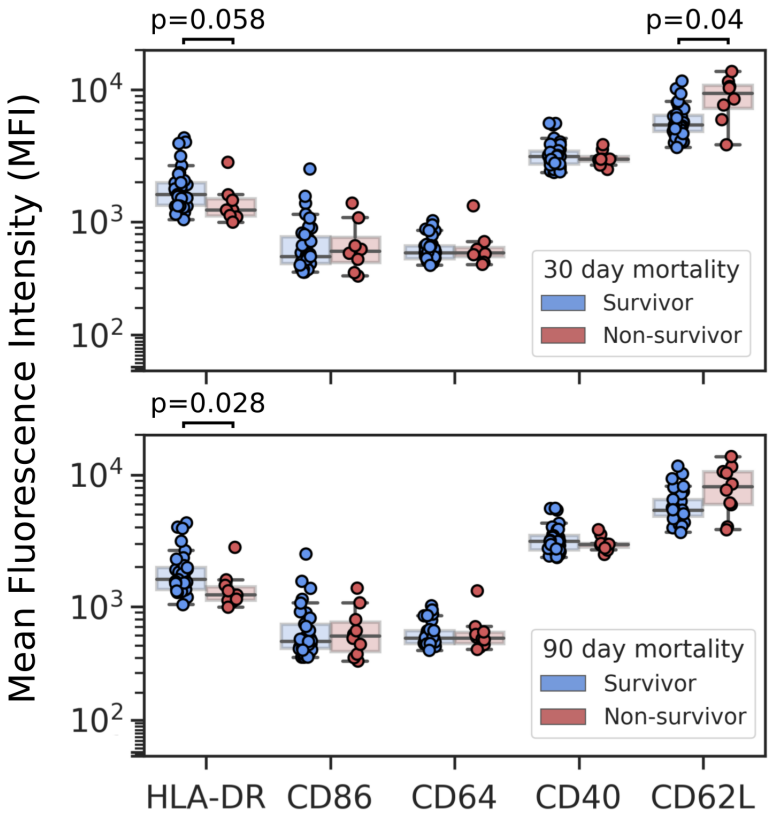

Figure 5.24: The mean fluorescence intensity (MFI) of HLA-DR, CD86, CD46, CD40 and CD62L on monocytes in sepsis.. Comparisons between survivors and non-survivors 30 (top) and 90 (bottom) days following a diagnosis of sepsis are shown. P-values were generated using a two-tailed Mann-Whitney U test with Bonferroni-Holm correction for multiple comparisons.

Unlike with the monocytes, GeoWaVe consensus clustering of neutrophils revealed simple clusters almost entirely differentiable on the expression of CD62L alone (Figure 5.25). CD62L expression formed a gradient across the four clusters identified, with the smallest cluster, cluster 1, having the lowest expression, followed by cluster 3, then cluster 2, and finally, the largest cluster, cluster 0, with the greatest expression of the cell adhesion molecule. Cluster 1 represented a smaller proportion of all neutrophils in survivors compared to non-survivors and was significantly different when comparing 90-day mortality.

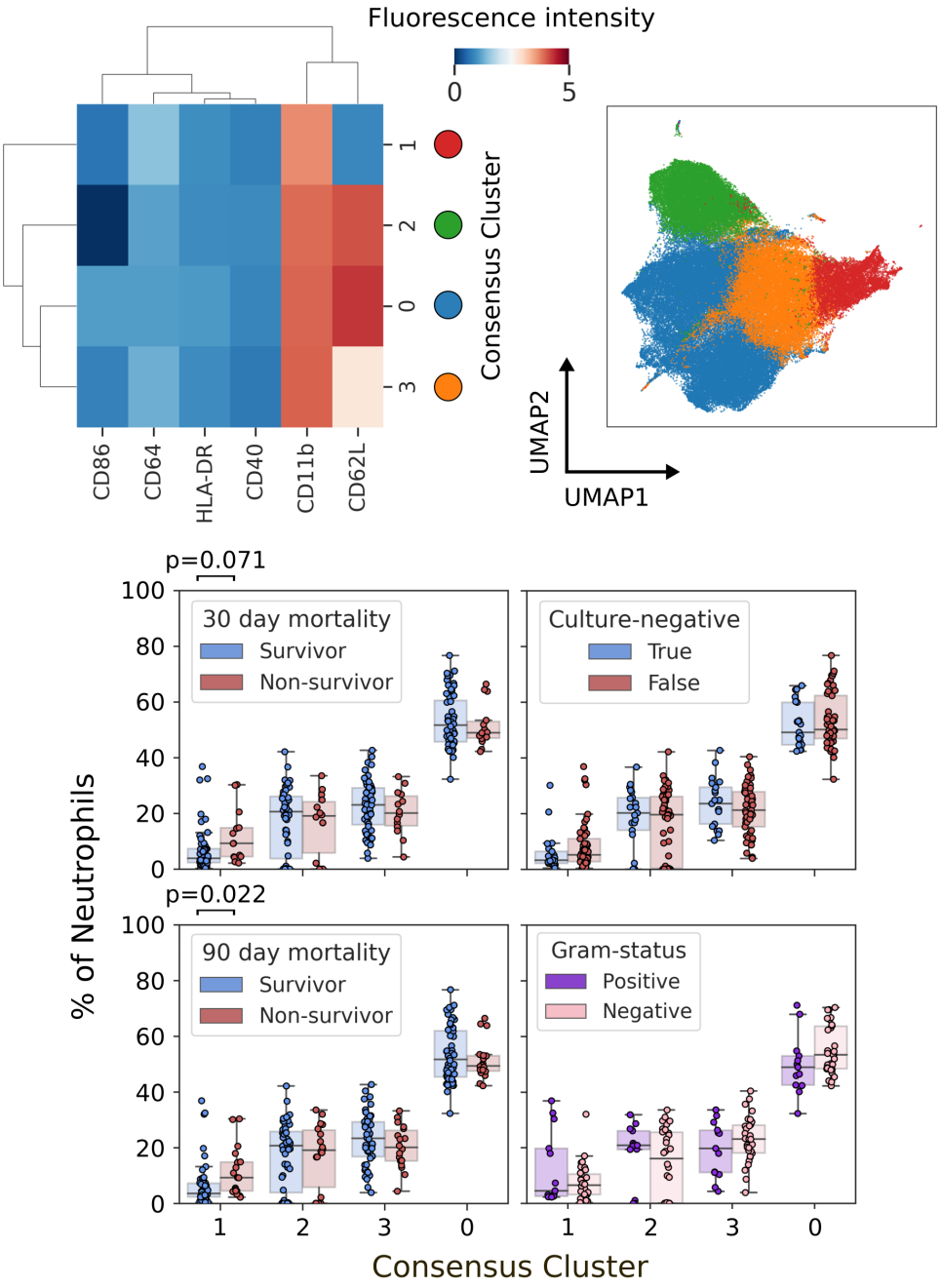

Figure 5.25: GeoWaVe ensemble clustering of neutrophils in sepsis. The heatmap and accompanying UMAP scatterplot (right) show the identified consensus clusters and their expression profile. Fluorescence intensity is shown as hyperbolic arcsine with a cofactor of 150. The proportion of each cluster as a percentage of total number of neutrophils (left) is shown, with comparisons between surviviors and non-survivors at 30 and 90 days post diagnosis with sepsis. Additionally, comparisons are shown for those with and without microbiologically confirmed infection (culture negative) and, for those with a confirmed infection, the difference between those with a Gram-positive versus a Gram-negative causative pathogen. P-values were generated using a two-tailed Mann-Whitney U test with Bonferroni-Holm correction for multiple comparisons.

	CD45RA	CD27	CCR7	CD57	Subtype
$CD4^+$					
Cluster 0	↓	↑	↑	↓	Central Memory [255]
Cluster 2	↓	↓	↓	↑	Effector Memory [255]
Cluster 4	↑	↑	↑	↓	Naïve [255]
Cluster 14	↑	↑	↑	↓	Naïve [255]
$CD8^+$					
Cluster 1	↑	↑	↑	↓	Naïve [256]
Cluster 3	↑	↓	↓	↑	Terminally differentiated effector memory [256]

Table 5.4: Summary of GeoWaVe T cell cluster phenotypes stained for differentiating memory subtypes. Arrows represent the expression of a cell surface marker relative to the expression amongst all other T cells. The complete expression profile is shown in the heatmap of Figure 5.26.

Figure 5.26 shows the GeoWaVe consensus clusters for T cells, stained for the identification of memory subsets (Figure 5.26 left heatmap and UMAP scatterplot) and activated subsets (Figure 5.26 right heatmap and UMAP scatterplot). In both staining panels, MAIT cells and $V\delta2^+$ $\gamma\delta$ T cells were identified; clusters 6 and 9, respectively, in the memory staining panel, and clusters 17 and 11, respectively, in the activation staining panel.

Within the panel for memory subsets, two clusters were identified as $CD8^+$ T cells (clusters 1 and 3) and four clusters as $CD4^+$ T cells (clusters 0, 2, 4, and 14). However, the relatively small cluster 4 had an interesting phenotype with high expression for $V\alpha7.2$ and CD161, meaning they might constitute a $CD4^+$ MAIT cell population similar to those described by Gherardin *et al.* [254]. A summary of the expression profiles shown in the left heatmap of Figure 5.26 is given in Table 5.4. Amongst the $CD8^+$ T cells, naïve and terminally differentiated effector cells were identified, whereas $CD4^+$ T cell clusters consisted of naïve, central memory, and effector memory T cells.

Within the activated subsets, six clusters were identified as $CD8^+$ T cells: 5, 16, 7, 1, 8, and 10. These subsets were distinguished primarily on the expression of CD69, CD25, and CXCR3. The clusters 5, 16, and 10 had high expression of CD69, an early activation marker induced by T cell receptor signalling [17]. Cluster 10 differed by high expression of the chemokine receptor CXCR3. The binding of the CXCR3 ligands CXCL9 and CXCL10 recruits cytotoxic CD8 T cells to the site of inflammation to coordinate the cell-mediated killing of intracellular pathogens [17]. Another six clusters were identified as $CD4^+$ T cells: 6, 2, 4, 0, 13, and 14, and CD69, HLA-DR, and CD161 expression appeared to be the

defining markers. The majority of T cells belonged to the inactive CD4 T cells cluster 0. The smaller yet still prevalent cluster 6 displayed an activated phenotype with high expression of HLA-DR and CD25. The second largest CD4 cluster was cluster 4, with low expression of activation markers but high expression of CD161, a hallmark for IL-17 producing cells [257].

T cell clusters 5 and 10 amongst the memory subsets had a double negative ($CD4^-$ $CD8^-$) phenotype and could not be fully characterised with the chosen staining panels. Two double negative clusters (other than V$\delta 2^+$ $\gamma\delta$ T cells) were identified amongst the activation subsets as well (cluster 3 and 9). The populations could constitute a natural killer T cell or double negative regulatory T cell population [258], however, further work is needed to fully characterise these populations.

The major subsets are outlined in the UMAP plots and were combined to provide proportions as a percentage of T cells (Figure 5.27A). Identifying major subsets provided additional validation and could be compared to the findings from autonomous gating as shown in Figures 5.14-5.17. The clustering results confirm observations from autonomous gating: $CD8^+$ T cells were significantly reduced in non-survivors 30 days after sepsis diagnosis, and MAIT cells and V$\delta 2^+$ $\gamma\delta$ T cells were reduced in Gram-negative infections compared to Gram-positive infections.

When observing memory subsets, the predominant cluster of $CD4^+$ T cells was a central memory phenotype (cluster 0). In contrast, similar proportions of TEMRA and naive phenotypes existed for $CD8^+$ T cells (clusters 3 and 1, respectively). For activated subsets, in both $CD4^+$ T cells and $CD8^+$ T cells, the dominant clusters were inactive populations (cluster 0 for $CD4^+$ T cells and cluster 1 for $CD8^+$ T cells). Comparisons between patient groups were made for each of the $CD4^+$, $CD8^+$, and smaller undefined T cell clusters as a percentage of total T cells (Figure 5.27 B and C), but no significant differences were observed. When comparisons were made within the groups of $CD4^+$ and $CD8^+$ clusters, however, $CD8^+$ T cells with a TEMRA phenotype (cluster 3) were decreased in culture-negative sepsis compared to those with an identified pathogen, whereas naive $CD8^+$ T cells (cluster 1) were increased (Figure 5.28). The same trend was observed when comparing Gram-positive versus Gram-negative causative pathogens, although this difference was not statistically significant.

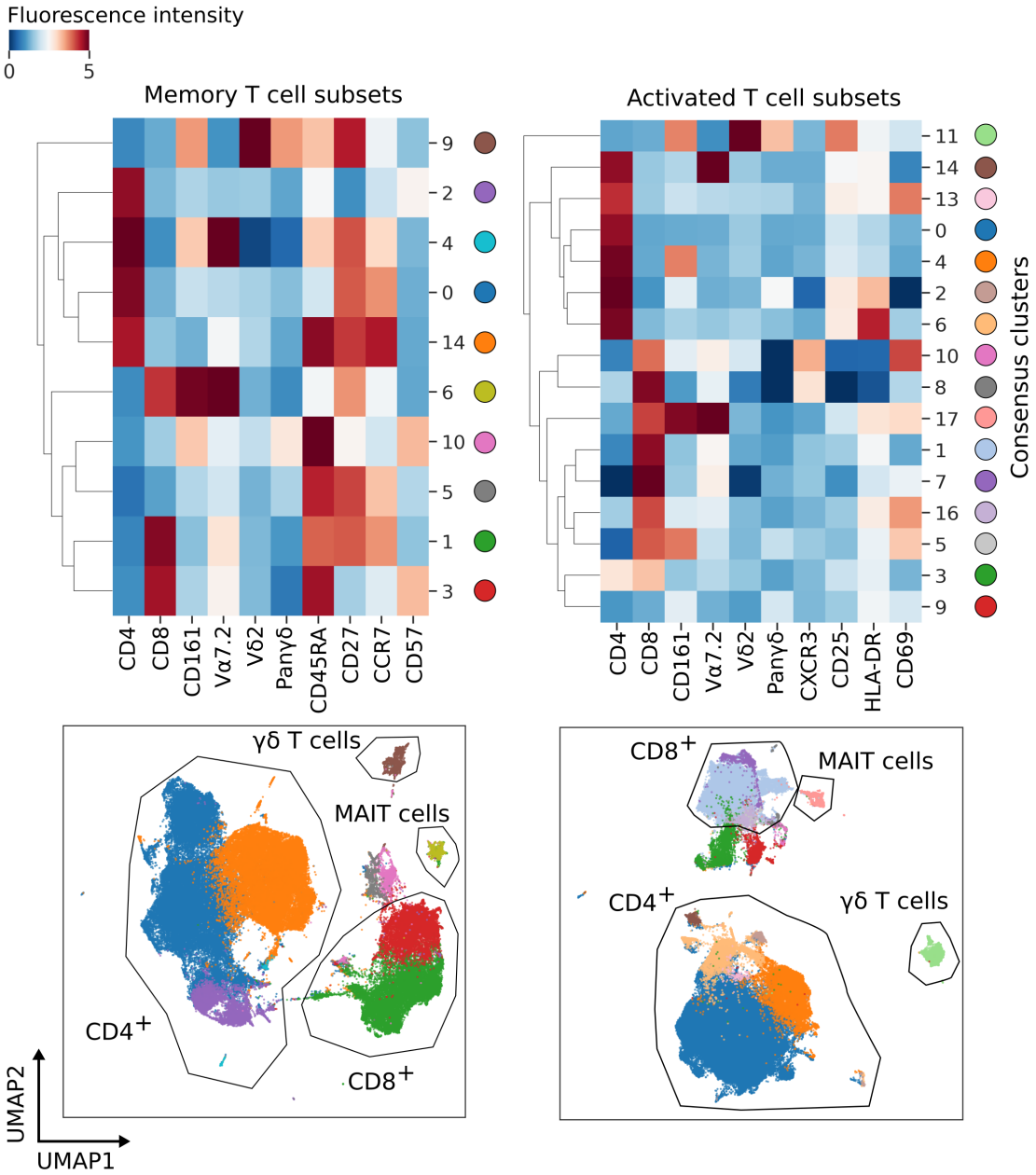

Figure 5.26: GeoWaVe ensemble clustering of T cells. The left heatmap and UMAP scatterplot shows T cells stained for identification of memory subsets whereas the right shows T cells stained for identification of activated subsets. The heatmaps and accompanying UMAP scatterplots show the identified consensus clusters and their expression profile. Fluorescence intensity is shown as hyperbolic arcsine with a cofactor of 150.

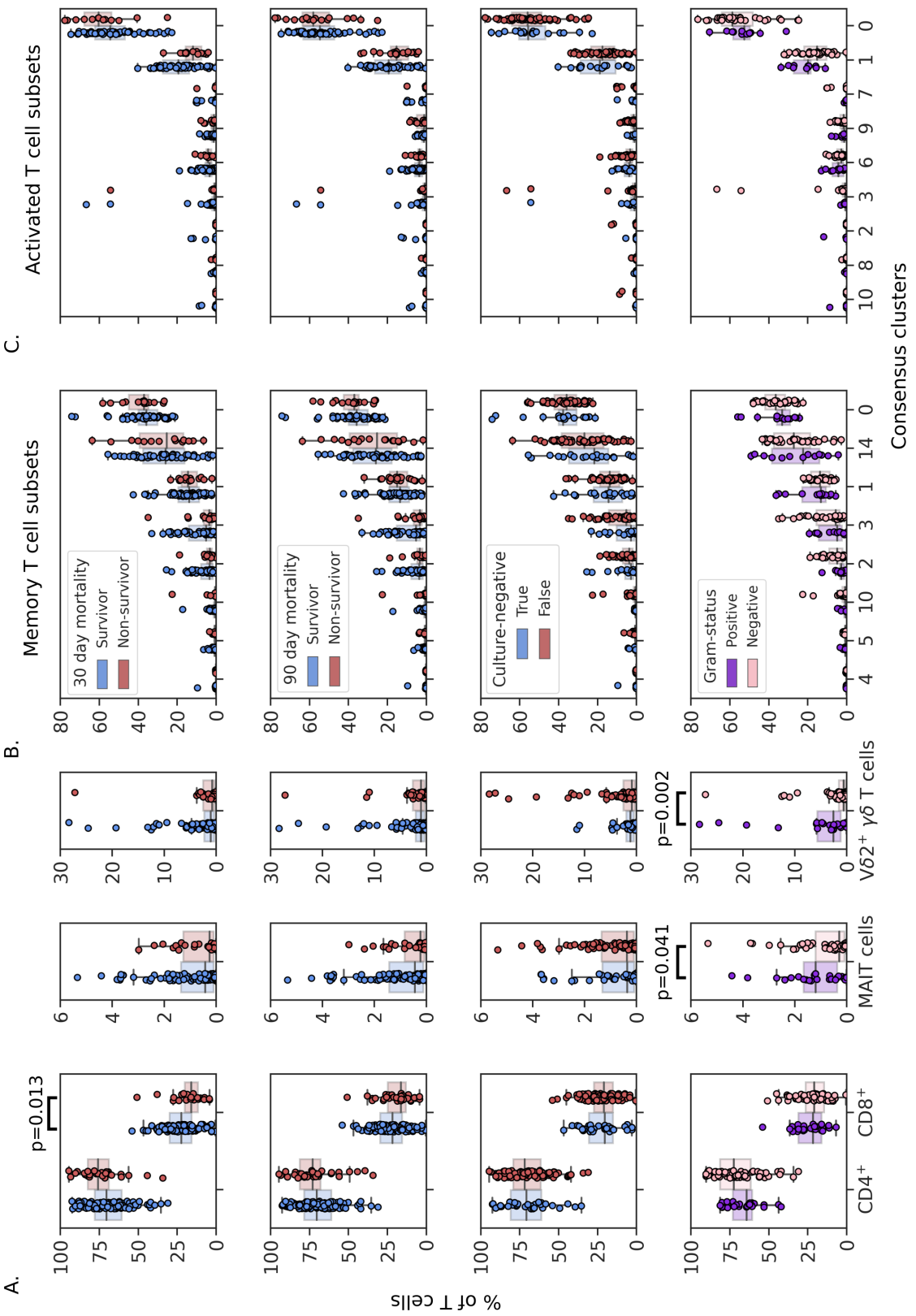

Figure 5.27: Proportion of GeoWaVe consensus clusters as a percentage of T cells. (Continued on the following page.)

Figure 5.27: Proportion of GeoWaVe consensus clusters as a percentage of T cells. Clusters were combined into major subsets and averaged across the two experiments (memory and activation subset staining) to provide proportions of $CD4^+$, $CD8^+$, MAIT cells, and $V\delta2^+$ $\gamma\delta$ T cells (A). Individual clusters from staining for the identification of memory subsets are shown (B), alongside clusters stained for the identification of activated subsets (C). Comparisons between survivors and non-survivors at 30 and 90 days post diagnosis with sepsis are shown (top row and second row, respectively). Additionally, comparisons are shown for those with and without microbiologically confirmed infection (culture negative; third row) and, for those with a confirmed infection, the difference between those with a Gram-positive versus a Gram-negative causative pathogen (bottom row). P-values were generated using a two-tailed Mann-Whitney U test, with Bonferroni-Holm correction for multiple comparisons.

Clustering of all T cells identified unconventional T cell subsets of MAIT cells and $\gamma\delta$ T cells but did not generate the resolution to identify subclusters of these populations. Unconventional T cell populations were of particular interest due to their ability to recognise metabolites of bacteria [28, 259, 250]. They were identified by autonomous gating (Figure 2.2) so that in downstream analysis, detailed clustering could be performed on these populations independent of all other T cells without the need for downsampling.

GeoWaVe consensus clusters of $V\delta2^+$ $\gamma\delta$ T cells were generated using the staining panel for memory subsets (Figure 5.29) and activated subsets (Figure 5.30). The majority of $V\delta2^+$ $\gamma\delta$ T cells expressed high levels of CD161, but a small cluster with distinctly low CD161 expression was found in both staining panels (clusters 5 in Figure 5.29; cluster 3 in Figure 5.30). The largest cluster amongst memory subsets was cluster 4 (Figure 5.29, blue cluster) with a $CD45RA^{hi}$ and $CCR7^{lo}$ phenotype, the next largest clusters are 1 and 3, cluster 3 is similar to cluster 4 but is distinct from all other clusters with expressing CD57. Cluster 1, however, can be characterised as $CD45RA^{lo}$, $CCR7^{lo}$, and $CD57^{lo}$. Comparisons between patient sub-groups for these clusters as a percentage of $V\delta2^+$ $\gamma\delta$ T cells (Figure 5.29 boxplots) did not demonstrate any significant differences.

Amongst $V\delta2^+$ $\gamma\delta$ T cells stained for activation markers (Figure 5.30), clusters exhibited moderate expression of CD25, a marker of activation on lymphocytes, also known as interleukin-2 receptor alpha chain. Meanwhile, all clusters exhibit low expression of the chemokine receptor CXCR3. The clusters were almost exclusively differentiated by their expression of CD161, HLA-DR, and CD69. The two largest clusters (cluster 0 and 2, the orange and blue clusters in Figure 5.30, respectively) had a phenotype of HLA-DR^{lo} and $CD69^{lo}$. Clusters 1 and 4 showed relatively higher expression of CD69 and HLA-DR, respectively. As with

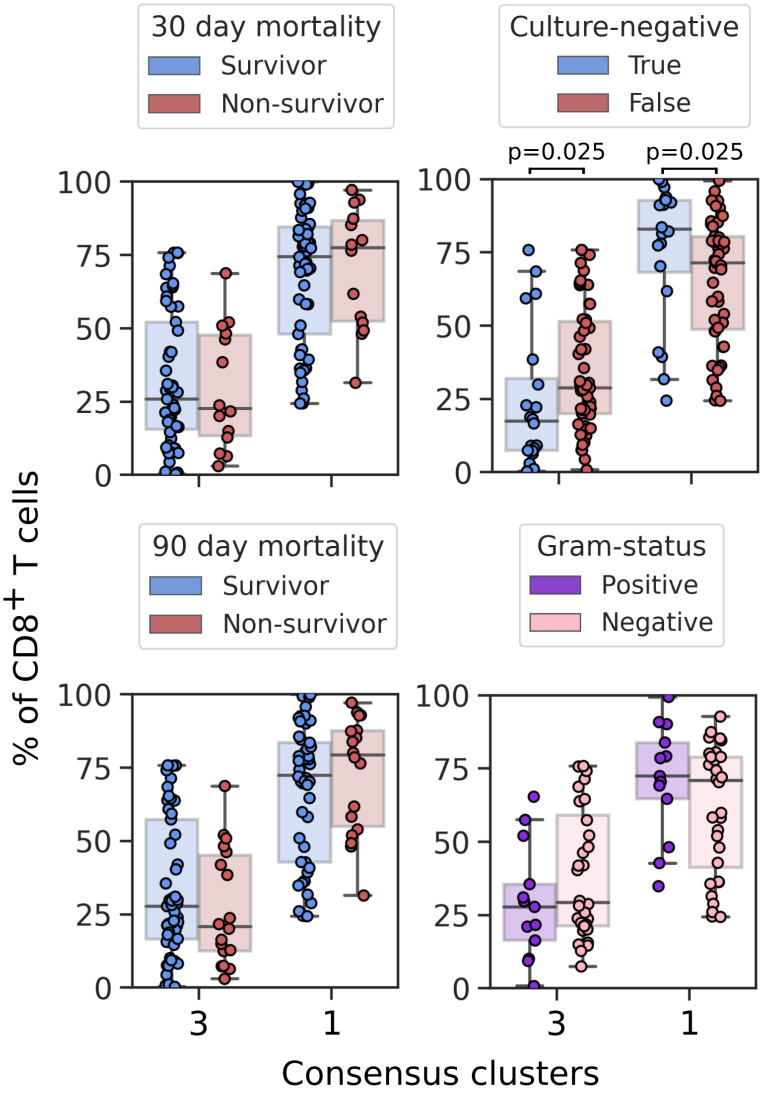

Figure 5.28: Proportion of $CD8^+$ GeoWaVe consensus clusters from PBMCs stained for the identification of memory T cell subsets, as a percentage of total $CD8^+$ T cells. Comparisons between surviviors and non-survivors at 30 and 90 days post diagnosis with sepsis are shown. Additionally, comparisons are shown for those with and without microbiologically confirmed infection (culture negative) and, for those with a confirmed infection, the difference between those with a Gram-positive versus a Gram-negative causative pathogen. P-values were generated using a two-tailed Mann-Whitney U test, with Bonferroni-Holm correction for multiple comparisons.

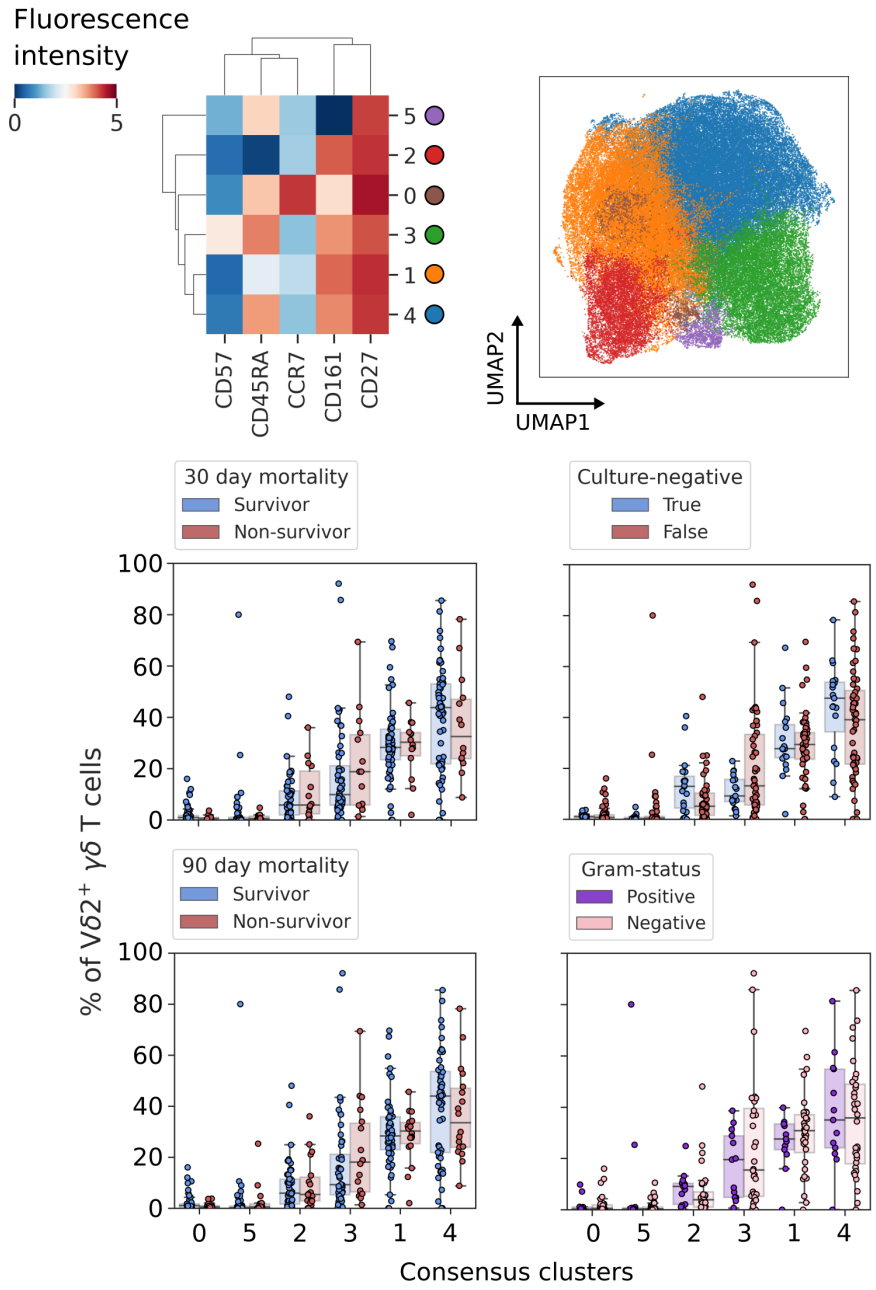

Figure 5.29: GeoWaVe ensemble clustering of V$\delta 2^+$ $\gamma\delta$ T cells, stained for identification of memory subsets. The heatmaps and accompanying UMAP scatterplots show the identified consensus clusters and their expression profile. Fluorescence intensity is shown as hyperbolic arcsine with a cofactor of 150. The proportion of consensus clusters as a percentage of the total V$\delta 2^+$ $\gamma\delta$ T cells are shown in accompanying box plots, with comparisons for survivors and non-survivors 30 and 90 days post sepsis diagnosis, patients with and without microbiologically confirmed infection, and patients with a Gram-negative vs Gram-positive causative pathogen.

V$\delta 2^+$ $\gamma\delta$ T cells stained for memory subsets, comparisons between patient sub-groups of interest for these clusters (Figure 5.30 boxplots) did not yield any significant differences.

GeoWaVe consensus clusters of MAIT cells were also generated using the staining panel for memory subsets (Figure 5.31) and activated subsets (Figure 5.32). MAIT cells were identified as CD3$^+$ Vα7.2$^+$ CD161$^+$ lymphocytes, and in both staining panels, a smaller CD4$^+$ CD8$^-$ population was identified alongside the majority CD4$^-$ CD8$^+$ population. These CD4$^+$ MAIT cells amongst memory subsets were identified as clusters 0 and 5 (Figure 5.31 red and pink clusters, respectively). These clusters differed by their expression of CD27 and CCR7, with lower expression in the much smaller cluster 5. All MAIT cells showed low expression of CD57. The CD8$^+$ MAIT cells, clusters 2, 1 6, and 4 differed by their CD45RA and CD27 expression, but the predominant cluster (blue cluster 2 Figure 5.31) had a phenotype of CD45RA^{mi} CCR7^{mi} CD27^{hi}. The smaller but moderately sized cluster 3 (orange in Figure 5.31) showed an identical phenotype to this majority cluster, except for not expressing CD4 or CD8.

When comparing patient sub-groups for differences in memory subset MAIT clusters as a percentage of total MAIT cells (Figure 5.31 boxplots), survivors and non-survivors were similar in the composition of clusters, as were those with and without microbiologically confirmed infections. Comparisons of Gram-negative versus Gram-positive pathogens in those with a confirmed infection showed interesting trends but ultimately no significant differences; cluster 0 (the CD4$^+$CD8$^-$ cluster) exhibits a trend of increased proportions amongst Gram-negative infections but with a p-value of 0.11 after correction for multiple comparisons. In contrast, cluster 2 was slightly decreased amongst Gram-negative infections but again without statistical rigour, with a p-value of 0.16 after correction for multiple comparisons. It is also worth noting that the analysis presented here described peripheral T cells, and the phenotype of tissue-resident MAIT cells in sepsis may differ significantly.

Only one CD4$^+$ CD8$^-$ cluster was identified amongst MAIT cells stained for activated subsets (green cluster 0 in Figure 5.32) and showed an inactive phenotype. The majority of CD4$^-$ CD8$^+$ clusters 2, 4, and 3 were largely subdued in the expression of activation markers CXCR3, HLA-DR and CD25 and only differed in the expression of CD69. Similar to the staining panel for memory subsets, a CD4$^-$ CD8$^-$ cluster was identified (red cluster 1 in

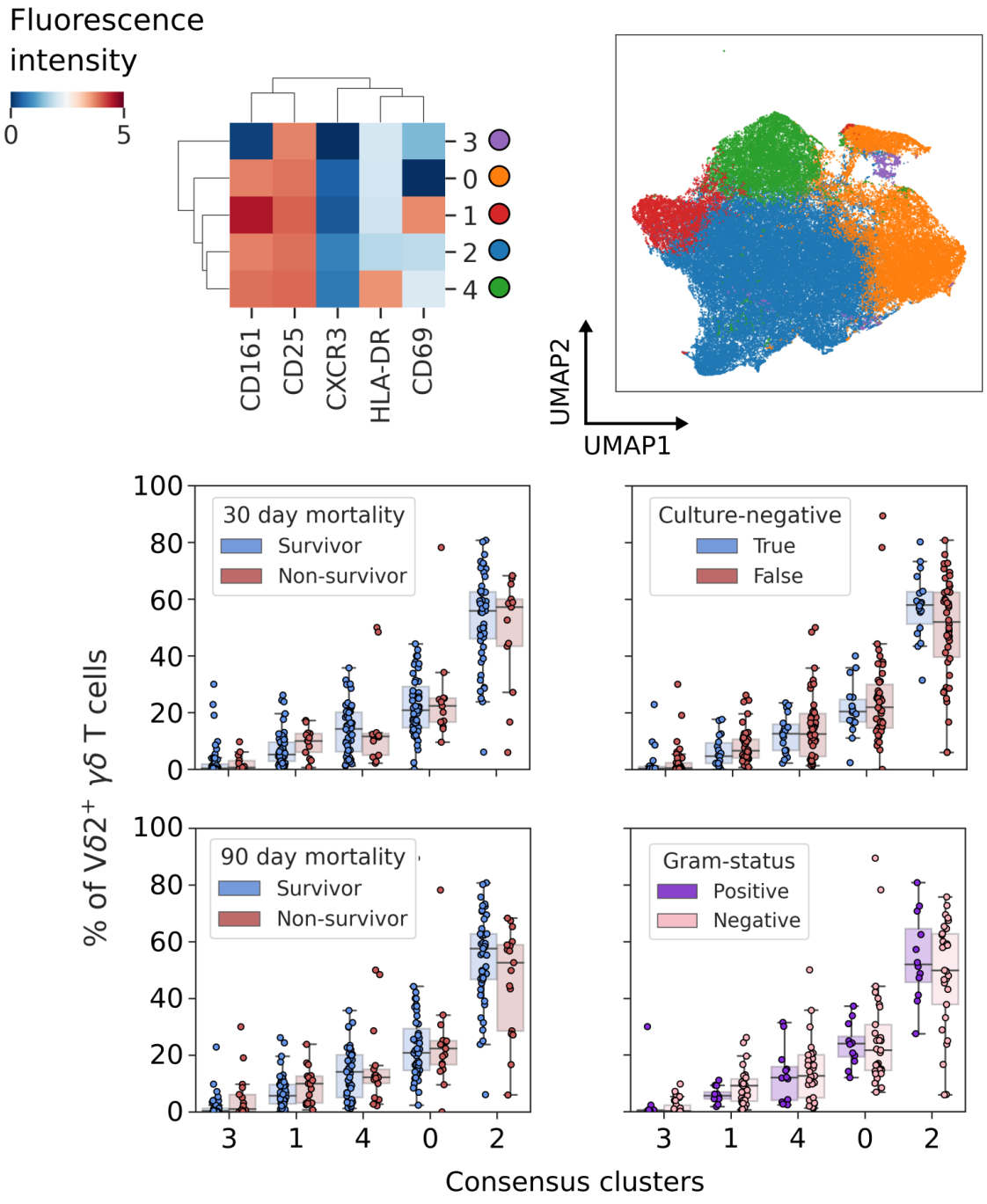

Figure 5.30: GeoWaVe ensemble clustering of V$\delta 2^+$ $\gamma\delta$ T cells, stained for identification of activated subsets. The heatmaps and accompanying UMAP scatterplots show the identified consensus clusters and their expression profile. Fluorescence intensity is shown as hyperbolic arcsine with a cofactor of 150. The proportion of consensus clusters as a percentage of the total V$\delta 2^+$ $\gamma\delta$ T cells are shown in accompanying box plots, with comparisons for survivors and non-survivors 30 and 90 days post sepsis diagnosis, patients with and without microbiologically confirmed infection, and patients with a Gram-negative vs Gram-positive causative pathogen.

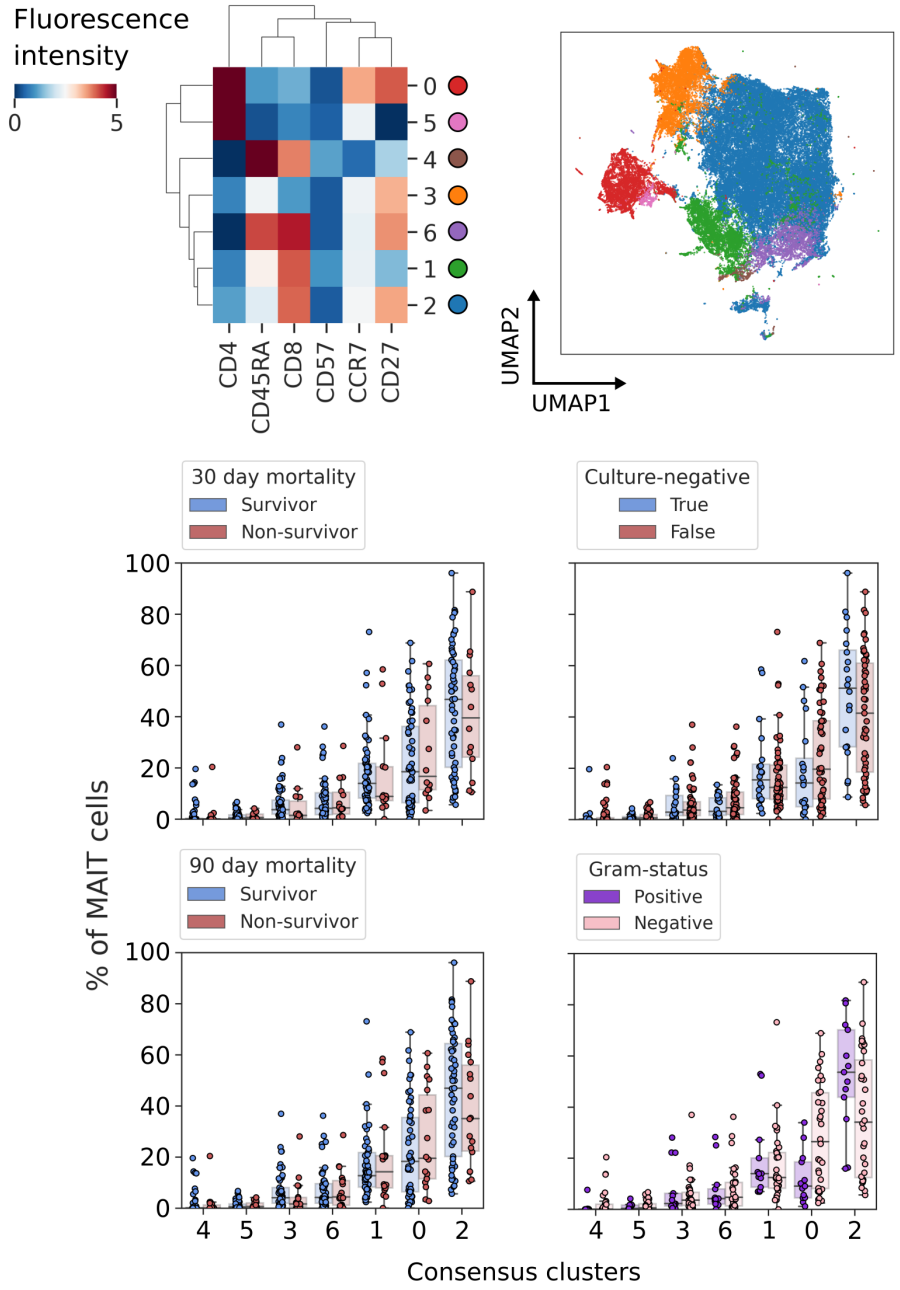

Figure 5.31: GeoWaVe ensemble clustering of MAIT cells ($CD3^+$ $V\alpha7.2^+$ $CD161^+$ lymphocytes), stained for identification of memory subsets. The heatmaps and accompanying UMAP scatterplots show the identified consensus clusters and their expression profile. Fluorescence intensity is shown as hyperbolic arcsine with a cofactor of 150. The proportion of consensus clusters as a percentage of the total MAIT cells are shown in accompanying box plots, with comparisons for survivors and non-survivors 30 and 90 days post sepsis diagnosis, patients with and without microbiologically confirmed infection, and patients with a Gram-negative vs Gram-positive causative pathogen.

Figure 5.32); this cluster showed low expression of all activation markers present within the staining panel.

When comparing patient sub-groups for differences in MAIT cells clustered by activation markers (Figure 5.32 boxplots), as with cells stained for memory subsets, survivors and non-survivors were similar in the composition of clusters, the same was true for those with and without microbiologically confirmed infections. However, comparisons of Gram-negative versus Gram-positive pathogens in those with confirmed infection showed a significant increase in the $CD4^+$ $CD8^-$ cluster 0 in Gram-negative infections. There was also a trend with a slight decrease in the $CD8^+$ $CD69^{lo}$ cluster 2 amongst Gram-negative infections. Albeit with a p-value of 0.14 after corrections for multiple comparisons, therefore it is not possible to claim any certainty about this difference.

In the observations of MAIT cells and V$\delta2^+$ $\gamma\delta$ T cell clustering of activation markers, it was evident that some markers had greater influence over clustering than others; CD161, HLA-DR, and CD69 appeared informative for V$\delta2^+$ $\gamma\delta$ T cell clustering whilst CXCR3 and CD25 did not convey any differences (Figure 5.30), and CD69 appeared to be the only informative activation marker amongst MAIT cell clusters (Figure 5.32). It was, therefore, questioned whether the mean fluorescent intensity (MFI) of activation markers considered in isolation would be informative variables. No significant differences were observed for comparisons of survivors and non-survivors. However, the MFI of CD69 on both MAIT cells and V$\delta2^+$ $\gamma\delta$ T cells appeared to differentiate Gram-positive from Gram-negative infections (Figure 5.32). The MFI of CD25 on MAIT cells was also informative, with greater expression in those with a Gram-negative causative pathogen, despite similar expression of CD25 across identified clusters.

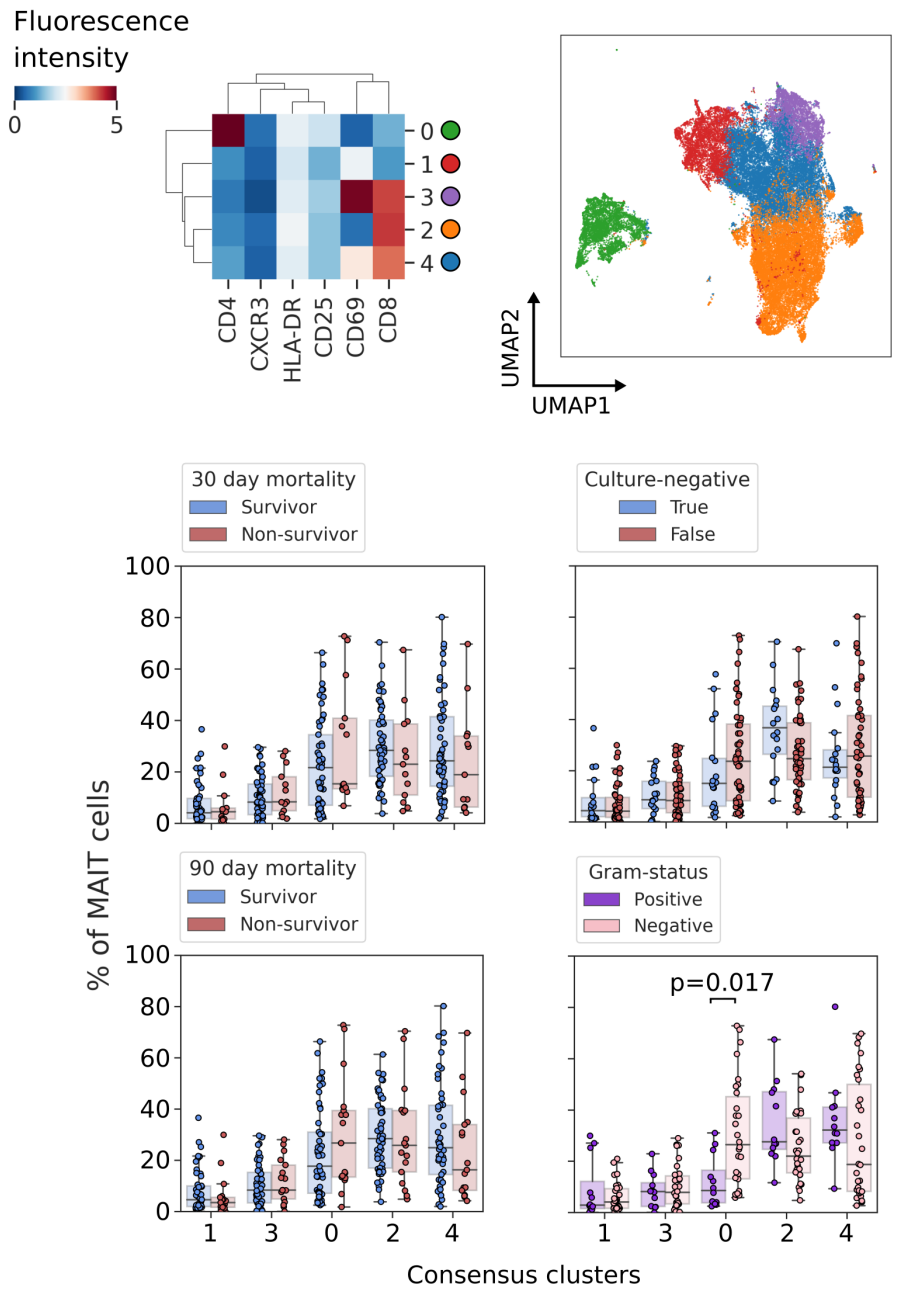

Figure 5.32: GeoWaVe ensemble clustering of MAIT cells ($CD3^+$ $V\alpha7.2^+$ $CD161^+$ lymphocytes), stained for identification of activated subsets. The heatmaps and accompanying UMAP scatterplots show the identified consensus clusters and their expression profile. Fluorescence intensity is shown as hyperbolic arcsine with a cofactor of 150. The proportion of consensus clusters as a percentage of the total MAIT cells are shown in accompanying box plots, with comparisons for survivors and non-survivors 30 and 90 days post sepsis diagnosis, patients with and without microbiologically confirmed infection, and patients with a Gram-negative vs Gram-positive causative pathogen. P-values were generated using a two-tailed Mann-Whitney U test, with Bonferroni-Holm correction for multiple comparisons.

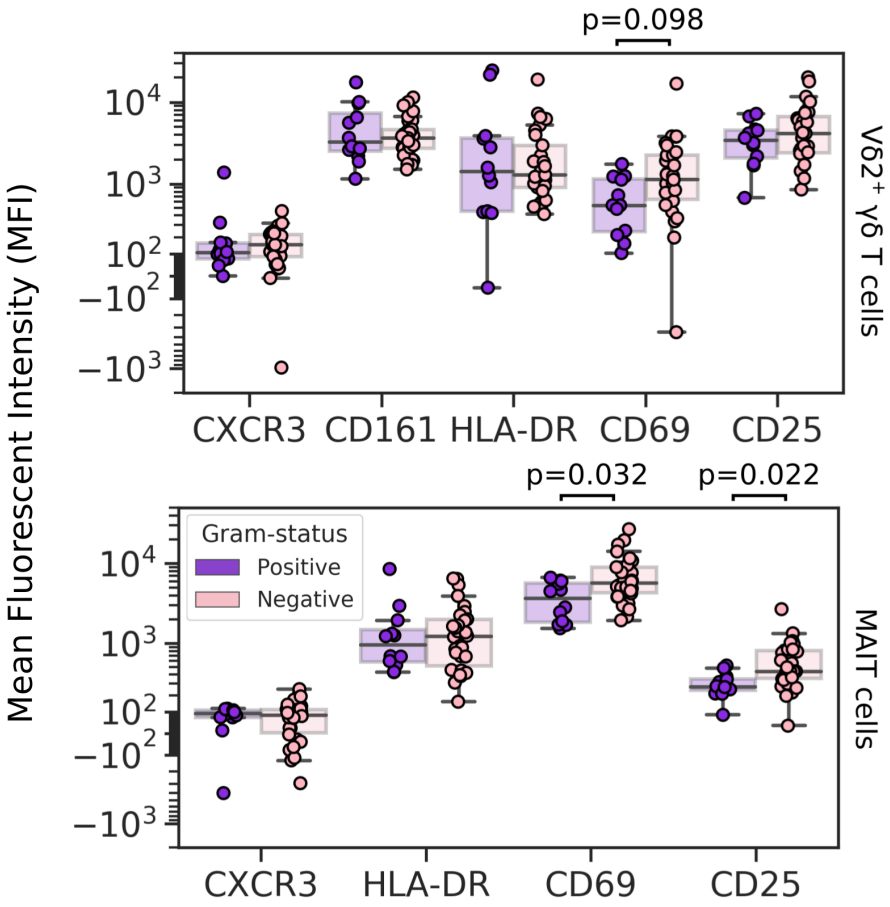

Figure 5.33: The mean fluorescence intensity (MFI) of HLA-DR, CD86, CD46, CD40 and CD62L on MAIT cells, with comparisons between sepsis patients with a Gram-positive versus a Gram-negative infection. P-values were generated using a two-tailed Mann-Whitney U test, with Bonferroni-Holm correction for multiple comparisons.

5.4 Discussion

In this chapter, the phenotype of 77 patients with a diagnosis of sepsis within 36 hours of the presumed onset of infective illness was detailed. The objective was to identify significant differences in the presentation of survivors compared to non-survivors and identify phenotypes that identify the causative pathogen. Survivors of sepsis in the ILTIS cohort consisted of a slightly younger population with a greater representation of females with fewer interventions such as mechanical ventilation and renal replacement therapy. However, the two population demographics did not significantly differ.

It is estimated that for anywhere between 28 to 89% of patients with sepsis, the causative pathogen is not identified and there are contradicting findings amongst retrospective studies as to how the severity of illness, length of stay, and in-hospital mortality compare between 'culture-negative' and 'culture-positive' sepsis [77]. It is still uncertain whether 'culture-negative' sepsis represents a different clinical entity with the possibility of the absence of infection entirely. Unfortunately, the picture is blurred by antimicrobial intervention reducing the yield of bacterial culture [75, 74]. An increased proportion of 'culture-negative' patients in the ILTIS study had been admitted with some traumatic injury or emergency surgery. Although not statistically significant (p-value of 0.320), this could have influenced antimicrobial intervention and contributed to the inability to identify the causative pathogen. The distinction between 'culture-negative' and 'culture-positive' sepsis within this cohort was an essential comparison before investigating the differences between Gram-negative and Gram-positive infection. Significant variation in the immunophenotypes of culture-negative and culture-positive sepsis would be expected if clinically distinct.

This thesis aimed to identify variations in immunophenotypes that would provide informative features for multi-variate models tasked with predicting mortality and the underlying cause (discussed in-depth in Chapter 6). Retrospective studies such as ILTIS allow deep clinical data mining of routine data collected prior to sepsis diagnosis. Routine clinical data provides additional information for models whilst also granting perspective on the value immunophenotyping provides beyond readily available clinical biomarkers.

A total of 63 variables derived from routine clinical data were compared amongst patient sub-groups. Nevertheless, amongst this large repository, only inspired oxygen (FiO2) differed between survivors and non-survivors after correcting for multiple comparisons. Differences in FiO2 were only realised for values measured closest to the diagnosis of sepsis; its significance was reduced when values were averaged across samples taken 48 hours prior to diagnosis. The observations are problematic as the sampling time could differ severely between patients. It is also difficult to ascertain whether FiO2 is an informative biomarker when considering confounding factors such as mechanical ventilation, hence the need for multi-variant modelling. No other clinically available data were particularly informative, including comprehensively studied biomarkers such as CRP and arterial lactate.

The lack of informative biomarkers amongst routine clinical data identifies the need for detailed phenotyping. Sepsis is understood as a dysregulation of the patient's response to infection, and therefore it makes sense that the early immune response could yield informative biomarkers. Multiplex assays and ELISAs were performed to investigate a broad range of soluble biomarkers, including cytokines and chemokines that regulate the immune response. Analytes included pro-inflammatory molecules, such as IL-1β, GM-CSF, IFN-γ, and TNF-α, and anti-inflammatory molecules, such as IL-4, IL-6, and IL-10. Procalcitonin was included for its reported prognostic utility [49] and potential value in differentiating Gram-positive and Gram-negative infections [85]. Other promising biomarkers were included, such as ferritin, the hallmark of hyperferritinemic sepsis that poses a greater risk of mortality [260, 248]. Ferritin is also a crucial biomarker for identifying Haemophagocytic lymphohistiocytosis (HLH). HLH is a syndrome of severe immune dysregulation with a very similar presentation to sepsis but is often under-recognised and exhibits a high rate of mortality [261].

Of the 19 analytes measured, only CXCL10 and IL-15 plasma levels showed a significant difference between survivors and non-survivors. CXCL10 levels were decreased in non-survivors when comparing 30-day mortality, whereas IL-15 levels were increased for non-survivors. The difference at this time-point is severe but then subdued when comparing 90-day mortality. The observed relationship could indicate that CXCL10 and IL-15 levels are only informative for early mortality. However, it is suspected that this effect could also result from a difference in the class imbalance between the two time points (given that more non-survivors are seen at 90 days compared to 30 days). CXCL10, the ligand for CXCR3, is produced in response to IFN-γ and is responsible for lymphocyte migration [246]. CXCL10 has been implicated in other types of severe infectious disease [264, 262, 263], and CXCR3 blockade has been suggested as a potential therapeutic target in sepsis [265]. The findings of this study do not support the claim that increased levels of CXCL10 in plasma are associated with increased mortality, but rather the opposite, and the loss of effect seen at the later 90-day time-point makes it challenging to conclude the findings here. The observations for IL-15 are equally suspect, given that most patient values for IL-15 were below the detection limit of the Luminex assay.

Flt3L was significantly increased in patients with a confirmed infection compared to those with culture-negative sepsis. As far as I am aware, Flt3L has not been studied as an etiologi-

cal biomarker in sepsis, but Flt3L is critical to the differentiation of DCs [247]. The absence of bacterial culture is not definitive, and therefore Flt3L could benefit from comparison with other indicators, such as severity scores. Flt3L was comparable between Gram-negative and Gram-positive infections. In fact, the only analyte that demonstrated a significant difference between Gram-negative and Gram-positive infections was ferritin, with decreased levels in Gram-negative infections. Increased ferritin levels are associated with higher mortality, resulting from “Hyperferritinemic Syndrome", a condition in which high concentrations of iron poor ferritin induce both pro-inflammatory cytokines and immunosuppression [248]. The relationship with the causative pathogen is unclear, and I failed to identify previous studies comparing pathogens in hyperferritinemic sepsis. Future studies should compare the underlying cause in sepsis and whether this impacts plasma ferritin concentrations. Amongst the other analytes, a trend was observed in PCT levels with lower values in Gram-negative compared to Gram-positive infection. However, the findings were not statistically significant, and therefore this study failed to confirm previous reports of PCT differentiating between a Gram-positive and Gram-negative pathogen [83, 86, 85].

The primary limitation in the analysis of cytokines, chemokines, and acute phase proteins in plasma was the detection limit on the chosen Luminex^{TM} multi-plex assay, resulting in all but six analytes being outside the detectable range for 20% or more of the tested samples. Future analysis should include greater sensitivity when attempting to quantify these analytes in plasma, especially when sampled from patients at such an early stage of sepsis when concentrations might be low and difficult to detect. Contrarily, ferritin and procalcitonin would benefit from individual assays with care taken to dilute samples sufficiently to avoid the risk of concentrations beyond the upper limit of detection.

Samples were separated into those above and below the assay detection limits to help capitalise on available data. Comparisons were made as if these assays generated a binary result with the upper and lower cut-off values. The resulting odds ratios showed large confidence intervals due to the small sample size in the resulting contingency table but still identified some familiar trends. Increased levels of IL-6 showed a trend toward increased odds of mortality at both 30 and 90 days post sepsis diagnosis, which aligns with previous observations for this anti-inflammatory cytokine [51, 66]. Higher levels of IL-15 showed a trend towards higher odds of mortality at 30 days after sepsis diagnosis and, to a lesser extent, at 90 days. IL-15 is described as a ‘bi-directional’ cytokine with both pro-inflammatory and immunoreg-

ulatory effects and promotes the proliferation of $CD8^+$ and $CD4^+$ memory T cells and NK cells [246]. This function has been implicated in the pathobiology of septic shock; therefore, IL-15 could be a potential indicator of severity [266, 267].

After quantifying soluble factors, the phenotype of $CD8^+$ and $CD4^+$ T cells, MAIT cells, V$\delta2^+$ $\gamma\delta$ T cells, monocytes, and neutrophils were described. The first observation was a profound reduction in circulating T cells in non-survivors, a phenomenon well documented in sepsis [268, 269, 270]. Amongst T cells, a trend towards a reduction in the $CD8^+$ T cell compartment is observed amongst non-survivors which supports the observation of an immunosuppressive phenotype dominating in sepsis [270, 271].

In this work, the focus was given to monocytes and neutrophils as antigen-presenting cells of the innate immune response, influenced by previous findings that Vγ9 V$\delta2$ $\gamma\delta$ T cells can induce an APC-like phenotype in neutrophils, with similar phenotypes observed in sepsis patients [23]. Decreased monocyte HLA-DR expression as an indicator of severity in sepsis is well understood [251, 252, 70] and confirmed by the findings in this study, albeit that the reduction in HLA-DR MFI was small; possibly reflecting the very early time point of sepsis that was captured in this work. CD62L (L-selectin) expression differentiated monocyte clusters, and CD62L monocyte MFI was slightly increased in non-survivors. This type-1 transmembrane glycoprotein functions in cell adhesion and has the unique feature of being rapidly shed from the cell surface upon activation [272]. Subsequently, high levels of soluble L-selectin in serum have been identified as a potential predictor of survival in sepsis [273].

Neutrophils were significantly increased in Gram-negative compared to Gram-positive infections. The opposite was observed for MAIT cells and $\gamma\delta$ T cells, unconventional T cell populations that have been described as having a microbe-specific response to bacterial infection [23, 274]. In both automated gating and clustering analysis, both MAIT cells and V$\delta2^+$ $\gamma\delta$ T cells were found to be significantly decreased in Gram-negative infection. The observations were surprising given that most Gram-negative pathogens are producers of HMB-PP, a well-described stimulant of $\gamma\delta$ T cell activation and proliferation [275, 26]. In previous studies, increased proportions of circulating $\gamma\delta$ T cells were observed amongst patients infected with HMB-PP^+ pathogens in acute peritonitis [175] and sepsis [23].

Given that unconventional T cell subsets have been found to accumulate at the site of infection [259, 277, 276], it is possible that we are observing the diffusion of these populations

into the site of infection. However, without local sampling and a comparison with PBMCs, it is impossible to clarify from this study alone. Interestingly, the MFI of the activation markers CD69 and CD25 were increased in MAIT cells in Gram-negative infections compared to Gram-positive. Again, it must be stressed that the differences observed are subtle. Therefore, a larger sample size would be needed to clarify whether the reduction of these subsets amongst circulating T cells and increased expression of activation markers is correlated with Gram-negative aetiology.

When investigating distinct clusters of MAIT and $\gamma\delta$ T cells, no significant differences were observed between patient subsets apart from a $CD4^+$ $V\alpha7.2^+$ $CD161^+$ cluster, which was increased in Gram-negative infections compared to Gram-positive. The majority of MAIT cells express a $CD8^+$ $CD4^-$ profile, but a subset of $CD4^+$ MAIT cells has been previously described and noted to produce more IL-2 than other subsets [254]. However, to the best of our knowledge, a $CD4^+$ $V\alpha7.2^+$ $C161^+$ population has not been described in acute sepsis or observed as increased in frequency in Gram-negative compared to Gram-positive infection.

The time points for all-cause mortality, 30 and 90-days, pose a significant limitation. Although they reflect a time point common amongst clinical trials and therefore offer comparable findings, mortality is complicated by unobserved confounding variables such as interventions and events occurring after the initial sepsis diagnosis. The methodology applied here also missed the opportunity for more granular comparisons of the time-to-death, which could have been achieved with survival analysis utilising either Kaplan-Meier or a Cox Proportional-Hazard models.

A significant limitation in this study and a possible explanation of the variation observed within patient sub-groups is the heterogeneity of the cohort. Amongst the patients studied, around 25% were admitted to ICU with trauma or following emergency surgery. There was insufficient data regarding patient co-morbidity or history of infectious disease, and less than 70% of patients had a confirmed infection by positive culture. Heterogeneity in sepsis, with its various patterns of presentation, has long been recognised as a barrier to the advancement of diagnosis and therapy [279, 278]. The Sepsis-3 definition [243] alone does not distinguish between the complex heterogeneity observed in the pathophysiology of sepsis [280]. Therefore future studies should seek to recruit patients with more restrictive inclusion criteria. Unsupervised clustering methods could drive the identification of endotypes for recruitment

[281, 282], or simple strategies employed to limit recruitment to those of comparable aetiology *e.g.* culture-positive urosepsis or pneumonia amongst patients in a defined age bracket.

In the next chapter, I will introduce multi-variant modelling that takes as input variables all data sources described in this chapter (routine clinical data, soluble analytes in plasma, and immune cell phenotypes), as well as some additional data from collaborators. Here, I will attempt to address the heterogeneity of sepsis by modelling the complex interaction of the observed variables to try and obtain generalised patterns predictive of mortality and cause.

6 | Machine learning models identify biomarker signatures correlated with mortality and causative pathogen in sepsis

6.1 Introduction

Sepsis is a life-threatening disorder with complex pathophysiology that has yet to be fully described. The multi-faceted nature of sepsis requires personalised care with rapid intervention. For example, the administration of early antimicrobial therapy has been shown to reduce mortality [233, 78, 76] and yet the inability to identify the causative pathogen could lead to inappropriate antibiotic use and the development of multiple-resistant organisms [235, 283]. The need to identify the cause and then direct urgent care has driven an interest in identifying diagnostic and prognostic biomarkers [155, 284, 96] and developing complex algorithms that leverage existing electronic health record systems [286, 288, 287, 289, 285, 290, 291, 101]. The former has primarily focused on single biomarker studies with mixed results [155, 96] and only in recent years has the value of employing multiple biomarkers in combination been demonstrated [293, 292, 294, 98]. The latter has focused mainly on diagnostic tools to identify sepsis patients early in their disease pathway. They rely almost entirely on routine clinical data that do not capture the complex pro-inflammatory and anti-inflammatory mechanisms that are now known to contribute significantly to sepsis pathology. The work described in this chapter will combine routine clinical data with immunological profiling, combining biomarkers in machine learning models to describe complex patterns predictive of mortality or underlying cause. Mortality prediction is a valuable tool for prioritising care in a resource-limited environment, and the ability to predict the underlying pathogen could help direct targeted care and improve antibiotic stewardship.

Studies have only started incorporating a combination of clinically available data with novel biomarkers to create predictive models for sepsis in recent years. Kofoed *et al.*[97] presented a logistic regression model combining six biomarkers, including novel markers not routinely collected and showed good diagnostic accuracy for differentiating a bacterial or non-bacterial cause of inflammation. Langley *et al.* [295] demonstrated a model for predicting mortality

in sepsis that combined clinical features with five metabolites, showing the potential for integrating metabolomics into predictive models. Taneja *et al.* [99] studied multiple machine learning algorithms that combine clinical data with non-traditional biomarkers (including pro- and anti-inflammatory biomarkers) and showed how a support vector machine could stratify adult patients with sepsis based on severity. A follow-up study in 2021 from the same authors [100] showed in a larger cohort that a combination of three non-routinely measured biomarkers combined with electronic health record (EHR) data had diagnostic and prognostic potential. Another recent multi-centre study of over 500 sepsis patients identified multiple machine learning models with prognostic capabilities, pooling EHR data with novel biomarkers [296].

The collective weight of this work demonstrates the advantage of sourcing a diverse feature space for predictive modelling. With the advent of multi-omics technology, there is a growing abundance of data, with the promise that a multi-layered approach to phenotyping the immunological response to sepsis could help identify diagnostic and prognostic signatures with direct application to the clinic [297, 298]. This landscape presents the challenge of analysing extensive high-dimensional data from which informative biomarker combinations must be found. Such a task is analogous to feature selection in machine learning. The minimal yet optimal variables are identified to help reduce model complexity, avert overfitting, and improve performance [299]. Numerous feature selection methodologies already exist, each with its benefits and disadvantages [159, 161, 160]. Typically, a single feature selection method is employed, yet the choice of features to include in a multivariate statistical model depends on the choice of algorithm. Since no single machine learning algorithm will be optimal for every task [301, 300], it is advised to search across multiple solutions and make conclusions based on the performance of observed data. Therefore, it is logical that no single feature selection algorithm will be optimal. Experimenting with multiple methodologies will reduce the risk of overlooking an informative signature or focusing on a single sub-optimal solution. In this chapter, multiple feature selection methodologies will be explored over a range of classification algorithms in a search strategy to identify optimal signatures for predicting outcomes and underlying pathogens in sepsis.

Data-driven pattern recognition with feature selection has successfully identified predictive signatures in cancer prognosis [302], pathogenic cause of peritonitis [199], diagnosis of psychiatric disorders [303], prognosis and treatment response in traumatic injury [304], prog-

nosis in COVID-19 [306, 305], and vaccine response [46]. This demonstrates the benefit of multi-omic data mining and is the motivator for the work described in this chapter. The work here will also intend to leverage the growing field of interpretable machine learning [166] to interrogate the decision-making behind the signatures identified, the influence of individual biomarkers on predictive models, and the identification of patterns that warrant further investigation into sepsis pathogenesis. The application of model agnostic methods for the measure of feature importance was recently demonstrated for predicting multiple organ dysfunction in paediatric sepsis [307], resistance to ionising radiation in cancer therapy [308], and the identification of risk factors in COVID-19 [309]. It is hoped that the inclusion of explainable machine learning algorithms in this work will help increase the accessibility and confidence in predictive models and encourage engagement with results for the generation of studies that intend to validate identified signatures or investigate causal relationships.

6.2 Aims

1. Develop an analytical workflow for generating interpretable machine learning models exposed to data with complex challenges including small sample size, missing data, and class imbalance.
2. Identify minimal and optimal feature sets that could present opportunities for predictive signatures and help derive hypotheses for future studies.
3. Create and critically assess the performance of binary classification models tasked with predicting:
   (a) Death 30 days after diagnosis with sepsis (30-day mortality).
   (b) Death 90 days after diagnosis with sepsis (90-day mortality).
   (c) Gram-negative sepsis vs Gram-positive sepsis.
4. Interrogate the influence of individual features on machine learning models and identify potential predictive signatures that may warrant additional study.

6.3 Preparing multi-omic clinical data for multivariate modelling

Throughout the analysis of the ILTIS data, detailed in Chapter 5, a MongoDB database was populated with clinical parameters, Luminex and ELISA results, and summary statistics of immunological populations acquired by flow cytometry[1]. In addition, collaboration with Ms. Linda Moet and Prof. Peter Ghazal provided lipid data for 52 patients, acquired with mass spectrometry analysis of cell-free plasma (see Methods & Materials 2.1.8 for details). Before the development of statistical models, the predictor variables generated from these activities (referred to as 'features' from this point forward) had to be collated and prepared. Data were combined into a table of 267 features (complete list available in Appendix Table A.2), which could be broadly categorised into physiology (e.g. age, gender), interventions (e.g. ventilation, renal replacement therapy), point of care testing (e.g. blood gas analysis), clinical laboratory results (e.g. full blood count, liver profile), cytokine and chemokine plasma concentrations (e.g. IL-6, IFNγ, CXCL10), proportions of immunological cell populations (T cells, monocytes, and neutrophils, and sub-clusters of each), mean fluorescence intensity of activation markers of immunological populations, and the aforementioned lipid measurements.

This data table, consisting of 77 patients (rows), also included three binary target variables: mortality 30 days after diagnosis with sepsis (abbreviated to '30 day mortality' in remaining text), mortality 90 days after diagnosis with sepsis (abbreviated to '90 day mortality' in remaining text), and Gram-negative sepsis (if a culture result was available). The four variables served as the target for prediction by binary classification models. All 77 patients had complete data available for mortality (both at 30 and 90 days), but only 46 patients had a value for the Gram status of causative pathogens. Patients missing the Gram-status outcome variable were excluded from models that predicted Gram-negative sepsis.

[1]Throughout this chapter, references are made to Chapter 5 and the clustering analysis that yielded variables included within predictive models. For convenience, a summarised version of clustering results is provided in appendix A.1 or accessed as a separate PDF from https://github.com/burtonrj/iltis_summary/blob/main/8_1.pdf

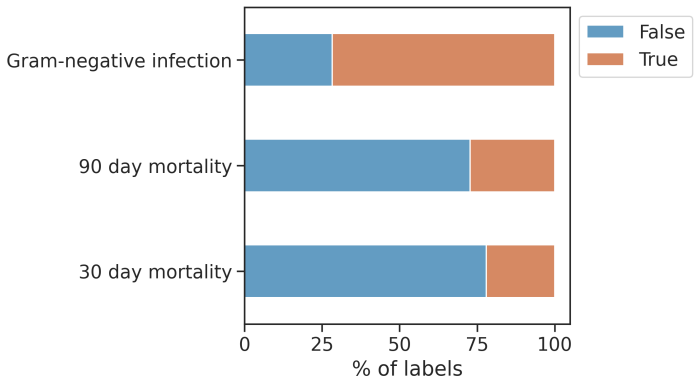

Figure 6.1: Class imbalance amongst target variables for binary classification models.

All outcome variables demonstrated severe class imbalance, with the majority class representing more than 70 per cent of the total patients for each target (Figure 6.1). Resampling methodologies such as SMOTEEN [310, 311] were considered. However, the need for imputation of missing values (addressed in section 6.4) combined with the small amount of available data introduced concerns that the minority class could not be inflated accurately, and this could introduce bias and reduce the ability of models to generalise. Equally, undersampling techniques were considered inappropriate, given the limited data. Alternatively, models were chosen that would allow for introducing a term to penalise the misclassification of the minority class. Take as a simple example Logistic Regression for binary classification, where the coefficients are optimised through an algorithm that seeks to minimise the negative log-likelihood (called the 'loss' function):

$$\frac{1}{N}\sum_{i=1}^{n}(y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)) \tag{6.1}$$

Where N is the number of values, y is the actual value of the target class, and $\hat{y}$ is the predicted value. In the case of binary classification, this loss function can be modified to introduce weights for the negative ($w0$) and positive class ($w1$):

$$\frac{1}{N}\sum_{i=1}^{n}(w0(y_i log(\hat{y}_i)) + w1((1-y_i) log(1-\hat{y}_i))) \tag{6.2}$$

In this analysis weights were balanced according to the formula:

$$w_i = \frac{N\ obs}{N\ classes\ \times\ Nobs_i} \tag{6.3}$$

This penalises the misclassification of the minority class and encourages the model to search for a more optimal solution. Class weights can be extended to other linear models and more complex classifiers such as support vector machines and ensembles of decision trees [312].

Metrics for evaluating model performance were chosen to account for class imbalance. The F1 score is the harmonic mean between precision and recall. It can provide a reliable evaluation of model performance in the case of class imbalance by calculating the F1 score for each class independently and taking an unweighted average (known as the 'macro' F1 score):

$$Precision = \frac{tp}{tp + fp} \tag{6.4}$$

$$Recall = \frac{tp}{tp + fn} \tag{6.5}$$

Where tp is the number of true positives, fp is the number of false positives, and fn is the number of false negatives.

$$F1_{Macro} = \frac{2}{C} \sum_{c \in C} \frac{Precision_c Recall_c}{Precision_c + Recall_c} \tag{6.6}$$

Where c is a class in all possible classes C. This procedure can be repeated for accuracy to give the 'balanced accuracy' which in the case of a binary classify is equal to the arithmetic mean of sensitivity and specificity. The average is widely understood and is therefore accessible to a broad audience:

$$Accuracy_{Balanced} = \frac{1}{2}(\frac{tp}{tp + fn} + \frac{tn}{tn + fp}) \tag{6.7}$$

The area under the receiver operating characteristic curve (ROC-AUC) is a metric common to the medical literature and well understood by scientists and clinicians for reporting biomarkers [313]. Due to its popularity in the biomarker literature and interpretability, it was included alongside the macro F1 score and balanced accuracy. The macro AUC can be calculated similarly to the F1 score and balanced accuracy, but the AUC score should still be interpreted with care.

Given the small sample size (77 patients in total), the 'curse of dimensionality' [314] would be a significant challenge with the 267 predictor variables. The aforementioned 'curse' refers to problems when analysing high-dimensional data. As the feature space expands, the volume of available data rapidly becomes sparse and difficult to sample reliably, requiring exponentially more data. A series of steps were taken prior to modelling to reduce the feature space (summarised in Figure 6.2) and are detailed in subsequent sections of this chapter:

1. Features that are obviously redundant were removed e.g. where information is duplicated or all values are equal for for all patients.
2. Features removed with excessive missing data or where the assumption of 'missing at random' is clearly violated.
3. Multicollinear features were identified and either replaced with an estimated latent variable or the variables with the greatest mutual information with the target variables were retained.
4. Feature selection methodologies were employed that would reduce the feature set to the most informative variables driven by model performance.

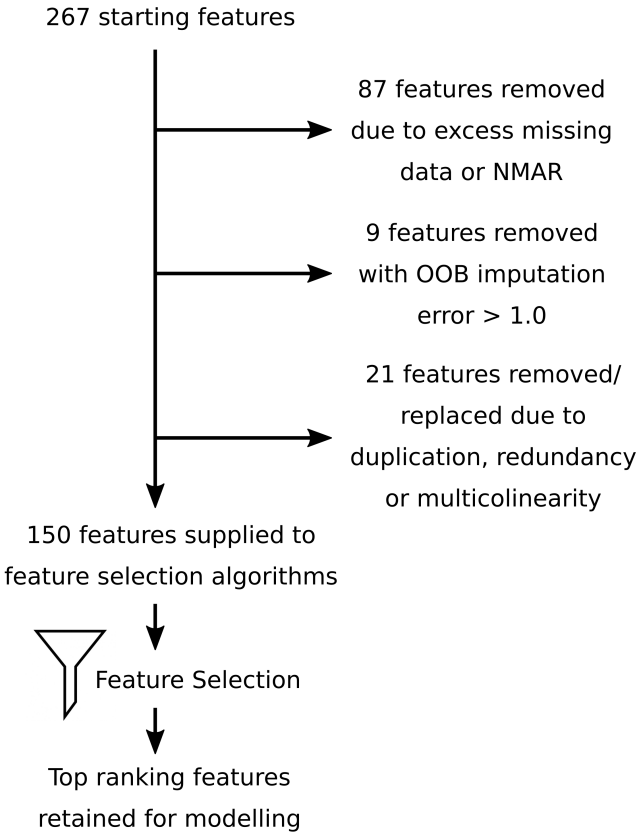

Figure 6.2: Summary of steps taken to reduce the complexity of the feature space prior to development of binary classification models.

6.4 Imputing missing values to maximise available training data

Most patients were missing data for over 50 of the 267 available features (Figure 6.3A). A combination of experimental error, issues of sample integrity, and the sporadic nature of clinical data collection all contributed to missing values, visualised for the entire cohort in the clustered heatmap of Figure 6.3B. The rows of this heatmap represent each unique patient, and the columns the features. Where cells are black, this indicates data are missing. It can be assumed that data are missing at random for most features. For example, only a sample of the cohort was available at the time of lipidomic analysis, some patients were missing data for immunological profiling due to sample integrity or laboratory error, and the technical issues on the ward made point-of-care testing results unavailable for some individuals. Other features present more pressing issues and cannot be assumed to be missing at random. The large cluster of features on the right-hand side of the clustered heatmap of Figure 6.3B (labelled 'Other laboratory measurements') consists of clinical laboratory measurements that have been obtained for only a few individuals. For example, the absence of Vancomycin and Gentamicin serum concentrations is not random and depends on their value. Such variables cannot be imputed with confidence because they are likely not missing at random, or the quantity of missing values is too severe. Therefore, they were excluded from further analysis.

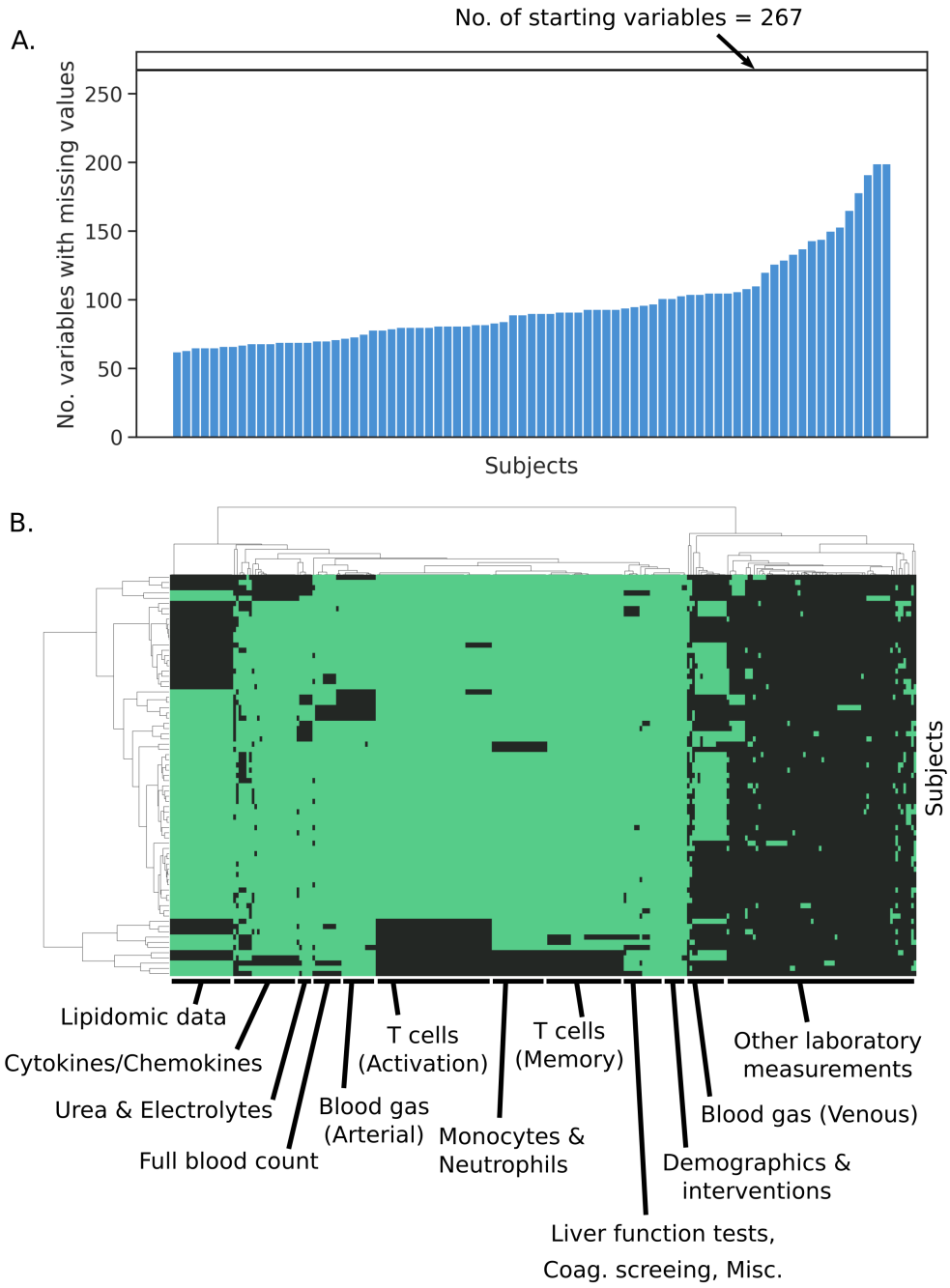

Figure 6.3: Visualisation of missing data in the ILTIS study. (A) Number of variables with a value missing for each patient, out of a total of 267 available variables; shown by the black horizontal line. (B) Clustered heatmap with the 267 available features as columns and each patient represented by a row. A black cell represents the absence of a variable for a given patient. Rows and columns were clustered using Ward hierarchical clustering.

The remaining missing values were imputed using the MissForest algorithm [315], an iterative imputation method that attempts to impute missing values using a series of Random Forest models. The Random Forest algorithm offers many advantages. Firstly, they are non-

parametric; therefore, there are no assumptions about the underlying sample distribution. It also can handle mixed-data types, and because results are derived from the average of many unpruned trees, it constitutes a multiple imputation methodology. Additionally, the out-of-bag (OOB) error estimates of Random Forest allows for the imputation error for each feature to be estimated. The OOB errors are reported as the proportion of incorrect classifications for categorical features. For continuous features, the normalised mean root squared error (NMRSE) is reported, normalised by the variance of each feature.

Another popular alternative for imputing missing values is Multiple Imputation by Chained Equations (MICE). MICE generates multiple complete datasets. Each dataset is exposed to any downstream statistical analysis, and the results are pooled using Rubin rules, capturing the uncertainty introduced by the imputation procedure [316]. The original authors of MissForest demonstrated how, on multiple datasets, their method outperformed MICE in terms of accuracy and computational performance, and OOB error estimates were comparable to the actual error [315]. MissForest also provides a single complete dataset, simplifying downstream analysis.

Figure 6.4A shows the distribution of OOB NMRSE for continuous features, where each dot is an individual feature. An NMRSE of 1.0 or greater indicates that a model is no better than random imputation of feature values. When imputed with the MissForest algorithm, most continuous features had an NMRSE greater than 0.8, and 40 features had an NMRSE greater than 1.0. Furthermore, for some categorical variables OOB estimates showed greater than 30% of values being falsely classified (Figure 6.4B).

In response to this poor performance, other methods were investigated. The MissRanger package [317, 318] in the R programming language is based on the same principles as MissForest but with improved computational performance. The MissRanger package also offers Predictive Mean Matching (PMM). In PMM, the imputed value is matched to the closest value amongst a 'donor pool' of complete data sampled from the original data, preventing imputation with unrealistic values not observed in the original sample distribution and helping increase the variance in the resulting conditional distributions to a realistic level. PMM is advisable for multi-modal or skewed distributions, a characteristic of many features in the ILTIS data. Imputation with MissRanger resulted in improved OOB error for both continuous (Figure 6.4A) and categorical (Figure 6.4B) features; the median NMRSE was reduced

from 0.92 to 0.68 and the percentage of categorical features misclassified reduced to less than 20%.

Features with an NMRSE greater than 1.0 were removed from further analysis, as were features with greater than 40 per cent missing data. A 40 per cent cutoff was chosen over concerns regarding the estimation of OOB where more than half of the observations were missing in the original data. This decision was also supported by a weak correlation (Pearson R^2 of 0.43; p-value < 0.001) observed between NMRSE and the percentage of missing data for each feature (Figure 6.4C), suggesting that the amount of missing data could be detrimental to the accuracy of imputation by MissRanger.

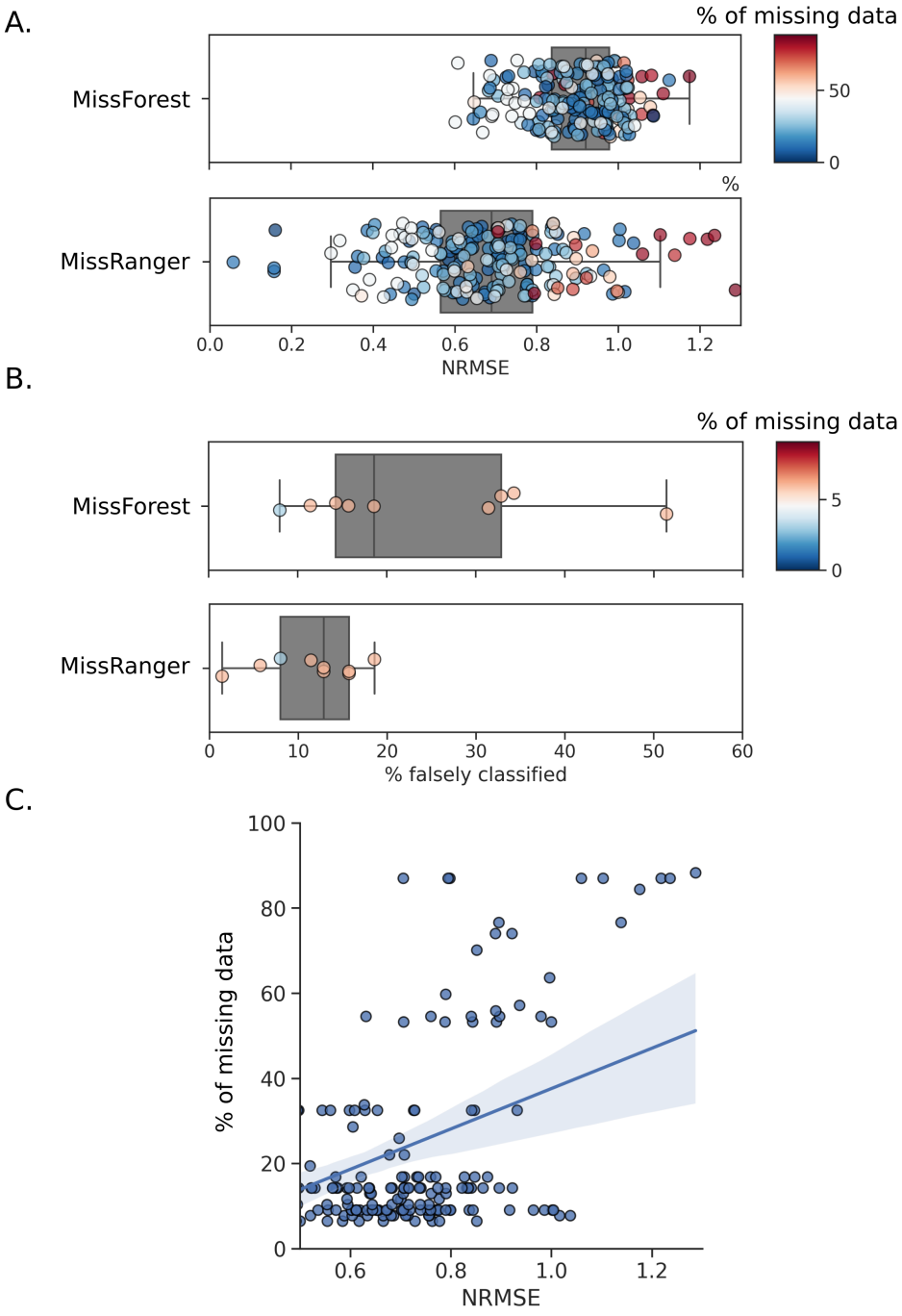

Figure 6.4: Out-of-bag (OOB) imputation error estimates when imputing missing values with MissForest and MissRanger. (A) OOB Normalised Root Mean Squared Error (NMRSE) estimates for continuous features imputed with MissForest (top) and MissRanger (bottom) algorithms. (B) OOB proportion of falsely classified values for categorical features imputed with MissForest (top) and MissRanger (bottom) algorithms. For both A and B, the gradient of colour for each data point represents the percentage of missing values in the original data. (C) The relationship between the OOB NMRSE of continuous features and the percentage of missing values in the original data.

6.5 Multicolinearity

At this stage in the analysis, missing values had been imputed, and features with excessive missing data or unsuitable for imputation were removed (Figure 6.2). A total of 174 features remained for predicting the binary target variables. The number of features at this point still vastly outweighed the number of observations. Multiomic and clinical data have many redundancies, with parameters often correlated. Therefore it was essential first to identify and replace strongly correlated features. Reducing the multicollinearity helped reduce the feature space and was also vital for model interpretation; highly correlated features can result in misinterpretation of feature importance since a feature might be included in a downstream model, not due to its relationship with the target of interest but rather its correlation with some other informative predictor.

The multi-colinearity of the feature space was visualised using a clustered matrix of pairwise Spearman's rank correlation coefficients (Figure 6.5A). There were multiple pairs of heavily correlated features as well as large clusters of both positive and negatively correlated features. A matrix of absolute Spearman's rank correlation coefficient was clustered (Figure 6.5B) to identify problematic groups of features. A total of 15 clusters were identified with one or more pairs of values with an absolute Spearman's rank correlation coefficient greater than or equal to 0.75.

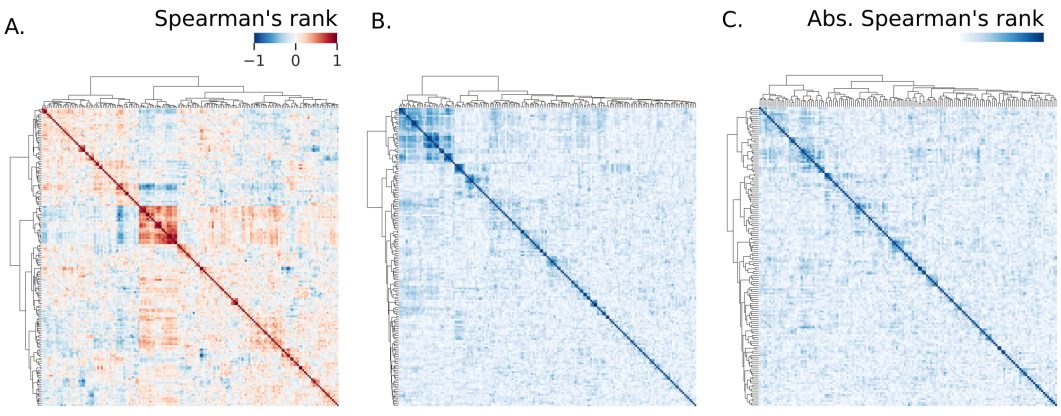

Figure 6.5: Multicolinearity amongst features visualised using pairwise Spearman's rank correlation coefficient. (A) Clustered matrix of pairwise Spearman's rank correlation coefficient shows groups of highly correlated features. The absolute pairwise Spearman's rank correlation coefficient matrix was clustered and is shown before (B) and after (C) removal of redundancies and replacement of highly correlated groups with latent variables.

Lipid data were highly correlated and were grouped into three main clusters with pairs of lipids with an absolute spearman rank of greater than 0.8. Each cluster was replaced with a latent variable generated using PCA. The component generated by PCA for each cluster described greater than 75 per cent of the total variance within each cluster of lipidomic features. Other redundant and duplicated features were identified and removed. Calcium serum level was removed in favour of albumin-adjusted calcium serum levels, considered the more reliable of these two features. Neutrophil count was heavily correlated with white blood cell count; neutrophils constitute the largest proportion of total circulating leukocytes, so it makes sense that a correlation was observed here. The neutrophil count was retained, along with features for other leukocytes, to offer more granularity in the feature space. Lastly, where two variables were heavily correlated, but it was not immediately apparent which variable should be retained, the variable with the greatest mutual information with respect to the target variables was retained. Overall, the collinearity between features was reduced (Figure 6.5C).

6.6 Feature selection

With the removal of redundant and highly correlated features, 150 features remained (Figure 6.2). Of these 150 features, a set had to be chosen small enough to reduce the risk of overfitting, improve classification accuracy [156], and ensure models could be easily interpreted. This problem is known as the 'minimal-optimal' problem [157], is well described, and has influenced the generation of many feature selection algorithms. For this study, five feature selection algorithms were chosen from the literature to obtain a reduced set of informative features for each target variable, resulting in five independent feature sets for each target. Including multiple feature selection algorithms prevented downstream analysis from being biased by one feature selection methodology. The methods were chosen to include filter and wrapper techniques and are popular in the biostatistics literature:

1. Univariate selection with permutation testing: a filter method that ranks features by the inverse of their p-value when testing for a significant difference between the positive and negative case of a target *e.g.* for 30-day mortality, the feature with the smallest p-value when comparing survivors and non-survivors would be the highest-ranking feature. Permutation testing was used for hypothesis testing, a non-parametric test that computes all possible values of the sample mean under possible rearrangements of the observed data using resampling. Due to the computational intensity of this task, an approximation method was used with 1000 rounds of resampling [319].

2. ReliefF: another filtering method, relief-based algorithms (RBAs), have gained popularity for their ability to capture feature dependencies whilst retaining the generalised advantages of filter methods. RBAs are computationally efficient, and selected features do not depend on the assumptions of a chosen model. RBAs generate a proxy statistic for each feature referred to as 'feature weights'. The feature weight scores the relevance of a feature to some target outcome and ranges between -1 and 1. The algorithm cycles through randomly selected training instances and calculates the distance between that instance and all other observations. Two nearest neighbour instances are chosen, one with the same class as the selected training instance (called the 'nearest hit') and one with the opposing class (called the 'nearest miss'). The feature weights are then updated to reflect the distance in the feature value between the nearest hit and

the nearest miss. Features with a different value between the randomly selected training instance and the nearest miss are assigned higher weights and therefore regarded as more informative. Conversely, a feature with values that differ between instances of the same class (the nearest hit) is assigned a lower weight. The ReliefF algorithm is the best-known variant of the RBAs and was adopted in this study. The default parameter for k of 10 (the number of neighbours used when scoring nearest hits and nearest misses) was used, a value based on preliminary empirical testing [320]. The ReliefF algorithm was implemented using the skrebate Python package. [321].

3. Minimum Redundancy - Maximum Relevance (MRMR): unlike other feature selection algorithms, the MRMR algorithm (first introduced by Ding & Peng [322]) focuses on identifying not only the most relevant features but also the minimal-optimal subset in the absence of redundant features. The objective of MRMR is not to identify all the relevant features that individually have predictive power but rather the smallest subset that collectively is most informative. MRMR iterates over each feature, computing the relevance to the target, weighted by the redundancy of the feature relative to all other features. The choice of relevance and redundancy functions varies [323]. However, the simplest form uses the F statistic to measure relevance and the Pearson correlation coefficient to measure redundancy. For categorical variables, mutual information can be used. Given the mixed data types and the likelihood of feature interaction, this study used a random forest classifier for the relevance function [323].

4. Boruta: the Boruta algorithm [324], unlike MRMR, concerns itself with identifying all-relevant features using a wrapper approach built around the popular random forest algorithm. In the Boruta algorithm, each feature is duplicated by a permuted copy called a 'shadow' feature, such that the original feature space is doubled in length. The z-score, calculated by dividing the average OOB accuracy by its standard deviation, captures feature importance. The feature importance between a real feature and its corresponding shadow feature is compared using a two-sided test of equality. Where a feature is found to be significantly greater than its shadow, it is marked as important. Unimportant features are removed, and the process is repeated until all features have been classified or the maximum number of runs is reached. The class imbalance was handled in the random forest model by assigning class weights as described in equation 6.3.

5. Recursive Feature Elimination with Support Vector Machines (RFE-SVM): recursive feature elimination (RFE) is a wrapper feature selection method developed for classification problems involving a small sample size [325]. Although not limited to one classification model in theory, this methodology was developed with and often is coupled to Support Vector Machines (SVM) [326, 327]. An SVM with a linear kernel seeks to separate classes by a linear negative and positive hyperplane. The decision boundary is determined by the instances (known as 'support vectors') on the edge of these hyperplanes. Soft-margin hyperplanes are preferred as they provide a degree of tolerance to the number of support vectors that violate the identified margin. Once the optimal hyperplanes are found, the coordinates of a vector orthogonal to the hyperplane provide the coefficients for each feature. Recursive feature elimination, in this context, is an iterative algorithm that eliminates the feature with the smallest absolute coefficient (and therefore, the lowest contribution to the separation of classes) on each cycle until the desired number of features is generated. Although a linear SVM assumes that the classes are linearly separable and alternatives exist with non-linear kernels [328], a linear SVM with a soft margin was chosen for this study for simplicity and to reduce the chance of overfitting. A linear SVM was reasonable since other feature selection methodologies were also employed that did not make such assumptions of a linear relationship between the target and covariates (*i.e.* ReliefF, MRMR, and Boruta). The class imbalance was handled in the SVM by assigning class weights as described in equation 6.3.

After running each feature selection algorithm for each target variable, the top ten features from each of the five algorithms were chosen to provide five independent feature sets for each target. Several factors influenced the choice of the number of features to retain: artificially reducing the available feature space helps reduce the risk of multicollinearity and prevents the construction of complex models that are less likely to generalise to a broader population. The downstream analysis also included validation of models on equivalent complete case data, and constraining the number of features increases the amount of data available during complete case analysis. Ultimately the objective of this study was to generate interpretable models that identify a minimal feature set that correlates with outcomes and would help generate new hypotheses. Limiting the feature space to a maximum of ten parameters ensures at least five observations are available per feature and makes the interpretation of model decisions easier. Nevertheless, using the top ten features could introduce redundancies. Since the minimal predictive set is the objective, classifiers were generated for the top three features to the top ten features and compared by classification performance (Figure 6.7 describes the entire modelling pipeline).

The overlap between feature selection methods was measured using the pairwise Jaccard index and visualised as a heatmap for each target (Figure 6.6C). A Jaccard index of 1 indicates that both sets are identical. Alternatively, if no common features are shared, the index is 0. The univariate filtering method is most distinct as having virtually no overlap with the feature sets generated by the other four methods, suggesting that the dependencies between features in this data, as opposed to individual predictive power alone, were crucial in identifying the optimal set of predictors. However, the overlap was still remarkably low among the other selection methods, resulting in five distinct feature sets and demonstrating a large variance between different feature selection algorithms when applied to this data. For 90-day mortality, most feature selection algorithms identified distinct sets. Although a moderate overlap of features selected by ReliefF and Boruta is observed, and to a lesser extent, some overlap is seen between Boruta and RFE-SVM. Moderate overlap of features from ReliefF and Boruta algorithms was also observable for the Gram-negative target variable.

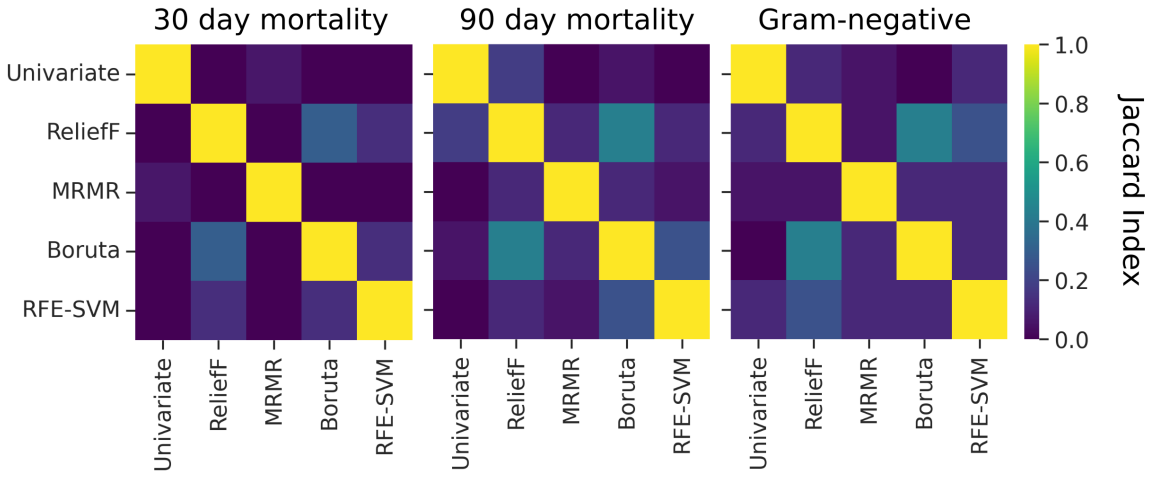

Figure 6.6: Pairwise Jaccard Index measures the overlap of feature sets generated by five independent feature selection algorithms. Feature selection was performed for three binary target variables: mortality 30 days after diagnosis with sepsis (left), 90 days after diagnosis with sepsis (middle), and a Gram-negative causative pathogen amongst those with a positive culture (right).

With a significantly reduced feature space, one last check for multicollinearity was made using the Variance Inflation Factor (VIF). An ordinary least squares regression model was generated for each feature using all other features as predictors. The VIF was then calculated as one divided by one minus the coefficient of determination (R^2) to give the ratio of the overall model variance to the variance of a model that includes only that single feature. A VIF of greater than five was considered to exhibit high multicollinearity. Amongst the five feature sets for each target, only the lipids C10:0 and C8:0 for 30-day mortality were of concern, with a VIF of 16.8 and 17.6, respectively. C8:0 was retained over C10:0 because it displayed more variance, had greater mutual information with 30-day mortality, and had a smaller p-value when comparing class means with a permutation test.

6.7 Multivariate models identify signatures that correlate with outcome and causative pathogen

An overview of the modelling pipeline and evaluation of model performance is provided in Figure 6.7. The first step was to obtain 'holdout' data that would be used to evaluate model performance after model and feature selection (Figure 6.7A). Generation of a holdout set is a crucial methodological decision that is often overlooked in biomarker and proteomics studies [330, 329]. By randomly selecting 20% of the data, and keeping this data independent of model and feature selection, over-fitting and inflated accuracy can be avoided.

Rather than limiting our analysis to a single classifier of interest, multiple classification algorithms were used (Figure 6.7B). The 'no free lunch theorem' presented by Wolpert and Macready [301], perpetuated by recent biomarker studies [46, 331, 332, 333], suggests that no single algorithm can be optimal for all problems. It is, therefore, a requirement to experiment with a range of diverse classifiers. In this study, eight groups of classifiers (referred to as 'classifier families' from here onward) were drawn upon for the task of binary classification:

1. Logistic regression: an extension of linear regression for classification tasks by modelling the probability of the outcome variable using a logistic function. Logistic regression supports lasso, ridge, and elastic net regularisation to reduce overfitting [158, 177].

2. Support vector machine with a linear kernel (Linear SVM): support vector machines (SVMs) encapsulate a popular family of classification algorithms that are proven effective in high dimensional spaces, are memory efficient, and are highly versatile. With a linear kernel, an SVM seeks to optimise a margin with two linear hyperplanes that separate the positive and negative class [158, 334].

3. Support vector machine with a non-linear kernel (Non-linear SVM): SVMs can be extended to perform non-linear classification by mapping samples onto a high-dimensional feature space in which linear classification is possible; this is known as the 'kernel trick'. Many kernel functions exist, but popular solutions include the polynomial kernel, the Gaussian radial basis function kernel, and the Sigmoid kernel [158, 334].

4. Naive Bayes: a family of supervised learning algorithms successfully applied in previous biomarker studies [173, 335]. Naive Bayes is based on Bayes' theorem but adopts the "naive" assumption that every feature is independent of all other features given the value of the target variable [177]. In this study, a Gaussian Naive Bayes algorithm was implemented, where the likelihood of the features is assumed to follow a Gaussian distribution.

5. K-Nearest Neighbours (KNN): neighbours-based classification is a type of instance-based learning because it does not construct a general model but rather stores training data instances. Classification results from a simple majority vote of the nearest neighbours of each point $i.e.$ a newly encountered observation are assigned the class with the most representation amongst k nearest neighbours. Therefore, this algorithm's performance is influenced by two hyperparameters, the choice of k and the distance metric used to construct the nearest-neighbour tree [334].

6. Random Forest: the Random Forest algorithm is a popular ensemble technique to improve the classification performance of decision trees. A perturb-and-combine strategy is employed to generate randomised decision trees, creating a diverse 'forest' of decision tree classifiers. The resulting prediction is drawn from the average of the individual classifiers[158, 334].

7. Extra Random Forest: extremely randomised or 'extra' Random Forest introduces additional randomness to the construction of decision trees. As with the random forest algorithm, a subset of features is used when constructing each decision tree. However, instead of splitting on the most discriminative thresholds, thresholds are drawn randomly for each candidate feature. The best randomly generated threshold is picked as the splitting rule, helping reduce the model variance at a slight expense of increased bias [177].

8. Extreme Gradient Boosting (XGBoost): the XGBoost algorithm is an extension of the gradient-boosted trees algorithm. Unlike a Random Forest approach, gradient boosting algorithms use successive weak learners to solve the classification problem. With each weak learner, more weight is put on instances that previous weak learners struggled to classify. Predictions are generated by a majority vote across the weak learners weighted by their accuracy [178].

Classifiers were chosen to include simple linear models ($i.e.$ Logistic regression and SVM with a linear kernel) and models with greater complexity that would capture non-linear relationships between the target and covariates ($i.e.$ SVM with a non-linear kernel, KNN, and ensembles of trees). All classifiers were implemented using the Scikit-Learn library [177].

Within each classifier family, a vast array of hyperparameters influence their behaviour. A grid search strategy was employed to tune optimal hyperparameters. Hyperparameters included L1 and L2 regularisation of varying strengths, polynomial and radial basis function kernels for non-linear SVMs (with multiple degrees for the former and a range of γ for the latter), different distance metrics and number of nearest neighbours for KNN, and multiple hyperparameters for ensembles of tree-based learners controlling parameters such as the depth of trees, number of splits, number of features, and sampling methods. The number of models derived from each family is shown in Figure 6.7B and resulted in 216 classification models.

Each model was trained on the 5 independent feature sets described in section 6.6 (Figure 6.7C). As previously discussed, the optimal number of features might be less than the top 10 ranked features presented by each feature selection algorithm. Therefore, iteratively, each classifier was trained on the top 3 through to the top 10 features. For each classifier, eight models were generated for each feature selection algorithm, totalling 40 models across all possible feature sets and 8,880 models trained across all possible classifiers, repeated for each target.

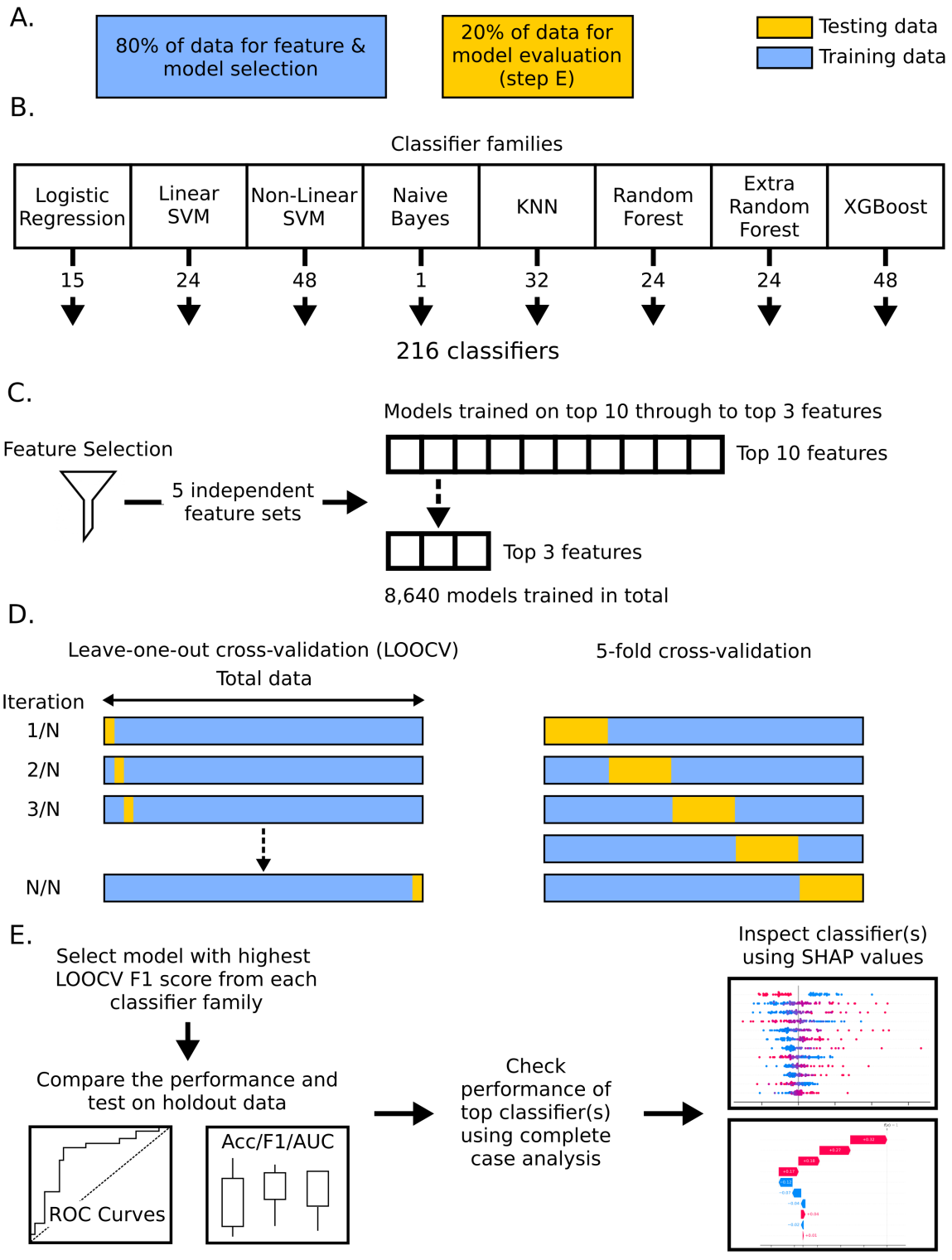

Figure 6.7: (Continued on the following page.)

Figure 6.7: Schematic of the modelling pipeline for selecting, comparing, and inspecting classification algorithms. Data were split (A) into training and holdout sets, retaining 20% of data for model evaluation. Data were exposed to eight classifier families (B), generating 216 classifiers in total after including multiple hyperparameters. Models were trained using five independent feature sets (C) iteratively on the top 3 through to the top 10 features from each feature set, producing 40 models for each feature set and 8,880 models for each target variable. Models were trained and tested using leave-one-out and 5-fold cross-validation (D), and the best performing model (measured by LOOCV macro F1 score) was selected for each classifier family. Cross-validation performance and performance on the holdout set were compared across classifiers using ROC curves, balanced accuracy, macro F1 score, and macro AUC score (E). The top performing classifier(s) were tested against their equivalent complete case data as an additional validation step before inspection of model decisions using SHapely Additive exPlanations (SHAP) values.

Cross-validation was used to select the optimal model before validation on independent hold-out data. Two cross-validation procedures were performed in parallel (Figure 6.7D). The cross-validation methods were chosen due to the small amount of training data available:

1. Leave-one-out cross-validation (LOOCV): a form of cross-validation where the number of folds equals the number of observations in the data. Therefore, on each fold, a single observation is kept out for testing, and all other observations are used as training data. Data splitting is repeated until every observation has been tested. LOOCV maximises the available number of training instances across folds and is appealing when the total sample size is small. The consequence is that the individual folds are very similar, resulting in performance estimates with low bias but possibly greater variance.

2. 5-fold cross-validation (5-fold CV): given that the condition of sepsis and the data obtained by ILTIS exhibit heterogeneity, the impact of resampling within the cohort on model performance was of interest. Therefore, stratified 5-fold cross-validation was used, splitting the data into five independent non-overlapping sets, and iterating over them using one set each round as testing data.

Within each classifier family, the model and feature set combined with the highest LOOCV F1 score was chosen for evaluation and model inspection (Figure 6.7E). The cross-validation and holdout performance for the optimal model from each classifier family were compared using ROC curves, balanced accuracy, macro F1 score, and macro AUC score. Models were first compared by 5-fold CV balanced accuracy using the non-parametric Friedman test, and Nemenyi post-hoc testing [336] to test whether the variation in performance across 5-folds

was significantly different between models. After selecting a model from each classifier family using cross-validation, their performance was validated on independent holdout data, and the best-performing model(s) were selected for inspection.

Before inspecting model(s) decisions, the performance of the chosen model(s) was compared to the equivalent complete case data. For practical reasons, imputation was performed before the generation of holdout data and is a possible source of data leakage. Additionally, imputation introduces additional error, as quantified by OOB estimates and discussed in section 6.4. Therefore, testing on complete case data in the reduced feature space serves as additional validation to offset these redundancies. Since complete case data substantially reduces the number of training instances, only LOOCV was used in the complete case analysis.

Finally, the top performing models were inspected using SHapely Additive exPlanations (SHAP) [167]. The best way to explain SHAP values is by using a simple example. Since, in this study, linear models were employed, we can use the simple example of explaining a Logistic Regression model tasked with predicting 30-day mortality (Figure 6.8). Traditionally, the standard way of interpreting a linear model is to examine the coefficients learned for each feature. Whilst these are informative, linear and non-linear models are compared in this study, and therefore a model-agnostic method for interpretation was desired.

To understand SHAP values, let us first try to understand how changing the value of a feature, such as the percentage of T cells, affects the model output. Figure 6.8A shows a partial dependency plot with T cells as a percentage of PBMCs on the x-axis and the expected value of the model given the percentage of T cells on the y-axis. The distribution of the percentage of T cells observed in the data is also visible as a histogram. The horizontal grey dotted line shows us the baseline model's expected value (equivalent to the observed mortality rate), and the vertical grey dotted line shows the average value of our distribution. The blue line describes the partial dependence and passes through the intersection of these two grey dotted lines, known as the centre of the partial dependence plot, with respect to the observed data distribution. Notice that as the percentage of T cells increases, the expected value with respect to the percentage of T cells decreases, and therefore the likelihood of mortality decreases.

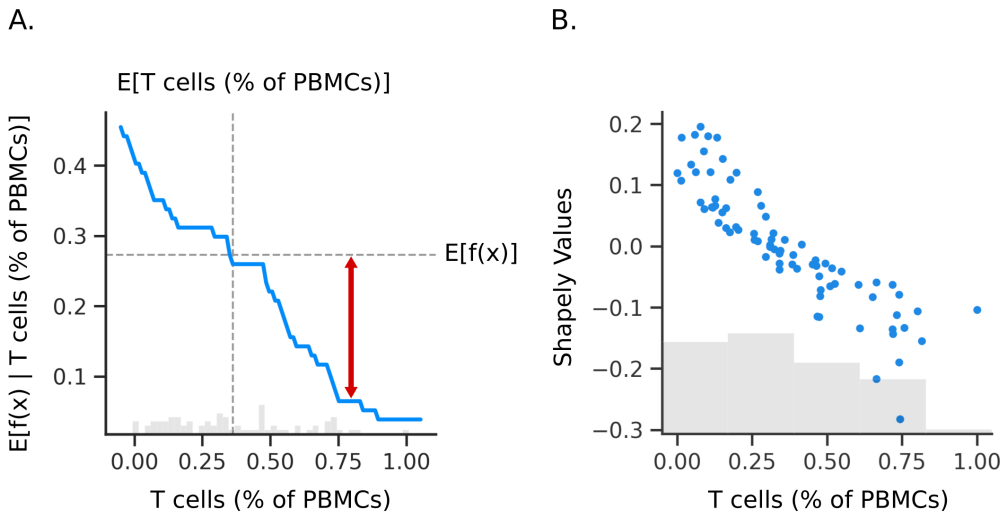

Figure 6.8: Partial dependency plot (A) shows the relationship between T cells (% of PBMCs) and the outcome of a Logistic Regression model. The red arrow shows the difference between the expected value given the proportion of T cells (% of PBMCs) and the baseline expected value for a single observation of T cells; for a linear model, this is equivalent to the SHAP value for this observation. The SHAP values for all observations are plotted as a scatterplot (B) where the same relationship as the partial dependency plot is observed.

SHAP values use cooperative game theory to allocate a score to each feature that reflects their contribution to a model's output. For a linear model, the SHAP values of a feature can be calculated by simply comparing the difference between the expected model output and the partial dependence plot at a feature value x_i (demonstrated by the red arrow in Figure 6.8A). If we plot the SHAP values across the entire data for a feature, the values closely follow the relationship shown by the partial dependency plot (Figure 6.8B).

A fundamental property is that the SHAP values of all input features will always sum to the difference between the expected model output (known as the baseline) and the observed output for a prediction. We can therefore use SHAP values to visualise the impact of features on the predictions of a machine learning model. Figure 6.9A and B show two observations predicted to be a member of the negative and positive class by the same Logistic Regression algorithm, respectively. The plots shown are known as 'waterfall plots' and start with the prior expected output of our model $E[f(X)]$ and add feature (y-axis) SHAP values until it reaches the observed outcome.

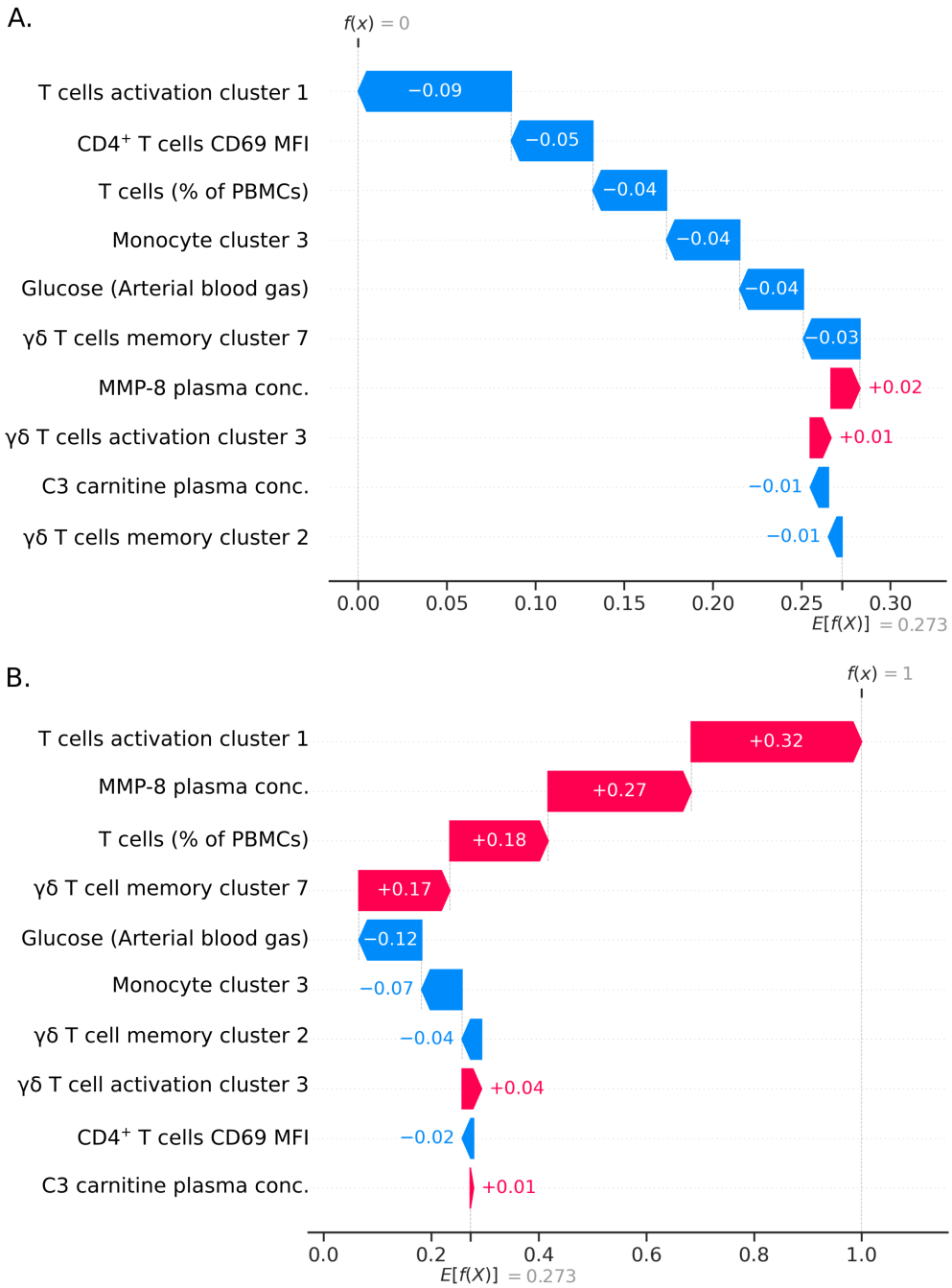

Figure 6.9: Waterfall plots for a negative (A) and positive (B) prediction by a Logistic Regression model predicting 30-day mortality. The baseline expected value (mortality rate observed in the training data; 27.3%) is shown as a grey horizontal dotted line. Features, shown on the y-axis, have increasing SHAP values from the bottom of the axis to the top, and their sum equals the difference between the baseline expected values and the predicted value.

6.7.1 A T cell dominant signature predict mortality at 90 days after diagnosis of sepsis.

The first target variable investigated was 30-day mortality. There was no significant difference (p=0.303) between the optimal models chosen for each classifier family when comparing 5-fold CV accuracy and F1 score (Figure 6.10A). Logistic regression and SVM's showed promise at first, with impressive LOOCV ROC curves (Figure 6.10B) and comparable performance between training and testing data within the model and feature selection process (Figure 6.10C). However, performance on holdout data was poor. None of the chosen models performed better than a random baseline (represented by the dotted diagonal line, Figure 6.10C) and individual performance metrics suggested that none of the models generalised well when exposed to holdout data (Table 6.1).

	Balanced Accuracy	Macro F1 Score	Macro AUC Score
Logistic regression - RFE-SVM - top 10	0.58 [0.50 – 0.67]	0.56 [0.49 – 0.60]	0.49 [0.44 – 0.58]
Linear SVM - RFE-SVM - top 10	0.59 [0.5 – 0.67]	0.57 [0.49 – 0.60]	0.52 [0.42 – 0.61]
SVM (cubic polynomial) - RFE-SVM - top 10	**0.63 [0.54 – 0.71]**	0.62 [0.53 – 0.66]	0.59 [0.53 – 0.64]
KNN - Boruta - top 7	0.63 [0.63 – 0.67]	**0.65 [0.64 – 0.712]**	0.68 [0.64 – 0.72]
Naive Bayes - RFE-SVM - top 9	0.50 [0.38 – 0.54]	0.50 [0.38 – 0.53]	0.52 [0.44 – 0.57]
Random Forest - Boruta - top 6	0.46 [0.46 – 0.50]	0.41 [0.40 – 0.42]	0.59 [0.55 – 0.64]
Extra Random Forest - Boruta - top 7	0.54 [0.42 – 0.58]	0.54 [0.40 – 0.58]	**0.73 [0.64 – 0.83]**
XGBoost - Boruta - top 7	0.50 [0.38 – 0.54]	0.50 [0.38 – 0.53]	0.48 [0.31 – 0.56]

Table 6.1: Holdout performance for the top-performing model selected within each classifier family for predicting 30-day mortality in sepsis. Each model is presented as the name of the classifier family, the feature selection method that generated the optimal feature set, and the number of features selected for the top-performing model. The highest ranking metrics are highlighted in bold font. Bootstrapped 95% confidence intervals are shown in square brackets, generated using 100 rounds of resampling.

Prediction of 90-day mortality was more reliable (Figure 6.11). Although a comparison of 5-fold CV accuracy and F1 score (Figure 6.11A) showed a significant difference in performance when applying Friedman's test, posthoc pairwise Nemenyi testing did not yield

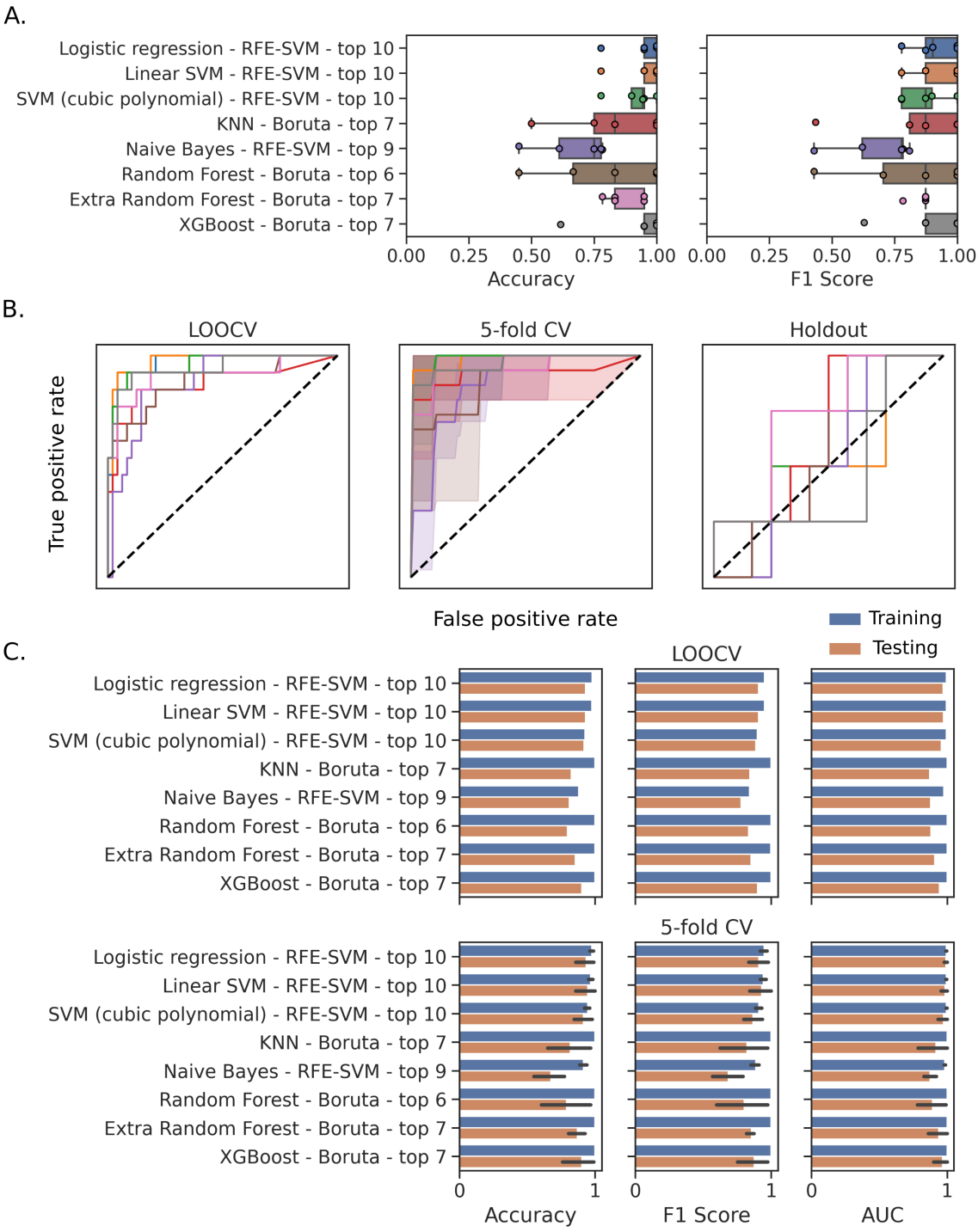

Figure 6.10: (Continued on the following page.)

Figure 6.10: Cross-validation and holdout performance for the top-performing model selected within each classifier family for predicting 30-day mortality in sepsis. Each model is presented as the name of the classifier family, the feature selection method that generated the optimal feature set, and the number of features selected for the top-performing model. (A) Balanced accuracy and macro F1 score for each fold of 5-fold cross-validation (5-fold CV) are shown as boxplots. Models receiver-operating-characteristic (ROC) curves are presented (B) for leave-one-out cross-validation (LOOCV), 5-fold CV, and testing on holdout data. The dotted diagonal line represents a model with a random performance level. The difference in training and testing performance within each cross-validation procedure is shown (C), where error bars for 5-fold CV represent 95% bootstrap confidence intervals with 1000 rounds of resampling.

significant p-values for any comparisons; the lowest p-value was 0.101 for the comparison of SVM (quartic polynomial) and Random Forest. Therefore, we cannot conclude that the models were significantly different. However, high 5-fold CV accuracy and F1 score were observed for many models, with lower variance across folds compared to models predicting 30-day mortality. The LOOCV and 5-fold CV performance was generally good across all classifiers (Figure 6.11B and C). However, more complex models such as KNN and ensembles of tree-based learners exhibited more over-fitting compared to the simpler Logistic Regression and Linear SVM, except for the Extra Random Forest model. The Extra Random Forest model showed superior accuracy, F1 score, and AUC scores compared to all other models when tested on holdout data (Table 6.2) and was chosen for complete case analysis and inspection.

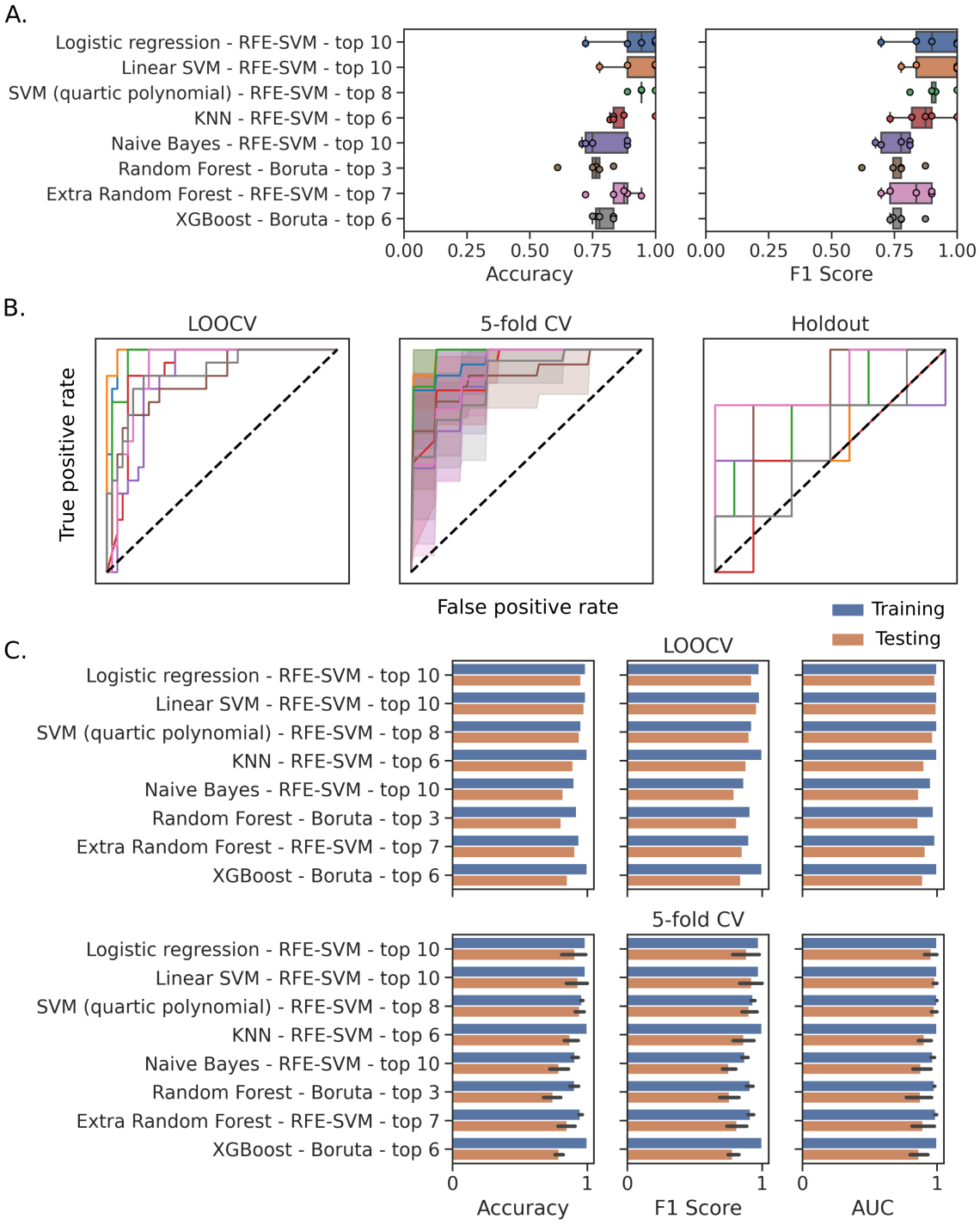

Figure 6.11: (Continued on the following page.)

Figure 6.11: Cross-validation and holdout performance for the top-performing model selected within each classifier family for predicting 90-day mortality in sepsis. Each model is presented as the name of the classifier family, the feature selection method that generated the optimal feature set, and the number of features selected for the top-performing model. (A) Balanced accuracy and macro F1 score for each fold of 5-fold cross-validation (5-fold CV) are shown as boxplots. Models receiver-operating-characteristic (ROC) curves are presented (B) for leave-one-out cross-validation (LOOCV), 5-fold CV, and testing on holdout data. The dotted diagonal line represents a model with a random performance level. The difference in training and testing performance within each cross-validation procedure is shown (C), where error bars for 5-fold CV represent 95% bootstrap confidence intervals with 1000 rounds of resampling.

	Balanced Accuracy	Macro F1 Score	Macro AUC Score
Logistic regression - RFE-SVM - top 10	0.71 [0.63 – 0.79]	0.73 [0.64 – 0.79]	0.69 [0.58 – 0.78]
Linear SVM - RFE-SVM - top 10	0.71 [0.63 – 0.79]	0.73 [0.64 – 0.79]	0.67 [0.58 – 0.78]
SVM (quartic polynomial) - RFE-SVM - top 8	0.70 [0.63 – 0.79]	0.72 [0.64 – 0.79]	0.72 [0.67 – 0.77]
KNN - RFE-SVM - top 6	0.54 [0.42 – 0.58]	0.54 [0.40 – 0.58]	0.54 [0.44 – 0.64]
Naive Bayes - RFE-SVM - top 10	0.71 [0.63 – 0.79]	0.73 [0.64 – 0.79]	0.71 [0.61 – 0.94]
Random Forest - Boruta - top 3	0.67 [0.58 – 0.75]	0.67 [0.58 – 0.72]	0.80 [0.77 – 0.89]
Extra Random Forest - RFE-SVM - top 7	**0.75 [0.67 – 0.83]**	**0.79 [0.71 – 0.88]**	**0.85 [0.81 – 0.86]**
XGBoost - Boruta - top 6	0.54 [0.42 – 0.58]	0.54 [0.4 – 0.58]	0.58 [0.44 – 0.72]

Table 6.2: Holdout performance for the top-performing model selected within each classifier family for predicting 90-day mortality in sepsis. Each model is presented as the name of the classifier family, the feature selection method that generated the optimal feature set, and the number of features selected for the top-performing model. The highest ranking metrics are highlighted in bold font. Bootstrapped 95% confidence intervals are shown in square brackets, generated using 100 rounds of resampling.

Figure 6.12 shows the LOOCV performance of the Extra Random Forest model when exposed to complete case data. The ROC curve for the imputed data was comparable to that of the complete case data, and the training LOOCV AUC was almost identical to the complete case AUC. Balanced accuracy and macro F1 score were decreased in complete case analysis compared to the training LOOCV scores, but both scores were still greater than 0.7.

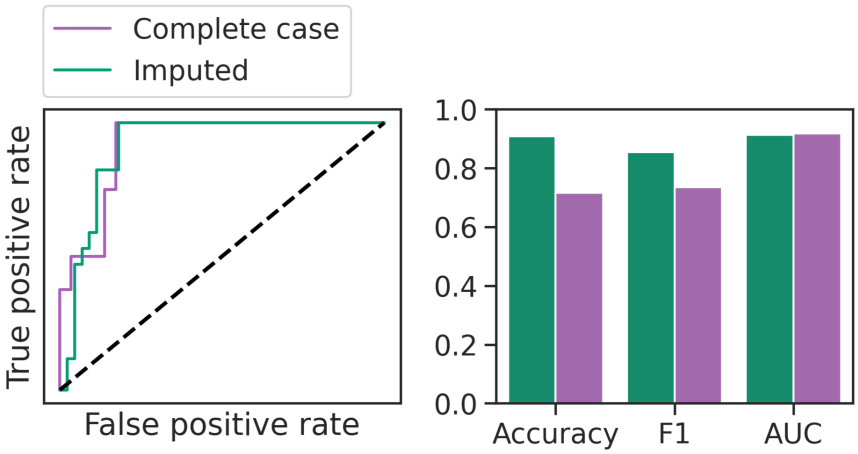

Figure 6.12: Complete case analysis for an Extra Random Forest model tasked with predicting 90-day mortality in sepsis. Performance is documented by a receiver-operating-characteristic (ROC) curve (left) and a bar plot (right) showing balanced accuracy, macro F1 score, and ROC area-under-curve (AUC) score. The dotted diagonal line accompanying the ROC curves represents a model with a random performance level.

The chosen features and their contribution to the Extra Random Forest are shown in Figure 6.13. The beeswarm plot (Figure 6.13, top) shows the features ranked from the most impactful on the model outcome to the least impactful. Each data point is a patient and is coloured according to the value of the respective feature. The x-axis shows the SHAP value, with greater values meaning a higher impact on a model's prediction of death 90-days after sepsis diagnosis. Lower SHAP values indicate that the feature had a higher impact on a model's survival prediction for that patient. The beeswarm plot is accompanied by a heatmap (Figure 6.13, bottom) with patients on the x-axis, the model features on the y-axis, and a bar plot on the right-hand y-axis showing the overall impact of the feature on model output. The individual cells of the heatmap show the SHAP values for each feature for a single patient. Patients (columns) are clustered by their explanation similarity, providing insight into what combination of features drives model predictions. Above the heatmap, the actual outcome of the patient (orange line) and the predicted outcome (black line) are shown as a line plot. It should be noted that the predictions reported here reflect performance on the complete training data and do not reflect how the model would perform when exposed to new data. The purpose of the heatmap is to understand the decision function learnt by the algorithm when exposed to training data.

The proportion of T cells (as a percentage of total PBMCs) was the most noteworthy feature of the Extra Random Forest model. Lower values for T cells influenced a prediction of 90-day mortality, as shown by the gradient for T cells on the beeswarm plot in Figure 6.13. Excluding the percentage of T cells, the most impactful features were: blood glucose, CXCR3 expression on $CD4^+$ T cells, and plasma concentration of Arachidonic acid (a 20-carbon chain polyunsaturated omega-6 fatty acid; C20:4). Increased levels of both blood glucose and $CD4^+$ T cell CXCR3 expression encouraged the model to predict 90-day mortality, whereas the inverse was true for Arachidonic acid, with lower values driving prediction of 90-day mortality. The percentage of T cells was the dominant factor in the Extra Random Forest model, but where the SHAP values were only moderately high, it appears that the influence of blood glucose, $CD4^+$ T cell CXCR3 expression, and Arachidonic acid encouraged the prediction of survival (Figure 6.13 heatmap). The remaining features in the Extra Random Forest model, magnesium plasma concentration, APACHE II Score, and CD25 expression on MAIT cells appear important for individuals rather than the wider training cohort.

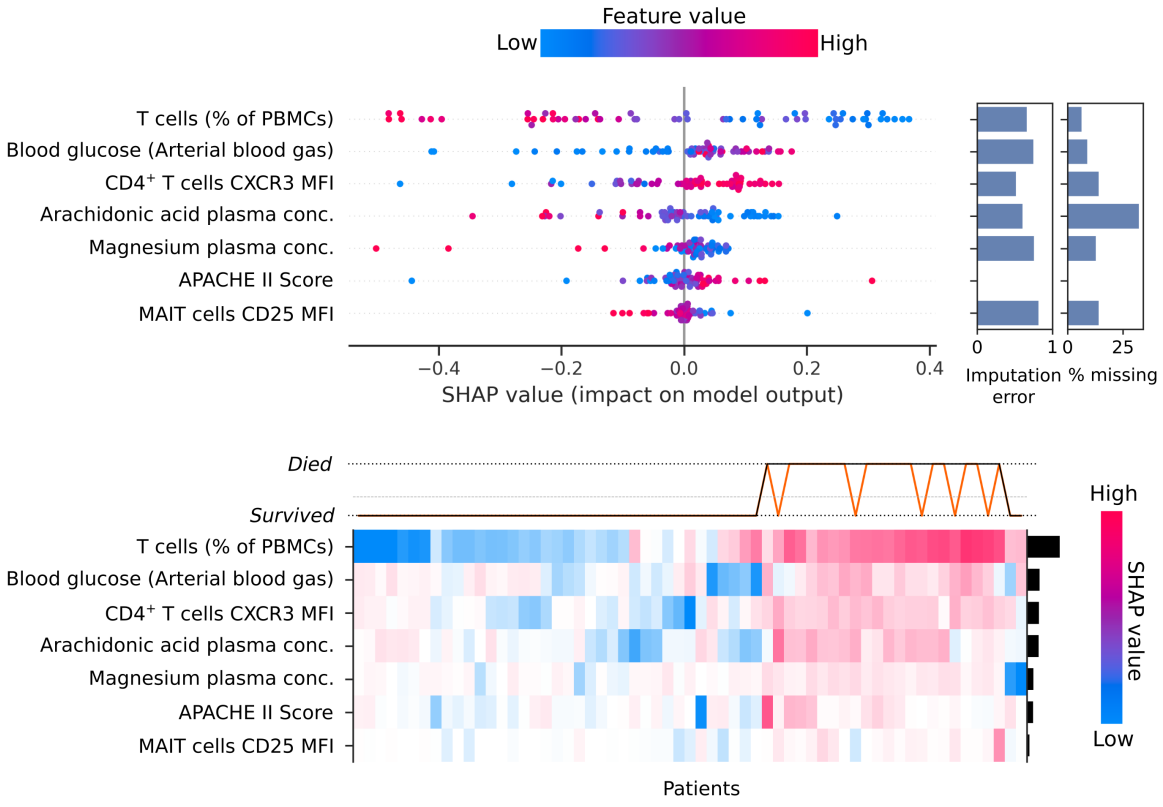

Figure 6.13: SHAP (SHapely Additive exPlanations) values for an Extra Random Forest model tasked with predicting mortality at 90 days after diagnosis with sepsis. The beeswarm plot (top) shows each observation as a single data point coloured by the value of the feature for that instance. On the x-axis is the SHAP value, a lower value corresponds to an instance having a more significant impact on the negative case for the model (i.e. prediction of survival), and a positive value corresponds to having a more significant impact on the positive case for the model (i.e. prediction of death). A barplot on the right-hand side of the beeswarm plot shows the imputation error (with a maximum value of 1) and the percentage of missing values observed in the original data. The heatmap (bottom) shows the SHAP values for each patient. The bar plot on the right-hand y-axis shows each feature's mean absolute SHAP value and measures the impact of a feature on model prediction. The line plot above the heatmap shares the x-axis and displays each patient's predicted outcome (black line) and the actual outcome (orange line). The dotted line between the possible outcomes is the expected value, equivalent to the observed mortality.

6.7.2 Neutrophils, CD8$^+$ T cells, and unconventional T cells form a predictive signature that differentiates Gram-negative and Gram-positive infection in sepsis.

The top-performing models for predicting Gram-negative cause in sepsis are shown in Figure 6.14. There was no significant difference when comparing the model's 5-fold CV balanced accuracy and macro F1 score. However, the median 5-fold CV balanced accuracy for Logistic Regression, and Linear SVM was superior to other models (Figure 6.14A). Logistic regression, SVM's, and the Extra Random Forest model demonstrated the best LOOCV performance (Figure 6.14B and C). The Logistic regression and Linear SVM models performed well on holdout data, each with a balanced accuracy of 0.83 and a macro F1 score of 0.86, but the Logistic regression model outperformed the Linear SVM in terms of ROC AUC score (Table 6.3). The performance of the Extra Random Forest model deteriorated when exposed to holdout data, however, the Random Forest model presented the best ROC AUC score overall. The Logistic regression model and Random Forest were chosen for complete case analysis. Both retained the same LOOCV performance on complete case data observed on the imputed training data (Figure 6.15).

Four features appeared dominant in the Logistic regression model (Figure 6.16): T cells activated cluster 1 (the largest of the CD8$^+$ T cell clusters and characterised by low expression of the activation markers CD69, CD25, HLA-DR, and CXCR3), neutrophil cluster 1 (characterised by its low expression of CD62L), neutrophil count, and sodium concentration in plasma. Higher values of T cells activated cluster 1 and neutrophil cluster 1 encouraged the prediction of Gram-positive cause. In contrast, higher values for the neutrophil count and sodium concentration in plasma influenced a prediction of Gram-negative cause. T cells activated cluster 1 was calculated as a proportion of total T cells and could have possibly been a surrogate marker for the percentage of CD8$^+$ T cells. This feature was replaced with the percentage of CD8$^+$ T cells, and the training and holdout data performance was unchanged. Therefore, T cells activated cluster 1 should not be judged on the expression profile but instead treated as an indication of the effect the proportion of CD8$^+$ T cells had on predicting Gram-negative cause.

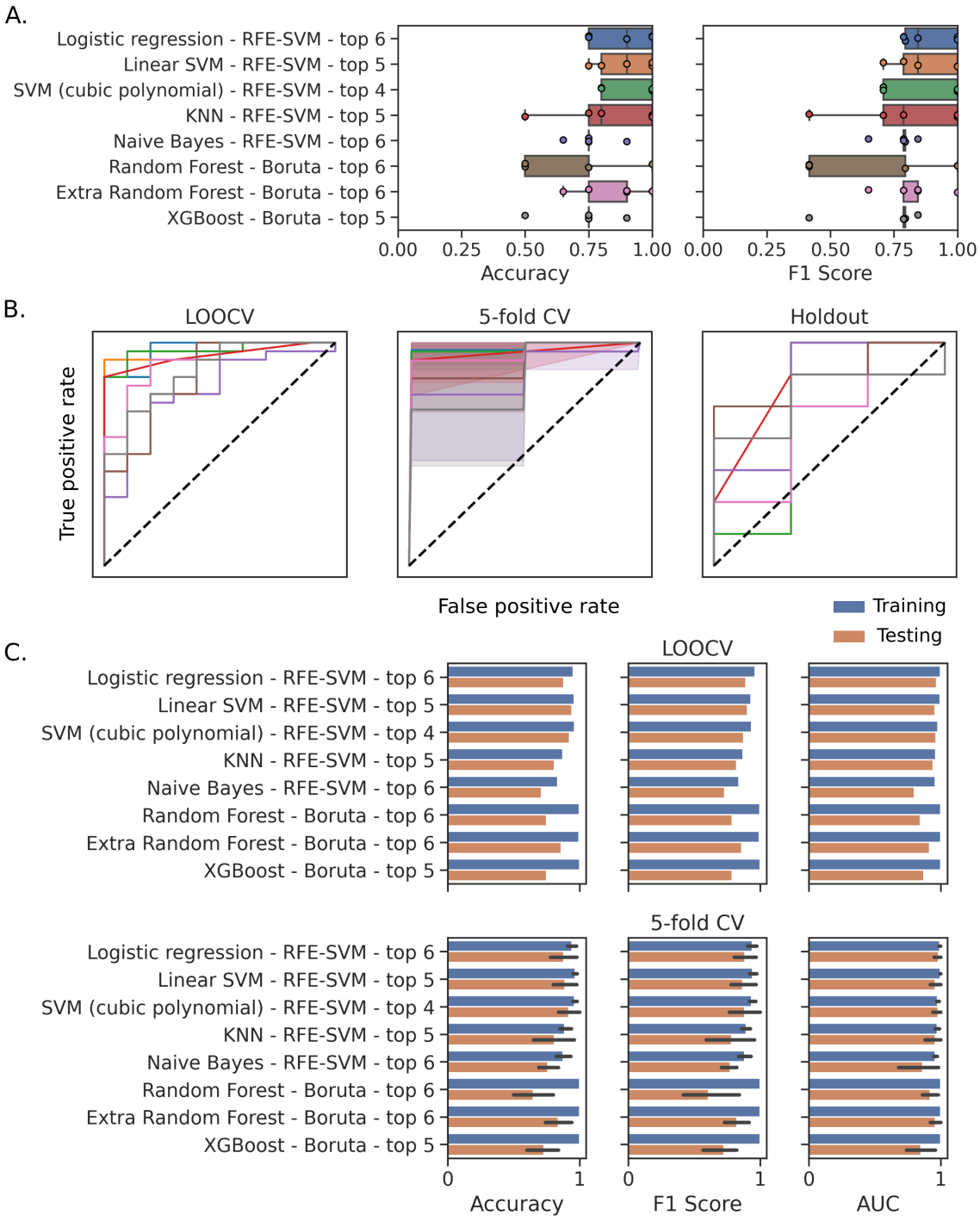

Figure 6.14: (Continued on the following page.)

Figure 6.14: Cross-validation and holdout performance for the top-performing model selected within each classifier family for predicting Gram-negative cause in sepsis. Each model is presented as the name of the classifier family, the feature selection method that generated the optimal feature set, and the number of features selected for the top-performing model. (A) Balanced accuracy and macro F1 score for each fold of 5-fold cross-validation (5-fold CV) are shown as boxplots. Models receiver-operating-characteristic (ROC) curves are presented (B) for leave-one-out cross-validation (LOOCV), 5-fold CV, and testing on holdout data. The dotted diagonal line represents a model with a random performance level. The difference in training and testing performance within each cross-validation procedure is shown (C), where error bars for 5-fold CV represent 95% bootstrap confidence intervals with 1000 rounds of resampling.

	Balanced Accuracy	Macro F1 Score	Macro AUC Score
Logistic regression - RFE-SVM - top 6	**0.83 [0.75 – 1.0]**	**0.86 [0.80 – 1.0]**	0.76 [0.64 – 1.0]
Linear SVM - RFE-SVM - top 5	0.83 [0.75 – 1.0]	0.86 [0.8 – 1.0]	0.71 [0.57 – 1.0]
SVM (cubic polynomial) - RFE-SVM - top 5	0.76 [0.68 – 0.93]	0.76 [0.68 – 0.86]	0.71 [0.57 – 0.92]
KNN - RFE-SVM - top 5	0.60 [0.43 – 0.68]	0.59 [0.4 – 0.68]	0.81 [0.71 – 0.93]
Naive Bayes - RFE-SVM - top 6	0.66 [0.50 – 0.75]	0.67 [0.44 – 0.80]	0.81 [0.71 – 1.0]
Random Forest - Boruta - top 6	0.67 [0.50 – 0.75]	0.68 [0.44 – 0.80]	**0.86 [0.79 – 0.94]**
Extra Random Forest - Boruta - top 6	0.69 [0.61 – 0.86]	0.67 [0.59 – 0.75]	0.62 [0.50 – 0.79]
XGBoost - Boruta - top 5	0.60 [0.43 – 0.68]	0.60 [0.4 – 0.68]	0.76 [0.71 – 0.89]

Table 6.3: Holdout performance for the top-performing model selected within each classifier family for predicting Gram-negative cause in sepsis. Each model is presented as the name of the classifier family, the feature selection method that generated the optimal feature set, and the number of features selected for the top-performing model. The highest ranking metrics are highlighted in bold font. Bootstrapped 95% confidence intervals are shown in square brackets, generated using 100 rounds of resampling.

The other two features were T cells memory cluster 3 (a $CD8^+$ cluster with a TEMRA phenotype of $CD45RA^{hi}$ $CD27^{lo}$ $CCR7^{lo}$ $CD57^{hi}$) and HLA-DR expression on MAIT cells, with high values for both encouraging a prediction of Gram-negative cause. The heatmap of SHAP values for the Logistic regression model in Figure 6.16 shows how many features exert a conflicting influence on the model output. Although T cells activated cluster 1 has the highest absolute mean SHAP value, we can see that other features drive the decision with confounding SHAP values for certain patients. For example, on the left-hand side of

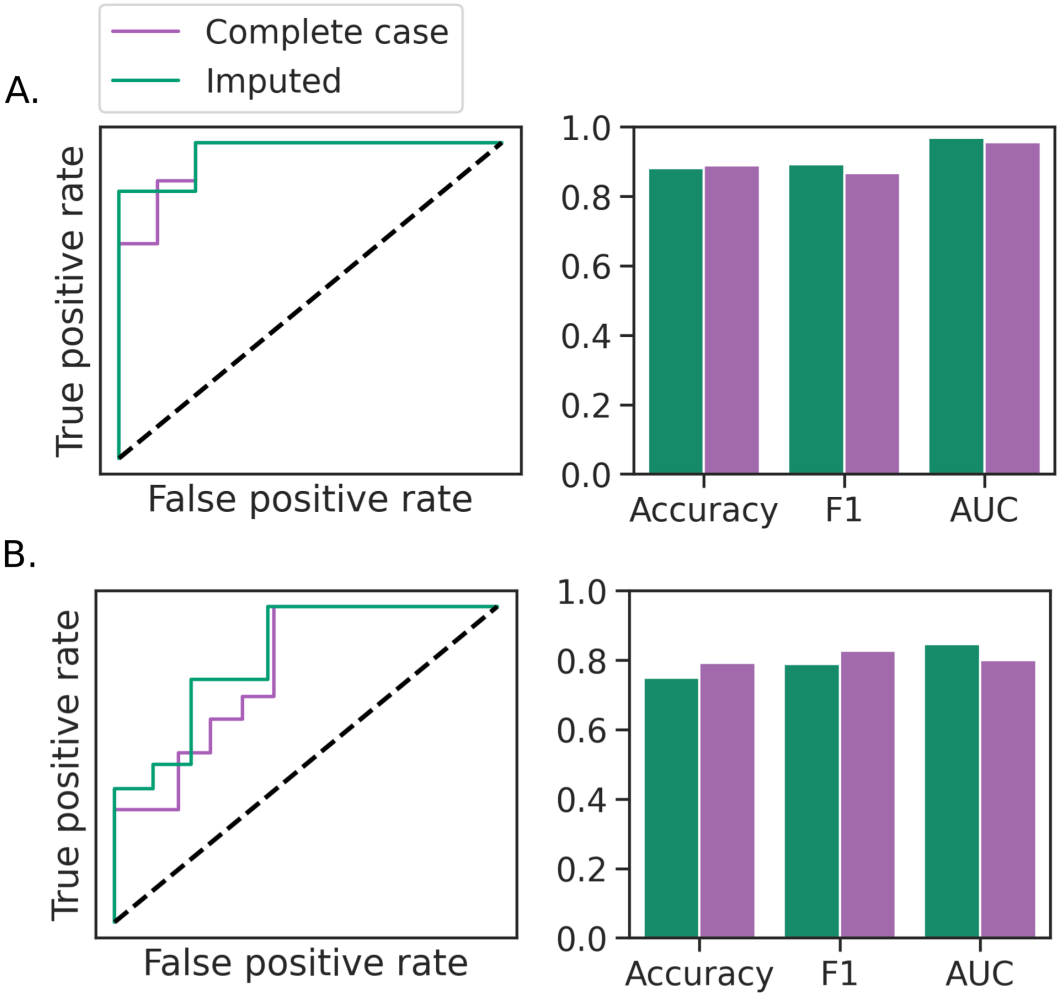

Figure 6.15: Complete case analysis for a Logistic regression model (A) and a Random Forest model (B) tasked with predicting Gram-negative cause in sepsis. Performance is documented by a receiver-operating-characteristic (ROC) curve (left) and a bar plot (right) showing balanced accuracy, macro F1 score, and ROC area-under-curve (AUC) score. The dotted diagonal line accompanying the ROC curves represents a model with a random performance level.

the heatmap, a cluster of patients with a gradient of low SHAP values for T cells activated cluster 1 has moderate to high values for neutrophil cluster 1, neutrophil count, and T cells memory cluster 3. Small clusters of patients have similar SHAP profiles, highlighting the training data's complexity and heterogeneity, hence the need for a multi-parameter model.

The Random Forest model selected features of T cells and the neutrophil count (Figure 6.17). The feature with the highest absolute mean SHAP value was the proportion of Vδ2$^+$ $\gamma\delta$ T cells (as a percentage of T cells). The relationship between the proportion of Vδ2$^+$ $\gamma\delta$ T

cells and their SHAP values is unclear on the beeswarm plot of Figure 6.17. It is better visualised as a scatterplot, as presented in Figure 6.18. A trend is obscured in the beeswarm plot that is revealed by plotting the proportion of V$\delta 2^+$ $\gamma\delta$ T cells versus the corresponding SHAP values. As the proportion of V$\delta 2^+$ $\gamma\delta$ T cells increases, the SHAP value decreases, and therefore the impact on the prediction of Gram-positive cause increases. However, there are two Gram-negative cases with high values for the proportion of V$\delta 2^+$ $\gamma\delta$ T cells. The model successfully identified the relatively abnormal relationship these outliers have with the proportion of V$\delta 2^+$ $\gamma\delta$ T cells, and this is reflected in their low absolute SHAP values.

Returning to the other features of the Random Forest model and their corresponding SHAP values described in Figure 6.17, we see a relatively straightforward relationship. Increased values for all other features were associated with higher SHAP values, influencing the model to predict a Gram-negative causative pathogen. The additional features in the Random Forest model included: neutrophil count, T cells memory cluster 2 (a $CD4^+$ cluster characterised by low expression of CD27 and CCR7, moderate expression of CD45RA, and high expression of CD57), CD25 expression on $CD8^+$ T cells, a $CD4^+$ $CD8^-$ MAIT cell cluster, and the proportion of T cells (as a percentage of PBMCs). The SHAP heatmap in Figure 6.17 shows that features were, for the majority, cooperative in their impact on model predictions. However, there were cases where features had contradictory SHAP values. For example, two patients on the far left-hand side of the heatmap had moderate to high SHAP values for the proportion of V$\delta 2^+$ $\gamma\delta$ T cells but were countered by low SHAP values for the neutrophil count, T cells memory cluster 2 and CD25 expression on $CD8^+$ T cells. Ultimately, the combination of the chosen features yields the correct prediction.

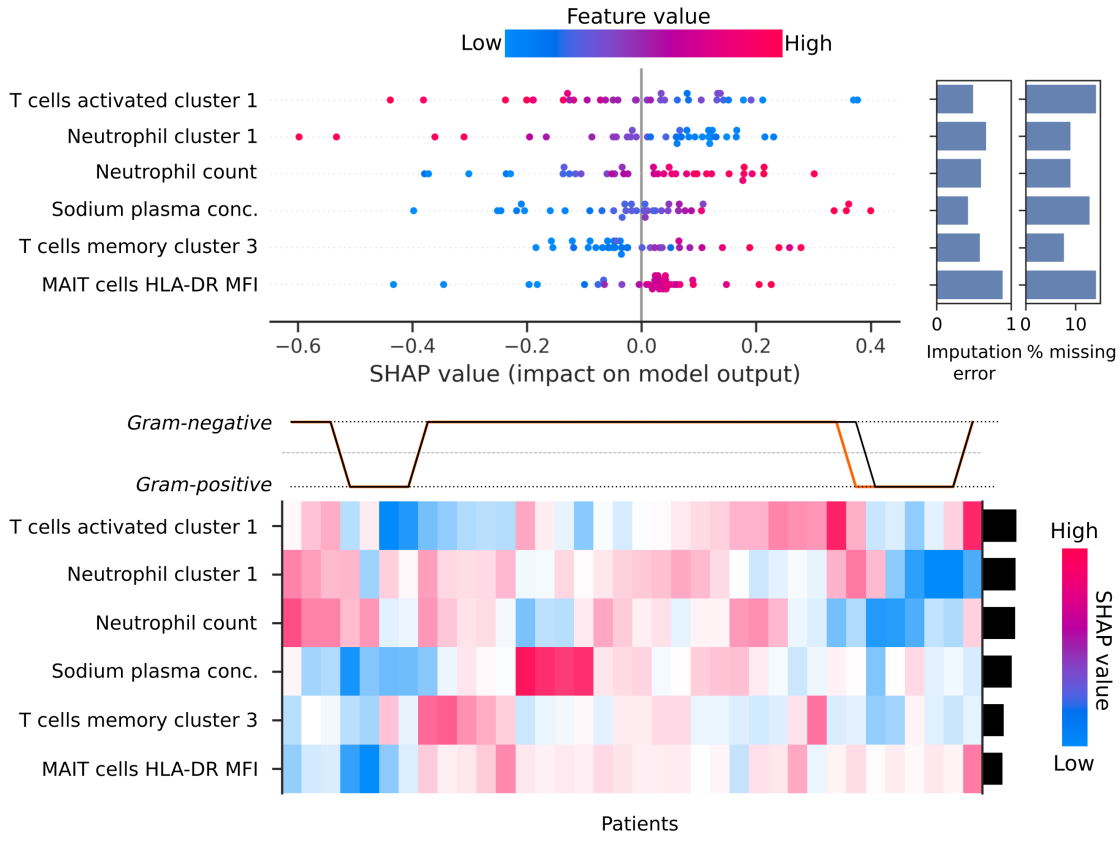

Figure 6.16: SHAP (SHapely Additive exPlanations) values for a Logistic regression model tasked with predicting Gram-negative cause in sepsis. The beeswarm plot (top) shows each observation as a single data point coloured by the value of the feature for that instance. On the x-axis is the SHAP value, a lower value corresponds to an instance having a more significant impact on the negative case for the model (i.e. prediction of Gram-positive sepsis), and a positive value corresponds to having a more significant impact on the positive case for the model (i.e. prediction of Gram-negative sepsis). A barplot on the right-hand side of the beeswarm plot shows the imputation error (with a maximum value of 1) and the percentage of missing values observed in the original data. The heatmap (bottom) shows the SHAP values for each patient. The bar plot on the right-hand y-axis shows each feature's mean absolute SHAP value and measures the impact of a feature on model prediction. The line plot above the heatmap shares the x-axis and displays each patient's predicted outcome (black line) and the actual outcome (orange line). The dotted line between the possible outcomes is the expected value, equivalent to the observed incidence of Gram-negative sepsis.

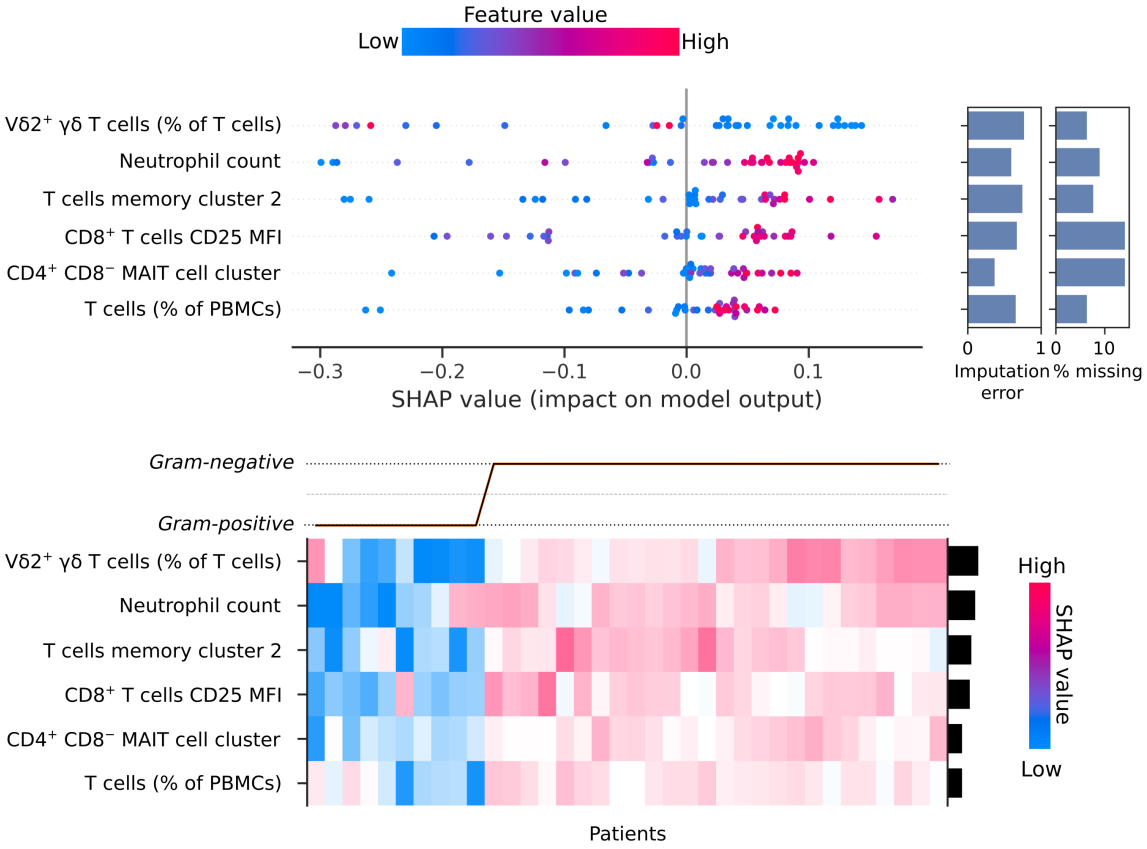

Figure 6.17: SHAP (SHapely Additive exPlanations) values for a Random Forest model tasked with predicting Gram-negative cause in sepsis. The beeswarm plot (top) shows each observation as a single data point coloured by the value of the feature for that instance. On the x-axis is the SHAP value, a lower value corresponds to an instance having a more significant impact on the negative case for the model (i.e. prediction of Gram-positive sepsis), and a positive value corresponds to having a more significant impact on the positive case for the model (i.e. prediction of Gram-negative sepsis). A barplot on the right-hand side of the beeswarm plot shows the imputation error (with a maximum value of 1) and the percentage of missing values observed in the original data. The heatmap (bottom) shows the SHAP values for each patient. The bar plot on the right-hand y-axis shows each feature's mean absolute SHAP value and measures the impact of a feature on model prediction. The line plot above the heatmap shares the x-axis and displays each patient's predicted outcome (black line) and the actual outcome (orange line). The dotted line between the possible outcomes is the expected value, equivalent to the observed incidence of Gram-negative sepsis.

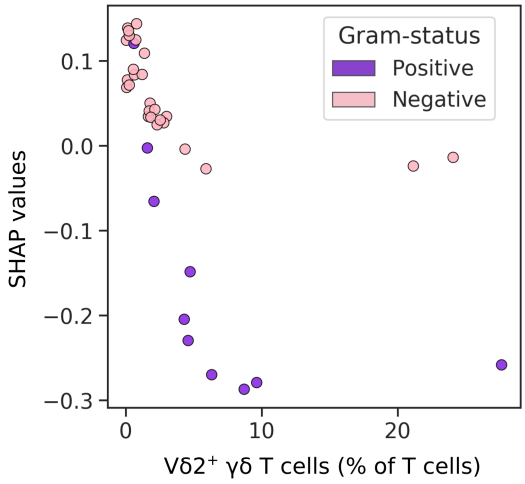

Figure 6.18: The proportion of Vδ2$^+$ $\gamma\delta$ T cells plotted against corresponding SHAP (SHapely Additive exPlanations) values that explain the impact on a Random Forest model tasked with predicting Gram-negative cause in sepsis. Each data point represents a unique patient, coloured by the causative pathogen of their acute infection.

6.8 Discussion

In this chapter, the ILTIS data (as described in Chapter 5) was exposed to classification algorithms to uncover multivariate patterns that correlated with outcomes of interest, namely mortality and the identification of a Gram-negative infection as the underlying cause. Many challenges arose during this analysis that reflect common issues when conducting a retrospective analysis of clinical and multi-omic data. The issue of class imbalance was addressed by penalising the misclassification of the minority case and using metrics balanced by class support. Missing data were also a significant issue, and this work proposed using the MissRanger algorithm (an extension of MissForest), which has the added advantage of capturing OOB error for each feature, reflecting the impact of imputation. Imputation error was reduced by using MissRanger over MissForest, but still exceeded an NMRSE of over 0.5 for most continuous variables. Therefore, as an additional validation, the identified multivariate patterns were validated on complete case data. Some models experienced a loss of sensitivity when exposed to complete case data, which could reflect the influence of class imbalance on the imputation mechanism. It could also result from the reduced quantity of training data available in complete case analysis.

Solutions proposed for handling class imbalance in imputation tasks suggest applying sampling techniques [337, 338]. However, it is questionable whether this would bias the training data when only a small sample is available. An alternative approach could be introducing class weights to the underlying Random Forest algorithm of MissRanger, but this is not possible in the current implementation. Another strategy proposed by Tomic *et al.* [46] is constructing complete case data by searching across all possible sets of feature and observation combinations, avoiding the need for imputation. However, this can limit the available training data and potentially bias the analysis, resulting in classification algorithms only applicable to a particular subset of the population, an issue that would be exacerbated in highly heterogeneous conditions such as sepsis.

A broad search strategy was employed to obtain informative feature sets, employing multiple feature selection algorithms and exposing all top-ranked combinations to many classification algorithms of ranging complexity and with a diverse choice of hyperparameters. Since no single classification algorithm is likely to be optimal for every problem [301], it is necessary to search across a wide range of classifiers, as is done in this study. Similarly, there are multiple feature selection algorithms [159, 161, 160], and no single approach is likely optimal for every task. In this study, five feature selection methods were explored, but one could stipulate that other methodologies could be included. It is essential to recognise, however, that no single feature set and classifier combination will be globally optimal, and there could be multiple patterns that correlate with the outcome.

The best-performing model and feature set combinations were chosen by LOOCV macro F1 score and validated against holdout and complete case data. The generation of independent holdout data is the only mechanism to ensure that the ascertained model and feature set combination is not overfitting to the chosen training data, other than the generation of an entirely new patient cohort for validation. Rigorous testing on validation data is a concern within the field of biomarker discovery, and feature selection in the absence of testing on a holdout set is a source of error that has been well-documented [339, 340, 330, 329]. A critical limitation of this study is the small sample size, increasing the risk that the holdout data was not representative of a larger population. An alternative approach could be a nested cross-validation methodology [341], with feature selection and hyperparameter tuning performed within each fold. The training data would differ within each round of cross-validation; therefore, the chosen features would likely differ. In order to draw any conclusion about the most informative

combination of biomarkers, a strategy would be required for combining the contribution of features within multiple independent models. Other suggestions include permutation studies, where models are trained and tested without feature selection or hyperparameter tuning. The performance is compared between the original data and a randomly permuted surrogate [329]. However, others suggest that independent holdout data are always necessary for meaningful evaluation of accuracy [339, 340, 329].

A model for predicting 30-day mortality could not be obtained, but an Extra Random Forest for predicting 90-day mortality was identified with a holdout AUC score of 0.85 (0.81 – 0.86), a significant improvement on the majority of previously reported prognosis biomarkers [96, 71]. It is not immediately apparent why models for predicting 30-day mortality did not generalise compared to those tasked with predicting 90-day mortality. The class imbalance observed for 90-day mortality was slightly less severe than 30-day mortality (Figure 6.1), which could be a contributing factor.

The proportion of T cells (as a percentage of PBMCs) was the main contributing feature to decision-making in the Extra Random Forest model (Figure 6.13). A comparison of T cells showed a significant difference between survivors and non-survivors at 90-days in Chapter 5 (Figure 5.15), and lymphopenia is well-documented as a sign of increased severity and associated with higher mortality [231]. Additional features in the Extra Random Forest model included CXCR3 expression on $CD4^+$ T cells, plasma concentrations of Arachidonic acid, and blood glucose. The importance of blood glucose levels is supported by the surviving sepsis campaign international guidelines, which recommend tight control of blood glucose levels, with hyperglycemia associated with increased mortality [76].

CXCR3 is a Th1-associated chemokine receptor upregulated rapidly upon cell activation and responsive to three interferon-inducible ligands: CXCL9, CXCL10, and CXCL11. CXCL10 concentrations have been shown to correlate with severity in sepsis [343, 342]. However, within the ILTIS cohort, concentrations of CXCL10 were decreased in non-survivors compared to survivors when measuring 30-day mortality (Figure 5.8). Higher values for CXCR3 expression on $CD4^+$ T cells contributed to a prediction of 90-day mortality, possibly suggesting the recruitment of $CD4^+$ T cells early in sepsis as being correlated with worse outcomes when considered with the other features included in the Extra Random Forest model.

Arachidonic acid had an inverse relationship with the prediction of 90-day mortality, with increased values encouraging the model to predict survival at 90-days. The pro-inflammatory eicosanoids, a family of bioactive lipids, are derived from arachidonic acid. Eicosanoid lipid mediators have been implicated in the pathogenesis of sepsis [344], and a reduction in arachidonic acid metabolism in sepsis patients compared to healthy controls has been described [345]. In this study, lower values of arachidonic acid in plasma contributed to the prediction of 90-day mortality in the Extra Random Forest model. The work discussed here highlights the benefits of including variables that describe lipid metabolism in multivariate models of sepsis.

Two models stood out from the rest when investigating the prediction of Gram-negative cause in sepsis: a Logistic regression model (holdout F1 score of 0.86 (0.8 – 1.0) and holdout AUC of 0.76 (0.64 – 1.0)) and a Random Forest model (holdout F1 score of 0.67 (0.59 – 0.75) and holdout AUC score of 0.86 (0.77 – 0.94)). Cluster 1, from the investigation of activated T cells, had the highest absolute mean SHAP score for the Logistic regression model. Lower values for this cluster were associated with a prediction of a Gram-negative cause. T cell activation cluster 1 encompassed the majority of $CD8^+$ T cells in the activation panel and was reported as the proportion of this cluster as a percentage of total T cells. Replacing this cluster with the proportion of $CD8^+$ T cells (as a percentage of total T cells) did not reduce training or holdout performance. T cells memory cluster 3, another $CD8^+$ T cell cluster identified from the study of memory subsets, was also implicated in the Logistic regression model. T cells memory cluster 3 demonstrated a "terminally differentiated" effector memory phenotype (TEMRA) [346] of $CD45RA^{hi}$ $CCR7^{lo}$ and was also characterised by high CD57 expression and low CD27 expression. T cells memory cluster 3, therefore, represents a replicatively senescent subset that could have a loss of functionality due to exhaustion, but this cannot be confirmed in the absence of additional markers such as programmed cell death 1 (PD-1) or mucin domain protein 3 (TIM-3) [347]. The other $CD8^+$ T cell parameter that appeared in the Gram-negative models was the expression of the activation marker CD25, which contributed to the Random Forest model, with increased expression influencing a prediction of Gram-negative cause. The surface marker CD25 (IL-2Rα) is the alpha chain of the IL-2 receptor, generated in response to specific antigen presentation along with co-stimulatory signalling [17]. It functions to respond to IL-2 during lymphocyte activation and remains elevated for several days [348]. Overall, the two models suggest that characteristics

of $CD8^+$ T cell activation could be informative for differentiating Gram-negative and Gram-positive infection in sepsis.

Neutrophil count was identified as an informative feature in both the Logistic regression and Random Forest model, with an increased neutrophil count associated with predicting Gram-negative cause. The relationship with neutrophil count corresponded with the trend observed in Chapter 5 and visualised in Figure 5.17. Neutrophil cluster 1 appeared in the Logistic regression model as an important feature, with lower values encouraging a Gram-negative prediction. Neutrophil cluster 1 was characterised by low expression of CD62L (L-selectin) and high expression of CD11b. Bacterial clearance by neutrophils depends on attachment to the microvasculature, controlled by the selectins and the integrins, two families of adhesion molecules [349]. Initially, binding is controlled by L-selectin, but sustained attachment requires the integrins CD11a and CD11b. Neutrophils from the blood of SIRS and sepsis patients have previously shown a decrease in CD62L expression and an increase in CD11b expression. The $CD62L^{lo}$ $CD11b^{hi}$ phenotype, combined with increased expression of CD64 (a high-affinity receptor for IgG) is associated with neutrophil activation [350]. Neutrophil cluster 1 exhibits an increased expression of CD64 relative to other neutrophil clusters, albeit fluorescence was low overall. Therefore, an increased neutrophil count but a decrease in circulating activated neutrophils contributed to a Gram-negative prediction in the Logistic regression model.

A compelling finding from this study was the importance attributed to parameters that describe unconventional T cells. In predicting 90-day mortality, increased expression of the activation marker CD25 on MAIT cells influenced survival prediction in the Extra Random Forest model. Previous studies have described MAIT cells as highly activated in clinical sepsis and protective during experimental sepsis [31]. Another activation marker on MAIT cells, HLA-DR, was associated with the prediction of Gram-negative cause in a Logistic regression model when the expression was increased. A semi-invariant TCR characterises MAIT cells with specificity for microbial riboflavin-derivative antigens presented by HLA-1b major histocompatibility complex (MHC)-related protein 1 (MR1) [30, 28, 29]. All causative Gram-negative pathogens implicated in this study have previously been described as capable of Vitamin B2 synthesis [23, 28]. In contrast, less than half of the Gram-positive causative pathogens observed were capable of Vitamin B2 synthesis. Subsequently, the biology that

underpins their specificity could explain the contribution of MAIT HLA-DR expression to the prediction of Gram-negative infection.

The proportion of V$\delta 2^+$ $\gamma\delta$ T cells (as a percentage of total T cells) was an essential feature in the Random Forest model. When combined with a high neutrophil count, parameters of CD8$^+$ T cells, and the total percentage of T cells, lower values for the proportion of V$\delta 2^+$ $\gamma\delta$ T cells encouraged the prediction of Gram-negative cause in sepsis. A reduction in circulating V$\delta 2^+$ $\gamma\delta$ T cells correlated with Gram-negative pathogens was also observed in Chapter 5.

Across all models, at least one parameter describing unconventional T cell biology has arisen in the top-ranking features and is included in the optimal models selected for further investigation. The inclusion of unconventional T cells in machine learning models helped differentiate the causative pathogen in patients with acute peritonitis [240] and have been identified as critical players in multi-parameter immune signatures with implications in COVID-19 prognosis [351, 352]. The work presented in this chapter provides additional evidence that extensive profiling of unconventional T cell populations could provide valuable contributions to predictive models of acute severe infectious disease.

It should be noted that the inclusion of features in the models described here does not imply causation, and their combined interaction in complex mathematical models identifies only correlations that could be valuable for the task of prediction. Their identification does offer the potential for new hypotheses and can help direct efforts of multi-parameter biomarker panels with application in routine care. The use of SHAP to interrogate machine learning models helps elevate the 'black box' nature of multivariate modelling [167, 166]. However, they do not answer causal questions, and future analysis should take the additional step of employing causal analysis methodologies. An example could be counterfactual explanations, a seemingly human-friendly approach to model explanations but subject to the 'Rashomon effect' that each instance will usually generate multiple counterfactual explanations [166, 353].

Further to the limitations regarding the interpretability of models, due to the small cohort studied, any observations must be confirmed with an additional validation cohort. If repeatable, any translational application of subsequent models would also require additional calibration to report on the confidence of subsequent predictions and to find the optimal prob-

ability threshold for positivity. Despite these limitations, the findings here show promise for a generalisable pattern for predicting 90-day mortality and underlying causes and exhibit the value of including parameters of unconventional T cells.

7 | Discussion and future work

7.1 The role of cytometry bioinformatics in biomarker discovery

At the beginning of this thesis, cytometry bioinformatics was introduced in response to the challenges faced by observational studies that collect large amounts of cytometry data over long periods to identify predictive biomarkers. The CytoPy software was developed, providing data structures for cytometry data analysis in Python and the first of its kind to anchor the analysis to a document-based database. The design proved valuable for the immunological profiling and biomarker discovery work discussed in later chapters because it allowed simple meta-data integration, and experimentation with different methodologies was easy to implement. Despite this, some improvements would help the usability and extend the capabilities of CytoPy, notably better integration with existing frameworks such as ScanPy [149].

The challenge of batch effects was addressed with the application of Harmony [196], implemented in Python [197] and integrated into the data structures of CytoPy. Batch effects are common amongst biomarker studies and were particularly challenging for the ILTIS study, where patients were recruited over several years, and flow cytometry was performed on fresh blood. If this work was to be repeated, there could be a potential benefit in the cryopreservation of cells and acquisition in a minimal number of batches. The cryopreservation process could risk the loss of cell populations and activation maker expression and should be accompanied by adequate validation of samples from sepsis patients. Alternatively, another technical intervention would be a repeat control, run with every sample and used as a reference during the post-hoc batch correction. The successful application of a reference control was demonstrated by Van Gassen *et al.* [354] and Ogishi *et al.* [355]. However, the practical implications of using a reference control throughout a prolonged observational study must be considered, and a suitable control must be chosen.

CytoPy provides an extensive toolbox for identifying cell populations in cytometry data. The autonomous gating methods are helpful but can still be quite labour intensive and may struggle where significant inter-sample variation is present. Therefore, autonomous gates should be limited to removing debris, artefacts, dead cells, or other simple gating strategies.

However, autonomous gating can extend to more complicated gating strategies through the use of dynamic time warping [122] if a more traditional analysis is preferred, and dynamic time warping is available in the most recent version of CytoPy.

Supervised classification methods are better suited when the user knows the precise number of populations to expect. New methodologies also continue to be developed at a staggering pace. For example, Blampey *et al.* [356] recently demonstrated a probabilistic model that can be directed with prior knowledge of the expected populations.

The exploratory nature of unsupervised clustering is valuable where the desired number of populations is unknown and has application for immunological biomarker discovery. In Chapter 4, a novel ensemble clustering method, GeoWaVe, was introduced. GeoWaVe was shown to be computationally efficient, had greater accuracy than graph-based ensemble clustering algorithms, and offered an interpretable visualisation step for the immunologist to introduce prior knowledge for selecting the final number of clusters. Using an ensemble clustering method such as GeoWaVe exposes the analysis to multiple clustering algorithms, each with different underlying mechanisms, therefore reducing the risk of overlooking a particular data partition or biasing the clustering results with the assumptions of a single algorithm.

Given more time, other paradigms of cytometry data analysis could be explored with potential application to biomarker discovery. Multiple instance learning was proposed by Arvaniti & Claassen [127], and the principles extended by Hu *et al.* [128] for the identification of latent cytomegalovirus infection. The concept removes the need to characterise the cell populations in cytometry data and instead focuses on predicting the disease state of interest. Inspecting the resulting model then identifies the cells that differentiate patients, which can then be inspected to determine their phenotype. A similar principle of focusing on the disease state was demonstrated by Weber *et al.* [135], alternatively employing clustering but intentionally identifying a large number of clusters (e.g. 100-400) and then selecting significant clusters or cluster-marker combinations with differential abundance and differential expression analysis.

Together, advancements in cytometry instruments and the methodologies for data analysis discussed here could help expand our understanding of the mechanisms driving sepsis. Current limitations regarding the number of parameters that can practically be included in cytometry panels result in a bias towards immune populations of interest to a particular

researcher $e.g.$ in the analysis presented in this thesis, prior work towards predicting underlying cause in acute infectious disease [259, 199] introduced a bias for unconventional T cell subsets. Advances in mass cytometry and spectral flow cytometry offer the opportunity to measure multiple paradigms of the immune response [135, 183], and imputation techniques offer the potential to characterise hundreds, perhaps even thousands, of markers in a single experiment [357]. Future biomarker studies that follow the inductive approach presented in this thesis could filter a vast immunological landscape without bias by utilising these new technologies to identify patterns with clinical application.

7.2 The immunopathology of sepsis and the role of unconventional T cells.

Before investigating the early immunological phenotype of sepsis patients, the available routine clinical data were reviewed. No individual routine biomarker successfully differentiated survivors and non-survivors (except FiO2, although this difference was absent when values were averaged over the 48-hour window before recruitment). The finding that routine clinical data alone were insufficient encourages more comprehensive phenotyping in sepsis.

Some interesting trends were observed when investigating soluble biomarkers such as cytokines, chemokines, and acute phase protein levels in plasma. CXCL10 was decreased in those patients who died within 30 days, increased levels of IL-15 and IL-6 showed a trend towards higher odds of mortality at 30 days, and IL-1α, OSM, and ferritin showed a trend towards higher plasma concentrations in Gram-positive infections compared to Gram-negative. Ultimately, however, limited data, class imbalance, and the detection limit of Luminex assays made it difficult to reconcile these findings.

Importantly, investigation of immune cell populations in whole blood did confirm previously well-described observations, such as a reduced proportion of circulating T cells and decreased HLA-DR expression on monocytes amongst non-survivors, and an increased proportion of circulating neutrophils in Gram-negative infections compared to Gram-positive infections.

The proportion of circulating populations of unconventional T cells was found to be decreased in Gram-negative infections compared to Gram-positive. Although it cannot be confirmed with the data presented in this thesis, one hypothesis is the possible recruitment of unconventional T cell populations to infected tissues. In support, Liuzzi *et al.* [259] demonstrated that Vγ9+Vδ2+ $\gamma\delta$ T cells and MAIT cells accumulated at the site of infection where HMB-PP and vitamin B2 producing pathogens were implicated. Future studies should try to extend such work to confirmed infections in sepsis patients. Such a study would require well-defined aetiology and a condition that warrants regular sampling, such as urinary sepsis or an acute upper respiratory infection.

In the ILTIS study described in this thesis, the sample obtained from sepsis patients was whole blood. Whilst whole blood is a minimally invasive and convenient sample from sepsis patients, it presents numerous limitations that limit our findings. Blood is a heterogeneous medium that might poorly reflect the complex immunological processes occurring within the tissue and at the site of infection. Biomarkers measured in the blood might be significantly diluted compared to the tissue in which they were initially produced. Their concentration in blood might be transient, making sampling time critical for their predictive value. Future studies should consider the potential for more targeted sampling strategies that obtain tissue from the site of infection or, alternatively, include multiple sampling points to model the transient nature of biomarkers in whole blood.

7.3 Machine learning models for identifying potential biomarker combinations.

The final results chapter collated the parameters described in Chapter 5, combined with additional data describing lipid concentrations in plasma, and created supervised machine learning models to predict mortality and underlying cause of infection. A modelling pipeline was created that considered the small cohort size, class imbalance, and missing data. Rigorous validation is a concern in biomarker discovery [339, 340, 330, 329] and it is essential to distinguish between the data used for evaluating a model and the data used for model development, especially the selection of biomarkers to be included. Although efforts were made to reduce the risk of overfitting, a critical limitation in this study was the size of the available dataset. If granted more time, an extensive resampling approach could be applied for feature selection, with repetition over thousands of permutations of the data, and each exposed to downstream models and tested against an independent holdout set. The computational intensity of such a procedure would need to be addressed and each permutation would likely derive a different set of interacting features, therefore a method would have to be devised to search this space for the optimal features. Regardless of any change to methodology, more data are needed, and any future study should expand on the existing data whilst identifying a validation population of equal size.

Models that generalised to the holdout data could not be obtained for 30-day mortality, but an Extra Random Forest model for 90-day mortality showed good performance with an AUC score of 0.854. The 90-day mortality model showed a diverse selection of input features, including parameters that quantified immune populations, activation profiles of T cells, lipid plasma concentrations and the APACHE II severity score. The diversity of the chosen features highlights the benefits of capturing variables that describe multiple systems and how their combination can contribute to the model's performance.

A question proposed in this work was whether including immunological biomarkers in a feature selection and supervised machine learning framework would uncover predictive patterns for determining the causative pathogen in sepsis. It is hypothesised that the specificity of APCs for pathogen-associated molecular patterns and the resulting signalling cascades and activation of the innate immune response would provide informative features for our models. The findings presented in this thesis are a step towards supporting this hypothesis, with models for identifying Gram-negative causative pathogens showing good performance. A logistic regression model with six parameters reported an AUC score of 0.76, and a Random Forest model with a different six parameters reported an AUC score of 0.86. A combination of features describing CD8 T cells, neutrophils, and unconventional T cells was valuable for predicting Gram-negative causative pathogens.

Even amongst the diversity of available variables to select from, parameters that described unconventional T cells arose in every model that performed well on holdout data. Increased CD25 expression on MAIT cells was associated with increased survival in the 90-day mortality model, increased HLA-DR expression on MAIT cells was associated with a prediction of Gram-negative causative pathogen, and the percentage of circulating V$\delta 2^+$ $\gamma\delta$ T cells were the most influential feature in the Random Forest model predicting Gram-negative causative pathogen. Therefore, parameters describing unconventional T cells should be considered a valuable contribution to the search for multi-parameter sepsis biomarker panels.

There was a particular focus on interpretability in the machine learning models described in Chapter 6. The interpretation was facilitated using SHAP values. However, it is essential to note that the selection of features in machine learning models and their associated SHAP values do not imply causation. Their combined interaction identifies a correlation with the predicted target. To identify the cause, additional experimentation and analysis would be re-

quired. Machine learning models can, however, assist in narrowing the list of target variables for investigation in subsequent experiments and is, therefore, a helpful hypothesis-generating exercise. Additionally, the interpretability of machine learning models offers the potential for identifying a more general sepsis model by identifying dysregulation patterns across multiple interconnected systems. When interpretable models are combined with systems of ordinary differential equations, there is the potential to gather insights that move beyond static dogmas of ‘cytokine storms’ and ‘immune paralysis’ [358].

7.4 Sepsis heterogeneity

A critical limitation in this study was the heterogeneity of the patients recruited to the ILTIS study, reflecting the complex and poorly defined nature of sepsis pathology. Fewer than 70 percent of the patients enrolled on the ILTIS study had a confirmed infection, a rate comparable to previous observations in sepsis [77]. It was impossible to identify whether the cause was the failure of microbiological culture or the genuine absence of any bacterial infection. Additionally, around 25 percent of patients were admitted to the ICU with trauma or following emergency surgery. Although this was included as a categorical variable in the machine learning pipeline to account for a confounding effect, the clinical condition and type of care for such patients would differ from those that had not experienced trauma. There was also insufficient data regarding patient co-morbidity and history of infectious disease before admission to the ICU. Such data form important confounding variables for both prediction of survival and the underlying cause of infection.

There is a solid case to be made that the current definition of sepsis is still inadequate. The current definition draws focus to the dysregulated host response as the characterising feature of sepsis. Nevertheless, all major clinical trials that seek to subdue the host response have either failed to show benefit or have proven harmful [359]. The complicated patterns of presentation have been recognised as a barrier to the advancement of diagnosis and therapy for some time [279, 278] and it is increasingly being recognised that the Sepsis-3 definition [243] cannot distinguish the complex heterogeneity observed in the pathophysiology of sepsis [280]. Research associated with COVID-19, a condition that has drawn many parallels to sepsis [360, 351], has reported success in uncovering immunological signatures associated with poor outcomes with links back to the underlying biological mechanisms [351, 43, 228]. The COVID-19 research has demonstrated that focusing on a particular pathology within sepsis can yield findings more readily associated with the underlying mechanism driving the immune response. Any future study expanding on the work discussed in this thesis should carefully consider the definition applied to recruitment. Reflecting on the success of immunophenotyping of COVID-19, simple strategies could be employed to limit recruitment to those of comparable aetiology, such as culture-positive urosepsis or acute lower respiratory infection. Alternatively, a robust recruitment approach might leverage unsupervised clustering and the identification of endotypes that should be treated as distinct yet overlap-

ping groups [281, 282]. In the future, sepsis will likely be recognised not as a syndrome but rather as a group of related diseases, each characterised by cellular alterations and related biomarkers [280, 281, 282].

7.5 Conclusion

In conclusion, the work presented in this thesis shows the interaction between three distinct but overlapping fields of study: cytometry bioinformatics, supervised machine learning, and sepsis biomarker discovery. A comprehensive framework for cytometry data analysis was introduced, along with a novel methodology for ensemble clustering. The methodology was then applied to an observational study of severe sepsis patients, characterising the early immune response and creating parameters for downstream statistical models. In recognition of the potential for multi-parameter biomarker panels, supervised machine learning was employed with a diverse range of input data, including routine clinical data, immunological parameters, and lipid data. The work here demonstrates the application of supervised machine learning for sepsis biomarker discovery and the potential contribution of parameters that describe unconventional T cell populations.

Bibliography

[1] *References and sources: Sepsis statistics*. Sept. 2022. URL: https://sepsistrust.org/about/about-sepsis/references-and-sources/.

[2] Michael Bauer et al. "Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019 - results from a systematic review and meta-analysis". In: *Critical Care* 24.1 (May 2020). DOI: 10.1186/s13054-020-02950-2. URL: https://doi.org/10.1186/s13054-020-02950-2.

[3] World Health Organization et al. "Global report on the epidemiology and burden of sepsis: current evidence, identifying gaps and future directions". In: (2020).

[4] Fernando Jose da Silva Ramos, Flavio Geraldo Rezende de Freitas, and Flavia Ribeiro Machado. "Sepsis in patients hospitalized with coronavirus disease 2019: how often and how severe?" In: *Current Opinion in Critical Care* 27.5 (2021), p. 474.

[5] *The cost of sepsis care in the UK*. Feb. 2017. URL: http://allcatsrgrey.org.uk/wp/wpfb-file/yhec-sepsis-report-17-02-17-final-pdf/.

[6] Celeste M Torio and Brian J Moore. "National inpatient hospital costs: the most expensive conditions by payer, 2013: statistical brief# 204". In: (2016).

[7] Manu Shankar-Hari and Gordon D Rubenfeld. "Understanding long-term outcomes following sepsis: implications and challenges". In: *Current Infectious Disease Reports* 18.11 (2016), pp. 1–9.

[8] Hallie C Prescott and Derek C Angus. "Enhancing recovery from sepsis: a review". In: *JAMA* 319.1 (2018), pp. 62–75.

[9] Raquel Bragante Gritte, Talita Souza-Siqueira, Rui Curi, Marcel Cerqueira Cesar Machado, and Francisco Garcia Soriano. "Why septic patients remain sick after hospital discharge?" In: *Frontiers in Immunology* (2021), p. 3873.

[10] *The Global Sepsis Alliance*. https://www.global-sepsis-alliance.org/. Accessed: 2022-09-19.

[11] *The UK sepsis trust*. https://sepsistrust.org/. Accessed: 2022-09-19.

[12] *WHA Adopts Resolution on Sepsis*. https://www.global-sepsis-alliance.org/news/2017/5/26/wha-adopts-resolution-on-sepsis. Accessed: 2022-09-19.

[13] Bishal Gyawali, Karan Ramakrishna, and Amit S Dhamoon. "Sepsis: The evolution in definition, pathophysiology, and management". In: *SAGE Open Medicine* 7 (2019).

[14] Duane J Funk, Joseph E Parrillo, and Anand Kumar. "Sepsis and septic shock: a history". In: *Critical Care Clinics* 25.1 (2009), pp. 83–101.

[15] Dominik Jarczak, Stefan Kluge, and Axel Nierhaus. "Sepsis—pathophysiology and therapeutic concepts". In: *Frontiers in Medicine* (2021), p. 609.

[16] Jeffrey E Gotts and Michael A Matthay. "Sepsis: pathophysiology and clinical management". In: *British Medical Journal* 353 (2016).

[17] Kenneth Murphy and Casey Weaver. *Janeway's Immunobiology*. Garland science, 2016.

[18] Melissa A Kovach and Theodore J Standiford. "The function of neutrophils in sepsis". In: *Current Opinion in Infectious Diseases* 25.3 (2012), pp. 321–327.

[19] Eric Vivier, Elena Tomasello, Myriam Baratin, Thierry Walzer, and Sophie Ugolini. "Functions of natural killer cells". In: *Nature Immunology* 9.5 (Apr. 2008), pp. 503–510. DOI: 10.1038/ni1582. URL: https://doi.org/10.1038/ni1582.

[20] José C Alves-Filho, Fernando Spiller, and Fernando Q Cunha. "Neutrophil paralysis in sepsis". In: *Shock* 34.7 (2010), pp. 15–21.

[21] Geneviève Drifte, Irène Dunn-Siegrist, Pierre Tissières, and Jérôme Pugin. "Innate immune functions of immature neutrophils in patients with sepsis and severe systemic inflammatory response syndrome". In: *Critical Care Medicine* 41.3 (2013), pp. 820–832.

[22] Xiaofei Shen, Ke Cao, Yang Zhao, and Junfeng Du. "Targeting neutrophils in sepsis: From mechanism to translation". In: *Frontiers in Pharmacology* 12 (2021), p. 644270.

[23] Martin S Davey et al. “Microbe-specific unconventional T cells induce human neutrophil differentiation into antigen cross-presenting cells”. In: *The Journal of Immunology* 193.7 (2014), pp. 3704–3716.

[24] Timothy SC Hinks and Xia-Wei Zhang. “MAIT cell activation and functions”. In: *Frontiers in Immunology* 11 (2020), p. 1014.

[25] Matthias Eberl, Ida M. Friberg, Anna Rita Liuzzi, Matt P. Morgan, and Nicholas Topley. “Pathogen-Specific Immune Fingerprints during Acute Infection: The Diagnostic Potential of Human gamma delta T-Cells”. In: *Frontiers in Immunology* 5 (2014). DOI: 10.3389/fimmu.2014.00572.

[26] Matthias Eberl et al. “A Rapid Crosstalk of Human $\gamma\delta$ T Cells and Monocytes Drives the Acute Inflammation in Bacterial Infections”. In: *PLOS Pathogens* 5.2 (Feb. 2009), pp. 1–16. DOI: 10.1371/journal.ppat.1000308. URL: https://doi.org/10.1371/journal.ppat.1000308.

[27] Neil E McCarthy and Matthias Eberl. “Human $\gamma\delta$ T-cell control of mucosal immunity and inflammation”. In: *Frontiers in Immunology* 9 (2018), p. 985.

[28] Anna Rita Liuzzi, James E McLaren, David A Price, and Matthias Eberl. “Early innate responses to pathogens: pattern recognition by unconventional human T-cells”. In: *Current Opinion in Immunology* 36 (2015), pp. 31–37.

[29] Dale I Godfrey, Hui-Fern Koay, James McCluskey, and Nicholas A Gherardin. “The biology and functional importance of MAIT cells”. In: *Nature Immunology* 20.9 (2019), pp. 1110–1128.

[30] Lars Kjer-Nielsen et al. “MR1 presents microbial vitamin B metabolites to MAIT cells”. In: *Nature* 491.7426 (2012), pp. 717–723.

[31] Shubhanshi Trivedi et al. “Mucosal-associated invariant T (MAIT) cells mediate protective host responses in sepsis”. In: *Elife* 9 (2020), e55615.

[32] Andreas Oberholzer, Caroline Oberholzer, and Lyle L Moldawer. “Interleukin-10: a complex role in the pathogenesis of sepsis syndromes and its potential as an anti-inflammatory drug”. In: *Critical Care Medicine* 30.1 (2002), S58–S63.

[33] Charalambos A Gogos, Eugenia Drosou, Harry P Bassaris, and Athanassios Skoutelis. “Pro-versus anti-inflammatory cytokine profile in patients with severe sepsis: a marker for prognosis and future therapeutic options”. In: *The Journal of Infectious Diseases* 181.1 (2000), pp. 176–180.

[34] Christian B Bergmann et al. “Potential targets to mitigate trauma-or sepsis-induced immune suppression”. In: *Frontiers in Immunology* 12 (2021), p. 622601.

[35] Matthew D Martin, Vladimir P Badovinac, and Thomas S Griffith. “CD4 T cell responses and the sepsis-induced immunoparalysis state”. In: *Frontiers in Immunology* 11 (2020), p. 1364.

[36] Caroline Landelle et al. “Low monocyte human leukocyte antigen-DR is independently associated with nosocomial infections after septic shock”. In: *Intensive Care Medicine* 36.11 (2010), pp. 1859–1866.

[37] Pénélope Bourgoin et al. “Toward monocyte HLA-DR bedside monitoring: a proof-of-concept study”. In: *Shock* 55.6 (2021), pp. 782–789.

[38] Guillaume Monneret and Fabienne Venet. “Monocyte HLA-DR in sepsis: shall we stop following the flow?” In: *Critical Care* 18.1 (2014), pp. 1–2.

[39] Karen J Quadrini et al. “A flow cytometric assay for HLA-DR expression on monocytes validated as a biomarker for enrollment in sepsis clinical trials”. In: *Cytometry Part B: Clinical Cytometry* 100.1 (2021), pp. 103–114.

[40] Jeffrey J Presneill, Trudi Harris, Alastair G Stewart, John F Cade, and John W Wilson. “A randomized phase II trial of granulocyte-macrophage colony-stimulating factor therapy in severe sepsis with respiratory dysfunction”. In: *American Journal of Respiratory and Critical Care Medicine* 166.2 (2002), pp. 138–143.

[41] Jonathan S Boomer et al. “Immunosuppression in patients who die of sepsis and multiple organ failure”. In: *JAMA* 306.23 (2011), pp. 2594–2605.

[42] Miguel Reyes et al. “An immune-cell signature of bacterial sepsis”. In: *Nature Medicine* 26.3 (2020), pp. 333–340.

[43] Nianping Liu et al. “Single-cell analysis of COVID-19, sepsis, and HIV infection reveals hyperinflammatory and immunosuppressive signatures in monocytes”. In: *Cell Reports* 37.1 (Oct. 2021), p. 109793. DOI: 10.1016/j.celrep.2021.109793. URL: https://doi.org/10.1016/j.celrep.2021.109793.

[44] Shinjiro Saito et al. “Epidemiology of disseminated intravascular coagulation in sepsis and validation of scoring systems”. In: *Journal of Critical Care* 50 (2019), pp. 23–30.

[45] Shu-Min Lin et al. “Serum thrombomodulin level relates to the clinical course of disseminated intravascular coagulation, multiorgan dysfunction syndrome, and mortality in patients with sepsis”. In: *Critical Care Medicine* 36.3 (2008), pp. 683–689.

[46] Adriana Tomic et al. “SIMON, an automated machine learning system, reveals immune signatures of influenza vaccine responses”. In: *The Journal of Immunology* 203.3 (2019), pp. 749–759.

[47] Sung-Yeon Cho and Jung-Hyun Choi. “Biomarkers of sepsis”. In: *Infection & Chemotherapy* 46.1 (2014), pp. 1–12.

[48] Apichot So-Ngern et al. “Prognostic value of serum procalcitonin level for the diagnosis of bacterial infections in critically-ill patients”. In: *Infection & Chemotherapy* 51.3 (2019), pp. 263–273.

[49] Dong Wook Jekarl et al. “Procalcitonin as a prognostic marker for sepsis based on SEPSIS-3”. In: *Journal of Clinical Laboratory Analysis* 33.9 (2019), e22996.

[50] Daniel G Remick, Gerald R Bolgos, Javed Siddiqui, Jungsoon Shin, and Jean A Nemzek. “Six at six: interleukin-6 measured 6 h after the initiation of sepsis predicts mortality over 3 days”. In: *Shock* 17.6 (2002), pp. 463–467.

[51] Juhyun Song et al. “Diagnostic and prognostic value of interleukin-6, pentraxin 3, and procalcitonin levels among sepsis and septic shock patients: a prospective controlled study according to the Sepsis-3 definitions”. In: *BMC Infectious Diseases* 19.1 (2019), pp. 1–11.

[52] Daniel J Henning et al. “Interleukin-6 improves infection identification when added to physician judgment during evaluation of potentially septic patients”. In: *The American Journal of Emergency Medicine* 38.5 (2020), pp. 947–952.

[53] Aikaterini Dimoula et al. “Serial determinations of neutrophil CD64 expression for the diagnosis and monitoring of sepsis in critically ill patients”. In: *Clinical Infectious Diseases* 58.6 (2014), pp. 820–829.

[54] Johannes JML Hoffmann. “Neutrophil CD64: a diagnostic marker for infection and sepsis”. In: *Clinical Chemistry and Laboratory Medicine* 47.8 (2009), pp. 903–916.

[55] Shigeatsu Endo et al. “Usefulness of presepsin in the diagnosis of sepsis in a multicenter prospective study”. In: *Journal of Infection and Chemotherapy* 18.6 (2012), pp. 891–897.

[56] Oh Joo Kweon, Jee-Hye Choi, Sang Kil Park, and Ae Ja Park. “Usefulness of presepsin (sCD14 subtype) measurements as a new marker for the diagnosis and prediction of disease severity of sepsis in the Korean population”. In: *Journal of Critical Care* 29.6 (2014), pp. 965–970.

[57] Jing Zhang, Zhi-De Hu, Jia Song, and Jiang Shao. “Diagnostic value of presepsin for sepsis: a systematic review and meta-analysis”. In: *Medicine* 94.47 (2015).

[58] Xin Zhang, Dan Liu, You-Ning Liu, Rui Wang, and Li-Xin Xie. “The accuracy of presepsin (sCD14-ST) for the diagnosis of sepsis in adults: a meta-analysis”. In: *Critical Care* 19.1 (2015), pp. 1–11.

[59] Chien-Chang Lee et al. “Prognostic value of mortality in emergency department sepsis score, procalcitonin, and C-reactive protein in patients with sepsis at the emergency department”. In: *Shock* 29.3 (2008), pp. 322–327.

[60] Shubhangi Arora, Prashant Singh, Preet Mohinder Singh, and Anjan Trikha. “Procalcitonin levels in survivors and nonsurvivors of sepsis: systematic review and meta-analysis”. In: *Shock* 43.3 (2015), pp. 212–221.

[61] Remi Porte et al. “The long pentraxin PTX3 as a humoral innate immunity functional player and biomarker of infections and sepsis”. In: *Frontiers in Immunology* 10 (2019), p. 794.

[62] Rui Tian et al. “Plasma PTX3, MCP1 and Ang2 are early biomarkers to evaluate the severity of sepsis and septic shock”. In: *Scandinavian Journal of Immunology* 90.6 (2019), e12823.

[63] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. “A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study”. In: *JAMA* 270.24 (1993), pp. 2957–2963.

[64] Chenggong Hu, Yongfang Zhou, Chang Liu, and Yan Kang. “Pentraxin-3, procalcitonin and lactate as prognostic markers in patients with sepsis and septic shock”. In: *Oncotarget* 9.4 (2018), p. 5125.

[65] Olga Livaditi et al. “Neutrophil CD64 expression and serum IL-8: sensitive early markers of severity and outcome in sepsis”. In: *Cytokine* 36.5-6 (2006), pp. 283–290.

[66] Shuangqing Liu et al. “Effects of neutrophil-to-lymphocyte ratio combined with interleukin-6 in predicting 28-day mortality in patients with sepsis”. In: *Frontiers in Immunology* 12 (2021), p. 639735.

[67] Jeffrey H Pruitt, EM Copeland 3rd, and Lyle L Moldawer. “Interleukin-1 and interleukin-1 antagonism in sepsis, systemic inflammatory response syndrome, and septic shock.” In: *Shock (Augusta, Ga.)* 3.4 (1995), pp. 235–251.

[68] Yun Ge, Man Huang, and Yong-ming Yao. “Recent advances in the biology of IL-1 family cytokines and their potential roles in development of sepsis”. In: *Cytokine & Growth Factor Reviews* 45 (2019), pp. 24–34.

[69] Serge Masson et al. “Circulating presepsin (soluble CD14 subtype) as a marker of host response in patients with severe sepsis or septic shock: data from the multicenter, randomized ALBIOS trial”. In: *Intensive Care Medicine* 41.1 (2015), pp. 12–20.

[70] Guillaume Monneret, Morgane Gossez, Nima Aghaeepour, Brice Gaudilliere, and Fabienne Venet. “How clinical flow cytometry rebooted sepsis immunology”. In: *Cytometry Part A* 95.4 (2019), pp. 431–441.

[71] Mi-Hee Kim and Jung-Hyun Choi. “An update on sepsis biomarkers”. In: *Infection & chemotherapy* 52.1 (2020), p. 1.

[72] Edward J Schenck et al. “Circulating cell death biomarker TRAIL is associated with increased organ dysfunction in sepsis”. In: *JCI Insight* 4.9 (2019).

[73] Katy Chun et al. “Sepsis pathogen identification”. In: *SLAS Technology* 20.5 (2015), pp. 539–561.

[74] Kenneth H Rand et al. “Hourly effect of pretreatment with IV antibiotics on blood culture positivity rate in emergency department patients”. In: *Open Forum Infectious Diseases*. Vol. 6. 5. Oxford University Press US. 2019, ofz179.

[75] Matthew P Cheng et al. “Blood culture results before and after antimicrobial administration in patients with severe manifestations of sepsis: a diagnostic study”. In: *Annals of Internal Medicine* 171.8 (2019), pp. 547–554.

[76] Laura Evans et al. “Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021”. In: *Intensive Care Medicine* 47.11 (2021), pp. 1181–1247.

[77] Jonathan Thorndike and Marin H. Kollef. “Culture-negative sepsis”. In: *Current Opinion in Critical Care* 26.5 (Aug. 2020), pp. 473–477. DOI: 10.1097/mcc.0000000000000751. URL: https://doi.org/10.1097/mcc.0000000000000751.

[78] Vincent X Liu et al. “The timing of early antibiotics and hospital mortality in sepsis”. In: *American journal of Respiratory and Critical Care Medicine* 196.7 (2017), pp. 856–863.

[79] Andrew F Shorr et al. “Inappropriate antibiotic therapy in Gram-negative sepsis increases hospital length of stay”. In: *Critical Care Medicine* 39.1 (2011), pp. 46–51.

[80] Bethany Tellor et al. “Inadequate source control and inappropriate antibiotics are key determinants of mortality in patients with intra-abdominal sepsis and associated bacteremia”. In: *Surgical Infections* 16.6 (2015), pp. 785–793.

[81] Marya D Zilberberg, Andrew F Shorr, Scott T Micek, Cristina Vazquez-Guillamet, and Marin H Kollef. “Multi-drug resistance, inappropriate initial antibiotic therapy and mortality in Gram-negative severe sepsis and septic shock: a retrospective cohort study”. In: *Critical Care* 18.6 (2014), pp. 1–13.

[82] Robert J Feezor et al. “Molecular characterization of the acute inflammatory response to infections with gram-negative versus gram-positive bacteria”. In: *Infection and Immunity* 71.10 (2003), pp. 5803–5813.

[83] Christian Leli et al. “Procalcitonin levels in gram-positive, gram-negative, and fungal bloodstream infections”. In: *Disease Markers* 2015 (2015).

[84] P Chalupa, O Beran, Heiko Herwald, N Kaspřiková, and M Holub. “Evaluation of potential biomarkers for the discrimination of bacterial and viral infections”. In: *Infection* 39.5 (2011), pp. 411–417.

[85] Daniel O Thomas-Ruddel et al. “Influence of pathogen and focus of infection on procalcitonin values in sepsis patients with bacteremia or candidemia”. In: *Critical Care* 22.1 (2018), pp. 1–11.

[86] Jian-Chang Lin, Zhao-Hong Chen, and Xiao-Dong Chen. “Elevated serum procalcitonin predicts Gram-negative bloodstream infections in patients with burns”. In: *Burns* 46.1 (2020), pp. 182–189.

[87] Pierre Emmanuel Charles et al. “Serum procalcitonin elevation in critically ill patients at the onset of bacteremia caused by either Gram negative or Gram positive bacteria”. In: *BMC Infectious Diseases* 8.1 (2008), pp. 1–8.

[88] Dominique J Pepper et al. “Procalcitonin-guided antibiotic discontinuation and mortality in critically ill adults: a systematic review and meta-analysis”. In: *Chest* 155.6 (2019), pp. 1109–1118.

[89] Igor Stoma et al. “Combination of sepsis biomarkers may indicate an invasive fungal infection in haematological patients”. In: *Biomarkers* 24.4 (2019), pp. 401–406.

[90] Jaap ten Oever et al. “Combination of biomarkers for the discrimination between bacterial and viral lower respiratory tract infections”. In: *Journal of Infection* 65.6 (2012), pp. 490–495.

[91] Xinjun Li, Xiaozhou Yuan, and Chengbin Wang. “The clinical value of IL-3, IL-4, IL-12p70, IL17A, IFN-γ, MIP-1β, NLR, P-selectin, and TNF-α in differentiating bloodstream infections caused by gram-negative, gram-positive bacteria and fungi in hospitalized patients: An Observational Study”. In: *Medicine* 98.38 (2019).

[92] Kuan-Ting Liu, Yao-Hua Liu, Chun-Yu Lin, Po-Lin Kuo, and Meng-Chi Yen. “Inflammatory molecules expression pattern for identifying pathogen species in febrile patient serum”. In: *Experimental and Therapeutic Medicine* 12.1 (2016), pp. 312–318.

[93] Tanja Vollmer, Cornelia Piper, Knut Kleesiek, and Jens Dreier. “Lipopolysaccharide-binding protein: a new biomarker for infectious endocarditis?” In: *Clinical Chemistry* 55.2 (2009), pp. 295–304.

[94] C Elsing, S Ernst, N Kayali, W Stremmel, and S Harenberg. “Lipopolysaccharide binding protein, interleukin-6 and C-reactive protein in acute gastrointestinal infections: value as biomarkers to reduce unnecessary antibiotic therapy”. In: *Infection* 39.4 (2011), pp. 327–331.

[95] Yasser Sakr, Ulricke Burgett, Flavio E Nacul, Konrad Reinhart, and Frank Brunkhorst. "Lipopolysaccharide binding protein in a surgical intensive care unit: a marker of sepsis?" In: *Critical Care Medicine* 36.7 (2008), pp. 2014–2022.

[96] Charalampos Pierrakos, Dimitrios Velissaris, Max Bisdorff, John C Marshall, and Jean-Louis Vincent. "Biomarkers of sepsis: time for a reappraisal". In: *Critical Care* 24.1 (2020), pp. 1–15.

[97] Kristian Kofoed et al. "Use of plasma C-reactive protein, procalcitonin, neutrophils, macrophage migration inhibitory factor, soluble urokinase-type plasminogen activator receptor, and soluble triggering receptor expressed on myeloid cells-1 in combination to diagnose infections: a prospective study". In: *Critical Care* 11.2 (2007), pp. 1–10.

[98] Hanah Kim, Mina Hur, Hee-Won Moon, Yeo-Min Yun, and Salvatore Di Somma. "Multimarker approach using procalcitonin, presepsin, galectin-3, and soluble suppression of tumorigenicity 2 for the prediction of mortality in sepsis". In: *Annals of Intensive Care* 7.1 (2017), pp. 1–9.

[99] Ishan Taneja et al. "Combining biomarkers with EMR data to identify patients in different phases of sepsis". In: *Scientific Reports* 7.1 (2017), pp. 1–12.

[100] Ishan Taneja et al. "Diagnostic and prognostic capabilities of a biomarker and EMR-based machine learning algorithm for sepsis". In: *Clinical and Translational Science* 14.4 (2021), pp. 1578–1589.

[101] Michael Moor, Bastian Rieck, Max Horn, Catherine R Jutzeler, and Karsten Borgwardt. "Early prediction of sepsis in the ICU using machine learning: a systematic review". In: *Frontiers in Medicine* 8 (2021), p. 607952.

[102] Lucas M Fleuren et al. "Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy". In: *Intensive Care Medicine* 46.3 (2020), pp. 383–400.

[103] Dickinson Becton and Company. *FlowJo*. Version 10.8. Sept. 19, 2022. URL: https://www.flowjo.com/.

[104] De Novo Software. *FCS Express*. Version 7.14.0020. Sept. 19, 2022. URL: https://www.flowjo.com/.

[105] Yvan Saeys, Sofie Van Gassen, and Bart N Lambrecht. "Computational flow cytometry: helping to make sense of high-dimensional immunology data". In: *Nature Reviews Immunology* 16.7 (2016), pp. 449–462.

[106] Ali Bashashati and Ryan R Brinkman. "A survey of flow cytometry data analysis methods". In: *Advances in Bioinformatics* 2009 (2009).

[107] *National Library of Medicine*. https://pubmed.ncbi.nlm.nih.gov/. Accessed: 2022-09-19.

[108] Florian Hahne et al. "flowCore: a Bioconductor package for high throughput flow cytometry". In: *BMC Bioinformatics* 10.1 (2009), pp. 1–8.

[109] Kenneth Lo, Florian Hahne, Ryan R. Brinkman, and Raphael Gottardo. "flowClust: A Bioconductor package for automated gating of flow cytometry data". In: *BMC Bioinformatics* 10 (2009), pp. 1–8. ISSN: 14712105. DOI: 10.1186/1471-2105-10-145.

[110] Nima Aghaeepour, Radina Nikolic, Holger H Hoos, and Ryan R Brinkman. "Rapid cell population identification in flow cytometry data". In: *Cytometry Part A* 79.1 (2011), pp. 6–13.

[111] Peng Qiu et al. "Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE". In: *Nature Biotechnology* 29.10 (2011), pp. 886–893. ISSN: 10870156. DOI: 10.1038/nbt.1991.

[112] Yu Qian et al. "Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data". In: *Cytometry Part B: Clinical Cytometry* 78.S1 (2010), S69–S82.

[113] Habil Zare, Parisa Shooshtari, Arvind Gupta, and Ryan R Brinkman. “Data reduction for spectral clustering to analyze high throughput flow cytometry data”. In: *BMC Bioinformatics* 11.1 (2010), pp. 1–16.

[114] Iftekhar Naim et al. “SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 1: Algorithm design”. In: *Cytometry Part A* 85.5 (2014), pp. 408–421. ISSN: 15524930. DOI: 10.1002/cyto.a.22446.

[115] Yongchao Ge and Stuart C Sealfon. “flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding”. In: *Bioinformatics* 28.15 (2012), pp. 2052–2058.

[116] Nima Aghaeepour et al. “Critical assessment of automated flow cytometry data analysis techniques”. In: *Nature Methods* 10.3 (2013), pp. 228–238. ISSN: 15487091. DOI: 10.1038/NMETH.2365.

[117] Mehrnoush Malek et al. “FlowDensity: Reproducing manual gating of flow cytometry data by automated density-based cell population identification”. In: *Bioinformatics* 31.4 (2015), pp. 606–607. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu677.

[118] Guenther Walther et al. “Automatic clustering of flow cytometry data with density-based merging”. In: *Advances in Bioinformatics* 2009 (2009).

[119] Fiona E. Craig, Ryan R. Brinkman, Stephen Ten Eyck, and Nima Aghaeepour. “Computational analysis optimizes the flow cytometric evaluation for lymphoma”. In: *Cytometry Part B: Clinical Cytometry* 30.9 (2013), n/a–n/a. DOI: 10.1002/cytob.21115.

[120] Greg Finak et al. “OpenCyto: An Open Source Infrastructure for Scalable, Robust, Reproducible, and Automated, End-to-End Flow Cytometry Data Analysis”. In: *PLoS Computational Biology* 10.8 (2014). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1003806.

[121] Kerstin Johnsson, Jonas Wallin, and Magnus Fontes. “BayesFlow: Latent modeling of flow cytometry cell populations”. In: *BMC Bioinformatics* 17.1 (2016), pp. 1–16. ISSN: 14712105. DOI: 10.1186/s12859-015-0862-z. URL: http://dx.doi.org/10.1186/s12859-015-0862-z.

[122] Markus Lux et al. “flowLearn: fast and precise identification and quality checking of cell populations in flow cytometry”. In: *Bioinformatics* 34.13 (2018), pp. 2245–2253.

[123] Nima Aghaeepour et al. “GateFinder: projection-based gating strategy optimization for flow and mass cytometry”. In: *Bioinformatics* 34.23 (2018), pp. 4131–4133.

[124] Daniel Commenges, Chariff Alkhassim, Raphael Gottardo, Boris Hejblum, and Rodolphe Thiébaut. “cytometree: A binary tree algorithm for automatic gating in cytometry analysis”. In: *Cytometry Part A* 93.11 (2018), pp. 1132–1140. ISSN: 15524930. DOI: 10.1002/cyto.a.23601.

[125] Huamin Li et al. “Gating mass cytometry data by deep learning”. In: *Bioinformatics* 33.21 (July 2017), pp. 3423–3430. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx448. eprint: https://academic.oup.com/bioinformatics/article-pdf/33/21/3423/25166108/btx448.pdf. URL: https://doi.org/10.1093/bioinformatics/btx448.

[126] Matthew Amodio et al. “Exploring single-cell data with deep multitasking neural networks”. In: *Nature Methods* 16.11 (2019), pp. 1139–1145. ISSN: 15487105. DOI: 10.1038/s41592-019-0576-7. URL: http://dx.doi.org/10.1038/s41592-019-0576-7.

[127] Eirini Arvaniti and Manfred Claassen. “Sensitive detection of rare disease-associated cell subsets via representation learning”. In: *Nature Communications* 8.1 (2017), p. 14825. ISSN: 2041-1723. DOI: 10.1038/ncomms14825. URL: https://doi.org/10.1038/ncomms14825.

[128] Zicheng Hu, Alice Tang, Jaiveer Singh, Sanchita Bhattacharya, and Atul J Butte. “A robust and interpretable end-to-end deep learning model for cytometry data”. In: *Proceedings of the National Academy of Sciences* 117.35 (2020), pp. 21373–21380.

[129] Karthik Shekhar, Petter Brodin, Mark M. Davis, and Arup K. Chakraborty. “Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE)”. In: *Pro-*

ceedings of the National Academy of Sciences of the United States of America 111.1 (2014), pp. 202–207. ISSN: 00278424. DOI: 10.1073/pnas.1321405111.

[130] Sofie Van Gassen et al. "FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data". In: *Cytometry Part A* 87.7 (2015), pp. 636–645. DOI: 10.1002/cyto.a.22625. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.a.22625. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.22625.

[131] Jacob H Levine et al. "Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis". In: *Cell* 162.1 (2015), pp. 184–197.

[132] Lukas M. Weber and Mark D. Robinson. "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data". In: *Cytometry Part A* 89.12 (2016), pp. 1084–1096. ISSN: 15524930. DOI: 10.1002/cyto.a.23030.

[133] Xiao Liu et al. "A comparison framework and guideline of clustering methods for mass cytometry data". In: *Genome Biology* 20.1 (2019), pp. 1–18.

[134] Nikolay Samusik, Zinaida Good, Matthew H. Spitzer, Kara L. Davis, and Garry P. Nolan. "Automated mapping of phenotype space with single-cell data". In: *Nature Methods* 13.6 (2016), pp. 493–496. ISSN: 15487105. DOI: 10.1038/nmeth.3863.

[135] Lukas M. Weber, Malgorzata Nowicka, Charlotte Soneson, and Mark D. Robinson. "diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering". In: *Communications Biology* 2.1 (2019). ISSN: 23993642. DOI: 10.1038/s42003-019-0415-5. URL: http://dx.doi.org/10.1038/s42003-019-0415-5.

[136] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of Machine Learning Research* 9.11 (2008).

[137] Etienne Becht et al. "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature Biotechnology* 37.1 (Jan. 2019), pp. 38–44. ISSN: 1546-1696. DOI: 10.1038/nbt.4314. URL: https://doi.org/10.1038/nbt.4314.

[138] Kevin R Moon et al. "Visualizing structure and transitions in high-dimensional biological data". In: *Nature Biotechnology* 37.12 (2019), pp. 1482–1492.

[139] Robert V. Bruggner, Bernd Bodenmiller, David L. Dill, Robert J. Tibshirani, and Garry P. Nolan. "Automated identification of stratifying signatures in cellular subpopulations". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.26 (2014). ISSN: 10916490. DOI: 10.1073/pnas.1408792111.

[140] Hao Chen et al. "Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline". In: *PLOS Computational Biology* 12.9 (Sept. 2016), pp. 1–17. DOI: 10.1371/journal.pcbi.1005112. URL: https://doi.org/10.1371/journal.pcbi.1005112.

[141] Yuting Dai et al. "CytoTree: an R/Bioconductor package for analysis and visualization of flow and mass cytometry data". In: *BMC Bioinformatics* 22.1 (2021), pp. 1–20.

[142] Gianni Monaco et al. "flowAI: automatic and interactive anomaly discerning tools for flow cytometry data". In: *Bioinformatics* 32.16 (2016), pp. 2473–2480. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw191. eprint: https://academic.oup.com/bioinformatics/article-pdf/32/16/2473/39715346/bioinformatics_32_16_2473.pdf. URL: https://doi.org/10.1093/bioinformatics/btw191.

[143] Kipper Fletez-Brant, Josef Špidlen, Ryan R Brinkman, Mario Roederer, and Pratip K Chattopadhyay. "flowClean: Automated identification and removal of fluorescence anomalies in flow cytometry data". In: *Cytometry Part A* 89.5 (2016), pp. 461–471.

[144] Annelies Emmaneel et al. "PeacoQC: Peak-based selection of high quality cytometry data". In: *Cytometry Part A* 101.4 (2022), pp. 325–338.

[145] *Stack Overflow Developer Survey 2019*. URL: https://insights.stackoverflow.com/survey/2019.

[146] *Stack Overflow Developer Survey 2020*. URL: https://insights.stackoverflow.com/survey/2020.

[147] *Stack Overflow Developer Survey 2021*. URL: https://insights.stackoverflow.com/survey/2021.

[148] Peter JA Cock et al. "Biopython: freely available Python tools for computational molecular biology and bioinformatics". In: *Bioinformatics* 25.11 (2009), pp. 1422–1423.

[149] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome Biology* 19.1 (2018), pp. 1–5.

[150] Scott White et al. "FlowKit: A Python Toolkit for Integrated Manual and Automated Cytometry Analysis Workflows". In: *Frontiers in Immunology* (2021), p. 4652.

[151] Brian Teague. "Cytoflow: A Python Toolbox for Flow Cytometry". In: *bioRxiv* (2022).

[152] Martın Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[153] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in Neural Information Processing Systems* 32 (2019).

[154] Savitri Kibe, Kate Adams, and Gavin Barlow. "Diagnostic and prognostic biomarkers of sepsis in critical care". In: *Journal of antimicrobial chemotherapy* 66.suppl_2 (2011), pp. ii33–ii40.

[155] Charalampos Pierrakos and Jean-Louis Vincent. "Sepsis biomarkers: a review". In: *Critical Care* 14.1 (2010), pp. 1–18.

[156] Ron Kohavi and George H John. "Wrappers for feature subset selection". In: *Artificial Intelligence* 97.1-2 (1997), pp. 273–324.

[157] Roland Nilsson, José M Pena, Johan Björkegren, and Jesper Tegnér. "Consistent feature selection for pattern recognition in polynomial time". In: *The Journal of Machine Learning Research* 8 (2007), pp. 589–612.

[158] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Vol. 112. Springer, 2013.

[159] Yvan Saeys, Inaki Inza, and Pedro Larranaga. "A review of feature selection techniques in bioinformatics". In: *Bioinformatics* 23.19 (2007), pp. 2507–2517.

[160] Beatriz Remeseiro and Veronica Bolon-Canedo. "A review of feature selection methods in medical applications". In: *Computers in Biology and Medicine* 112 (2019), p. 103375.

[161] Lipo Wang, Yaoli Wang, and Qing Chang. "Feature selection methods for big data bioinformatics: a survey from the search perspective". In: *Methods* 111 (2016), pp. 21–31.

[162] Zhiao Shi, Bo Wen, Qiang Gao, and Bing Zhang. "Feature selection methods for protein biomarker discovery from proteomics or multiomics data". In: *Molecular & Cellular Proteomics* 20 (2021).

[163] Pengyi Yang, Hao Huang, and Chunlei Liu. "Feature selection revisited in the single-cell era". In: *Genome Biology* 22.1 (2021), pp. 1–17.

[164] Sabyasachi Bandyopadhyay et al. "Discovery and validation of urinary molecular signature of early sepsis". In: *Critical Care Explorations* 2.10 (2020).

[165] U Parthasarathy et al. "Novel neutrophil subsets associated with sepsis, vascular dysfunction and metabolic alterations identified using systems immunology". In: *bioRxiv* (2022).

[166] Christoph Molnar. *Interpretable Machine Learning*. Lulu.com, 2020.

[167] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[168] Chang Hu et al. "Interpretable Machine Learning for Early Prediction of Prognosis in Sepsis: A Discovery and Validation Study". In: *Infectious Diseases and Therapy* (2022), pp. 1–16.

[169] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[170] Leonard Richardson. “Beautiful soup documentation”. In: *Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018]* (2007).

[171] Pauli Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature Methods* 17.3 (2020), pp. 261–272.

[172] Paul G. Gottschalk and John R. Dunn. “The five-parameter logistic: A characterization and comparison with the four-parameter logistic”. In: *Analytical Biochemistry* 343.1 (2005), pp. 54–65. ISSN: 0003-2697. DOI: https://doi.org/10.1016/j.ab.2005.04.035. URL: https://www.sciencedirect.com/science/article/pii/S0003269705003313.

[173] Adriana Tomic, Ivan Tomic, Cornelia L Dekker, Holden T Maecker, and Mark M Davis. “The FluPRINT dataset, a multidimensional analysis of the influenza vaccine imprint on the immune system”. In: *Scientific Data* 6.1 (2019), pp. 1–10.

[174] Chan C. Whiting et al. “Large-Scale and Comprehensive Immune Profiling and Functional Analysis of Normal Human Aging”. In: *PLOS ONE* 10.7 (July 2015), pp. 1–21. DOI: 10.1371/journal.pone.0133627. URL: https://doi.org/10.1371/journal.pone.0133627.

[175] Chan-Yu Lin et al. “Pathogen-Specific Local Immune Fingerprints Diagnose Bacterial Infection in Peritoneal Dialysis Patients”. In: *Journal of the American Society of Nephrology* 24.12 (2013), pp. 2002–2009. DOI: 10.1681/asn.2013040332.

[176] Nima Aghaeepour et al. “A benchmark for evaluation of algorithms for identification of cellular correlates of clinical outcomes”. In: *Cytometry Part A* 89.1 (2016), pp. 16–21. ISSN: 15524930. DOI: 10.1002/cyto.a.22732.

[177] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[178] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.

[179] Raphael Vallat. “Pingouin: statistics in Python.” In: *Journal of Open Source Software* 3.31 (2018), p. 1026.

[180] J. D. Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[181] Michael L. Waskom. “seaborn: statistical data visualization”. In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

[182] LM Hollestein et al. “MULTIPLE ways to correct for MULTIPLE comparisons in MULTIPLE types of studies”. In: *British Journal of Dermatology* 185.6 (2021), pp. 1081–1083.

[183] Tamim Abdelaal et al. “Predicting Cell Populations in Single Cell Mass Cytometry Data”. In: *Cytometry Part A* 95.7 (2019), pp. 769–781. ISSN: 15524930. DOI: 10.1002/cyto.a.23738.

[184] Zicheng Hu, Benjamin S. Glicksberg, and Atul J. Butte. “Robust prediction of clinical outcomes using cytometry data”. In: *Bioinformatics* 35.7 (2019), pp. 1197–1203. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty768.

[185] Florian Mair et al. “The end of gating? An introduction to automated analysis of high dimensional cytometry data”. In: *European Journal of Immunology* 46.1 (2016), pp. 34–43. ISSN: 15214141. DOI: 10.1002/eji.201545774.

[186] Ryan R. Brinkman et al. “Automated analysis of flow cytometry data comes of age”. In: *Cytometry Part A* 89.1 (2016), pp. 13–15. ISSN: 15524930. DOI: 10.1002/cyto.a.22810.

[187] Jeff Reback et al. *pandas-dev/pandas: Pandas 1.2.2*. Feb. 2021. DOI: 10.5281/ZENODO.4524629. URL: https://zenodo.org/record/4524629.

[188] MongoDB Inc. *MongoDB: The most popular database for modern apps*. URL: %5Curl%7Bhttps://www.mongodb.com/%7D.

[189] MongoEngine Inc. *MongoEngine: an document-object wrapper for MongoDB*. URL: %5Curl%7Bhttps://github.com/mongoengine%7D.

[190] Scott White, Dav Hau, and Lorenz Gerber. *whitews/FlowUtils: 0.9.5*. Version 0.9.5. Sept. 2021. DOI: 10.5281/zenodo.5510713. URL: https://doi.org/10.5281/zenodo.5510713.

[191] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. "On the shape of a set of points in the plane". In: *IEEE Transactions on Information Theory* 29.4 (1983), pp. 551–559. DOI: 10.1109/TIT.1983.1056714.

[192] Tommy Odland. *tommyod/KDEpy: Kernel Density Estimation in Python*. Dec. 2018. DOI: 10.5281/zenodo.2392268.

[193] Florian Hahne et al. "Per-channel basis normalization methods for flow cytometry data". In: *Cytometry Part A* 77A.2 (2010), pp. 121–131. DOI: https://doi.org/10.1002/cyto.a.20823. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.a.20823. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.20823.

[194] Greg Finak et al. "High-throughput flow cytometry data normalization for clinical trials". In: *Cytometry Part A* 85.3 (2014), pp. 277–286. ISSN: 15524922. DOI: 10.1002/cyto.a.22433.

[195] Hoa Thi Nhu Tran et al. "A benchmark of batch-effect correction methods for single-cell RNA sequencing data". In: *Genome Biology* 21.1 (2020), p. 12. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1850-9. URL: https://doi.org/10.1186/s13059-019-1850-9.

[196] Ilya Korsunsky et al. "Fast, sensitive and accurate integration of single-cell data with Harmony". In: *Nature Methods* 16.12 (2019), pp. 1289–1296. ISSN: 15487105. DOI: 10.1038/s41592-019-0619-0. URL: http://dx.doi.org/10.1038/s41592-019-0619-0.

[197] Kamil Slowikowski. *Harmonypy*. 2020. URL: http://doi.org/10.5281/zenodo.4531401.

[198] François Chollet et al. *Keras*. https://keras.io. 2015.

[199] Jingjing Zhang et al. "Machine-learning algorithms define pathogen-specific local immune fingerprints in peritoneal dialysis patients with bacterial infections". In: *Kidney International* 92.1 (2017), pp. 179–191. DOI: 10.1016/j.kint.2017.01.017.

[200] Chia-Te Liao et al. "Peritoneal macrophage heterogeneity is associated with different peritoneal dialysis outcomes". In: *Kidney International* 91.5 (2017), pp. 1088–1103. DOI: 10.1016/j.kint.2016.10.030.

[201] Melissa Cheung et al. "Current trends in flow cytometry automated data analysis software". In: *Cytometry Part A* 99.10 (2021), pp. 1007–1021. DOI: https://doi.org/10.1002/cyto.a.24320. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.a.24320. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.24320.

[202] Christina Bligaard Pedersen and Lars Rønn Olsen. "Algorithmic Clustering Of Single-Cell Cytometry Data—How Unsupervised Are These Analyses Really?" In: *Cytometry Part A* 97.3 (2020), pp. 219–221. DOI: https://doi.org/10.1002/cyto.a.23917. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.a.23917. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.23917.

[203] Joydeep Ghosh and Ayan Acharya. "Cluster ensembles". In: *WIREs Data Mining and Knowledge Discovery* 1.4 (2011), pp. 305–315. DOI: https://doi.org/10.1002/widm.32. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.32. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.32.

[204] Tom Ronan, Zhijie Qi, and Kristen M. Naegle. "Avoiding common pitfalls when clustering biological data". In: *Science Signaling* 9.432 (2016), re6–re6. DOI: 10.1126/scisignal.aad1932. eprint: https://www.science.org/doi/pdf/10.1126/scisignal.aad1932. URL: https://www.science.org/doi/abs/10.1126/scisignal.aad1932.

[205] Tossapon Boongoen and Natthakan Iam-On. "Cluster ensembles: A survey of approaches with recent extensions and applications". In: *Computer Science Review* 28 (2018), pp. 1–25. ISSN: 1574-0137. DOI: https://doi.org/10.1016/j.cosrev.2018.01.003. URL: https://www.sciencedirect.com/science/article/pii/S1574013717300692.

[206] Sandro Vega-Pons and José Ruiz-Shulcloper. "A Survey of Clustering Ensemble Algorithms". In: *Int. J. Pattern Recognit. Artif. Intell.* 25 (2011), pp. 337–372.

[207] Alexander Strehl and Joydeep Ghosh. "Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions". In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 583–617. ISSN: 1532-4435. DOI: 10.1162/153244303321897735. URL: https://doi.org/10.1162/153244303321897735.

[208] Xiaoli Zhang Fern and Carla E. Brodley. "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning". In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada: Association for Computing Machinery, 2004, p. 36. ISBN: 1581138385. DOI: 10.1145/1015330.1015414. URL: https://doi.org/10.1145/1015330.1015414.

[209] Xiaoshu Zhu, Jian Li, Hong-Dong Li, Miao Xie, and Jianxin Wang. "Sc-gpe: A graph partitioning-based cluster ensemble method for single-cell". In: *Frontiers in Genetics* 11 (2020), p. 604790.

[210] Vladimir Yu Kiselev et al. "SC3: consensus clustering of single-cell RNA-seq data". In: *Nature Methods* 14.5 (2017), pp. 483–486. ISSN: 1548-7105. DOI: 10.1038/nmeth.4236. URL: https://doi.org/10.1038/nmeth.4236.

[211] Yuchen Yang et al. "SAFE-clustering: Single-cell Aggregated (from Ensemble) clustering for single-cell RNA-seq data". In: *Bioinformatics* 35.8 (Sept. 2018), pp. 1269–1277. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty793. eprint: https://academic.oup.com/bioinformatics/article-pdf/35/8/1269/28500735/bty793.pdf. URL: https://doi.org/10.1093/bioinformatics/bty793.

[212] Dale Roberts, Norman Mueller, and Alexis Mcintyre. "High-Dimensional Pixel Composites From Earth Observation Time Series". In: *IEEE Transactions on Geoscience and Remote Sensing* 55.11 (2017), pp. 6254–6264. DOI: 10.1109/TGRS.2017.2723896.

[213] Matthew D. Wilkerson and D. Neil Hayes. "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking". In: *Bioinformatics* 26.12 (Apr. 2010), pp. 1572–1573. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq170. eprint: https://academic.oup.com/bioinformatics/article-pdf/26/12/1572/16893326/btq170.pdf. URL: https://doi.org/10.1093/bioinformatics/btq170.

[214] Shobana V Stassen et al. "PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells". In: *Bioinformatics* 36.9 (Jan. 2020), pp. 2778–2786. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa042. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/9/2778/33539635/btaa042.pdf. URL: https://doi.org/10.1093/bioinformatics/btaa042.

[215] Florian Mair and Martin Prlic. "OMIP-044: 28-color immunophenotyping of the human dendritic cell compartment". In: *Cytometry Part A* 93.4 (2018), pp. 402–405.

[216] Ross Jake Burton, Simone Cuff, Matthew Morgan, Andreas Artemiou, and Matthias Eberl. *GeoWaVe Cytometry Benchmark Data*. Version 1.0. Oct. 2022. DOI: 10.5281/zenodo.7134723. URL: https://doi.org/10.5281/zenodo.7134723.

[217] Takehiro Sano. *ClusterEnsembles*. Aug. 2021. URL: https://github.com/tsano430/ClusterEnsembles.

[218] Vivek Mehta, Seema Bawa, and Jasmeet Singh. "Analytical review of clustering techniques and proximity measures". In: *Artificial Intelligence Review* 53.8 (2020), pp. 5995–6023.

[219] Xingyu Yang and Peng Qiu. "Automatically generate two-dimensional gating hierarchy from clustered cytometry data". In: *Cytometry Part A* 93.10 (2018), pp. 1039–1050. ISSN: 15524930. DOI: 10.1002/cyto.a.23577.

[220] Yaxuan Cui et al. "Consensus clustering of single-cell RNA-seq data by enhancing network affinity". In: *Briefings in Bioinformatics* 22.6 (June 2021). bbab236. ISSN: 1477-4054. DOI: 10.1093/bib/bbab236. eprint: https://academic.oup.com/bib/article-pdf/22/6/bbab236/41088895/bbab236.pdf. URL: https://doi.org/10.1093/bib/bbab236.

[221] Fu Xiang Quah and Martin Hemberg. "SC3s - efficient scaling of single cell consensus clustering to millions of cells". In: *bioRxiv* (2021). DOI: 10.1101/2021.05.20.445027. eprint:

https://www.biorxiv.org/content/early/2021/05/22/2021.05.20.445027.full.pdf. URL: https://www.biorxiv.org/content/early/2021/05/22/2021.05.20.445027.

[222] Soufiane Khedairia and Mohamed Tarek Khadir. "A multiple clustering combination approach based on iterative voting process". In: *Journal of King Saud University - Computer and Information Sciences* 34.1 (2022), pp. 1370–1380. ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2019.09.013. URL: https://www.sciencedirect.com/science/article/pii/S131915781930597X.

[223] Tiffany Purcell Pellathy, Michael R Pinsky, and Marilyn Hravnak. "Intensive Care Unit Scoring Systems". In: *Critical Care Nurse* 41.4 (2021), pp. 54–64.

[224] Robert Goulden et al. "qSOFA, SIRS and NEWS for predicting inhospital mortality and ICU admission in emergency admissions treated as sepsis". In: *Emergency Medicine Journal* 35.6 (2018), pp. 345–349.

[225] Rodrigo Serafim, Jose Andrade Gomes, Jorge Salluh, and Pedro Póvoa. "A comparison of the quick-SOFA and systemic inflammatory response syndrome criteria for the diagnosis of sepsis and prediction of mortality: a systematic review and meta-analysis". In: *Chest* 153.3 (2018), pp. 646–655.

[226] Özkan Devran et al. "C-reactive protein as a predictor of mortality in patients affected with severe sepsis in intensive care unit". In: *Multidisciplinary Respiratory Medicine* 7.1 (2012), pp. 1–6.

[227] Z Zhang and H Ni. "C-reactive protein as a predictor of mortality in critically ill patients: a meta-analysis and systematic review". In: *Anaesthesia and Intensive Care* 39.5 (2011), pp. 854–861.

[228] Emanuela Sozio et al. "MR-proADM as prognostic factor of outcome in COVID-19 patients". In: *Scientific Reports* 11.1 (2021), pp. 1–7.

[229] Qiqi Chen et al. "Neutrophil CD64 expression is a predictor of mortality for patients in the intensive care unit". In: *International journal of Clinical and Experimental Pathology* 7.11 (2014), p. 7806.

[230] Wan Fadzlina Wan Muhd Shukeri, Azrina Md Ralib, Nor Zamzila Abdulah, and Mohd Basri Mat-Nor. "Sepsis mortality score for the prediction of mortality in septic patients". In: *Journal of Critical Care* 43 (2018), pp. 163–168.

[231] Joshua David Farkas. "The complete blood count to diagnose septic shock". In: *Journal of Thoracic Disease* 12.Suppl 1 (2020), S16.

[232] Jean-Marc Cavaillon and Minou Adib-Conquy. "Immune status in sepsis: the bug, the site of infection and the severity can make the difference". In: *Critical Care* 14.3 (2010), pp. 1–2.

[233] Ricard Ferrer et al. "Effectiveness of treatments for severe sepsis: a prospective, multicenter, observational study". In: *American journal of Respiratory and Critical Care Medicine* 180.9 (2009), pp. 861–866.

[234] Margaret Disselkamp, Angel O Coz Yataco, and Steven Q Simpson. "POINT: should broad-spectrum antibiotics be routinely administered to all patients with sepsis as soon as possible? Yes". In: *Chest* 156.4 (2019), pp. 645–647.

[235] Jayshil J Patel and Paul A Bergl. "COUNTERPOINT: should broad-spectrum antibiotics be routinely administered to all patients with sepsis as soon as possible? No". In: *Chest* 156.4 (2019), pp. 647–649.

[236] Sarah B Doernberg. "Will biomarkers be the answer for antibiotic stewardship?" In: *The Lancet Respiratory Medicine* 8.2 (2020), pp. 130–132.

[237] Shuhua Li et al. "Serum procalcitonin levels distinguish Gram-negative bacterial sepsis from Gram-positive bacterial and fungal sepsis". In: *Journal of Research in Medical Sciences* 21 (2016).

[238] Shun Yuan Guo, Yin Zhou, Qing Feng Hu, Jiong Yao, and Hong Wang. "Procalcitonin is a marker of gram-negative bacteremia in patients with sepsis". In: *The American Journal of the Medical Sciences* 349.6 (2015), pp. 499–504.

[239] Erin W Meermeier, Melanie J Harriff, Elham Karamooz, and David M Lewinsohn. "MAIT cells and microbial immunity". In: *Immunology and Cell Biology* 96.6 (2018), pp. 607–617.

[240] Jingjing Zhang et al. "Machine-learning algorithms define pathogen-specific local immune fingerprints in peritoneal dialysis patients with bacterial infections". In: *Kidney International* 92.1 (2017), pp. 179–191.

[241] Francesca Parrott. *Length of stay, survival and organ support of admissions with septic shock to adult, general critical care units in England, Wales and Northern Ireland*. Tech. rep. Napier House, 24 High Holborn, London, WC1V 6AZ: Intensive Care National Audit Research Centre (ICNARC), Aug. 2014.

[242] Tamas Szakmany et al. "Sepsis prevalence and outcome on the general wards and emergency departments in Wales: results of a multi-centre, observational, point prevalence study". In: *PLOS One* 11.12 (2016), e0167230.

[243] Mervyn Singer et al. "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)". In: *JAMA* 315.8 (Feb. 2016), pp. 801–810. ISSN: 0098-7484. DOI: 10.1001/jama.2016.0287. eprint: https://jamanetwork.com/journals/jama/articlepdf/2492881/jsc160002.pdf. URL: https://doi.org/10.1001/jama.2016.0287.

[244] Benjamin Nolt et al. "Lactate and immunosuppression in sepsis". en. In: *Shock* 49.2 (Feb. 2018), pp. 120–125.

[245] RM Dahl et al. "Variability in targeted arterial oxygenation levels in patients with severe sepsis or septic shock". In: *Acta Anaesthesiologica Scandinavica* 59.7 (2015), pp. 859–869.

[246] Zlatko Dembic. *Cytokines of the Immune System*. Elsevier Science, Jan. 2015.

[247] Ken Shortman and Shalin H Naik. "Steady-state and inflammatory dendritic-cell development". In: *Nature Reviews Immunology* 7.1 (2007), pp. 19–30.

[248] Joseph A. Carcillo, Kate K. Kernan, Christopher M. Horvat, Dennis W. Simon, and Rajesh K. Aneja. "Why and How Is Hyperferritinemic Sepsis Different From Sepsis Without Hyperferritinemia?∗". In: *Pediatric Critical Care Medicine* 21.5 (May 2020), pp. 509–512. DOI: 10.1097/pcc.0000000000002285. URL: https://doi.org/10.1097/pcc.0000000000002285.

[249] Steven Tenny and Mary R Hoffman. *Odds Ratio - StatPearls [Internet]*. May 2021. URL: https://www.ncbi.nlm.nih.gov/books/NBK431098/.

[250] Hannah Kaminski, Lionel Couzi, and Matthias Eberl. "Unconventional T cells and kidney disease". In: *Nature Reviews Nephrology* 17.12 (Dec. 2021), pp. 795–813. ISSN: 1759-507X. DOI: 10.1038/s41581-021-00466-8. URL: https://doi.org/10.1038/s41581-021-00466-8.

[251] A Lekkou, M Karakantza, A Mouzaki, F Kalfarentzos, and CA Gogos. "Cytokine production and monocyte HLA-DR expression as predictors of outcome for patients with community-acquired severe infections". In: *Clinical and Vaccine Immunology* 11.1 (2004), pp. 161–167.

[252] Yugang Zhuang, Hu Peng, Yuanzhuo Chen, Shuqin Zhou, and Yanqing Chen. "Dynamic monitoring of monocyte HLA-DR expression for the diagnosis, prognosis, and prediction of sepsis". In: *Frontiers in Bioscience-Landmark* 22.8 (2017), pp. 1344–1354.

[253] Jean-Marc Cavaillon and Minou Adib-Conquy. "Monocytes/macrophages and sepsis". In: *Critical care medicine* 33.12 (2005), S506–S509.

[254] Nicholas A Gherardin et al. "Human blood MAIT cell subsets defined using MR1 tetramers". In: *Immunology and Cell Biology* 96.5 (2018), pp. 507–525.

[255] Nadia Caccamo, Simone A Joosten, Tom HM Ottenhoff, and Francesco Dieli. "Atypical human effector/memory CD4+ T cells with a naive-like phenotype". In: *Frontiers in Immunology* 9 (2018), p. 2832.

[256] Matthew D Martin and Vladimir P Badovinac. "Defining memory CD8 T cell". In: *Frontiers in immunology* 9 (2018), p. 2692.

[257] Joannah R Fergusson, Vicki M Fleming, and Paul Klenerman. "CD161-expressing human T cells". In: *Frontiers in immunology* 2 (2011), p. 36.

[258] Fulvio D'Acquisto and Tessa Crompton. "CD3+ CD4- CD8-(double negative) T cells: saviours or villains of the immune response?" In: *Biochemical pharmacology* 82.4 (2011), pp. 333–340.

[259] Anna Rita Liuzzi et al. "Unconventional human T cells accumulate at the site of infection in response to microbial ligands and induce local tissue remodeling". In: *The Journal of Immunology* 197.6 (2016), pp. 2195–2207.

[260] Cristina Rosario, Gisele Zandman-Goddard, Esther G Meyron-Holtz, David P D'Cruz, and Yehuda Shoenfeld. "The Hyperferritinemic Syndrome: macrophage activation syndrome, Still's disease, septic shock and catastrophic antiphospholipid syndrome". In: *BMC Medicine* 11.1 (Aug. 2013). DOI: 10.1186/1741-7015-11-185. URL: https://doi.org/10.1186/1741-7015-11-185.

[261] Kris Bauchmuller et al. "Haemophagocytic lymphohistiocytosis in adult critical care". In: *Journal of the Intensive Care Society* 21.3 (2020), pp. 256–268.

[262] Monika Gudowska-Sawczuk and Barbara Mroczko. "What Is Currently Known about the Role of CXCL10 in SARS-CoV-2 Infection?" In: *International Journal of Molecular Sciences* 23.7 (2022), p. 3673.

[263] Xianan Wu, Xiaofei Lai, Hongmei Tu, Hua Zou, and Ju Cao. "Elevated serum CXCL10 in patients with Clostridium difficile infection are associated with disease severity". In: *International Immunopharmacology* 72 (2019), pp. 92–97.

[264] Mengyao Li et al. "Serum CXCL10/IP-10 may be a potential biomarker for severe Mycoplasma pneumoniae pneumonia in children". In: *BMC Infectious Diseases* 21.1 (2021), pp. 1–8.

[265] Daniela S Herzig et al. "Regulation of lymphocyte trafficking by CXC chemokine receptor 3 during septic shock". In: *American journal of respiratory and critical care medicine* 185.3 (2012), pp. 291–300.

[266] Yin Guo et al. "IL-15 enables septic shock by maintaining NK cell integrity and function". In: *The Journal of Immunology* 198.3 (2017), pp. 1320–1333.

[267] Masafumi Saito et al. "IL-15 improves aging-induced persistent T cell exhaustion in mouse models of repeated sepsis". In: *Shock* 53.2 (2020), pp. 228–235.

[268] Shigeaki Inoue et al. "Reduction of immunocompetent T cells followed by prolonged lymphopenia in severe sepsis in the elderly". In: *Critical Care Medicine* 41.3 (2013), pp. 810–819.

[269] Javier Cabrera-Perez, Stephanie A Condotta, Vladimir P Badovinac, and Thomas S Griffith. "Impact of sepsis on CD4 T cell immunity". In: *Journal of Leukocyte Biology* 96.5 (2014), pp. 767–777.

[270] Thomas Rimmelé et al. "Immune cell phenotype and function in sepsis". In: *Shock (Augusta, Ga.)* 45.3 (2016), p. 282.

[271] Derek B Danahy, Robert K Strother, Vladimir P Badovinac, and Thomas S Griffith. "Clinical and experimental sepsis impairs CD8 T-cell-mediated immunity". In: *Critical Reviews in Immunology* 36.1 (2016).

[272] Jakob B Seidelin, Ole H Nielsen, and Jens Strøm. "Soluble L-selectin levels predict survival in sepsis". In: *Intensive Care Medicine* 28.11 (2002), pp. 1613–1618.

[273] Stefano Gambardella et al. "ccf-mtDNA as a potential link between the brain and immune system in neuro-immunological disorders". In: *Frontiers in Immunology* 10 (2019), p. 1064.

[274] Joana Dias, Edwin Leeansyah, and Johan K Sandberg. "Multiple layers of heterogeneity and subset diversity in human MAIT cell responses to distinct microorganisms and to innate cytokines". In: *Proceedings of the National Academy of Sciences* 114.27 (2017), E5434–E5443.

[275] Matthias Eberl et al. “Microbial isoprenoid biosynthesis and human $\gamma\delta$ T cell activation”. In: *FEBS Letters* 544.1-3 (2003), pp. 4–10.

[276] Nadine Hartmann et al. “Role of MAIT cells in pulmonary bacterial infection”. In: *Molecular Immunology* 101 (2018), pp. 155–159.

[277] Zhenjun Chen et al. “Mucosal-associated invariant T-cell activation and accumulation after in vivo infection depends on microbial riboflavin synthesis and co-stimulatory signals”. In: *Mucosal Immunology* 10.1 (2017), pp. 58–68.

[278] Harm-Jan de Grooth et al. “Unexplained mortality differences between septic shock trials: a systematic analysis of population characteristics and control-group mortality rates”. In: *Intensive Care Medicine* 44.3 (2018), pp. 311–322.

[279] Natalja L Stanski and Hector R Wong. “Prognostic and predictive enrichment in sepsis”. In: *Nature Reviews Nephrology* 16.1 (2020), pp. 20–31.

[280] Tom van der Poll, Manu Shankar-Hari, and W Joost Wiersinga. “The Immunology of Sepsis”. In: *Immunity* 54.11 (2021), pp. 2450–2464.

[281] Alison E Fohner et al. “Assessing clinical heterogeneity in sepsis through treatment patterns and machine learning”. In: *Journal of the American Medical Informatics Association* 26.12 (2019), pp. 1466–1477.

[282] Christopher W Seymour et al. “Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis”. In: *JAMA* 321.20 (2019), pp. 2003–2017.

[283] Chanu Rhee et al. “Prevalence of antibiotic-resistant pathogens in culture-proven sepsis and outcomes associated with inadequate and broad-spectrum empiric antibiotic use”. In: *JAMA Network Open* 3.4 (2020), e202899–e202899.

[284] Tjitske SR Van Engelen, Willem Joost Wiersinga, Brendon P Scicluna, and Tom van der Poll. “Biomarkers in sepsis”. In: *Critical Care Clinics* 34.1 (2018), pp. 139–152.

[285] Meera Joshi et al. “Digital alerting and outcomes in patients with sepsis: systematic review and meta-analysis”. In: *Journal of Medical Internet Research* 21.12 (2019), e15166.

[286] Anil N Makam, Oanh K Nguyen, and Andrew D Auerbach. “Diagnostic accuracy and effectiveness of automated electronic sepsis alert systems: a systematic review”. In: *Journal of Hospital Medicine* 10.6 (2015), pp. 396–402.

[287] Poushali Bhattacharjee, Dana P Edelson, and Matthew M Churpek. “Identifying patients with sepsis on the hospital wards”. In: *Chest* 151.4 (2017), pp. 898–907.

[288] Thomas Desautels et al. “Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach”. In: *JMIR Medical Informatics* 4.3 (2016), e5909.

[289] Shamim Nemati et al. “An interpretable machine learning model for accurate prediction of sepsis in the ICU”. In: *Critical Care Medicine* 46.4 (2018), p. 547.

[290] Md Mohaimenul Islam et al. “Prediction of sepsis patients using machine learning approach: a meta-analysis”. In: *Computer Methods and Programs in Biomedicine* 170 (2019), pp. 1–9.

[291] Ryan J Delahanty, JoAnn Alvarez, Lisa M Flynn, Robert L Sherwin, and Spencer S Jones. “Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis”. In: *Annals of Emergency Medicine* 73.4 (2019), pp. 334–344.

[292] Nathan I Shapiro et al. “A prospective, multicenter derivation of a biomarker panel to assess risk of organ dysfunction, shock, and death in emergency department patients with suspected sepsis”. In: *Critical Care Medicine* 37.1 (2009), pp. 96–104.

[293] Fernando A Bozza et al. “Cytokine profiles as markers of disease severity in sepsis: a multiplex analysis”. In: *Critical Care* 11.2 (2007), pp. 1–8.

[294] Vimal Grover et al. “A biomarker panel (Bioscore) incorporating monocytic surface and soluble TREM-1 has high discriminative value for ventilator-associated pneumonia: a prospective observational study”. In: *PLOS One* 9.10 (2014), e109686.

[295] Raymond J Langley et al. “An integrated clinico-metabolomic model improves prediction of death in sepsis”. In: *Science Translational Medicine* 5.195 (2013), 195ra95–195ra95.

[296] Hsiao-Yun Chao et al. “Using Machine Learning to Develop and Validate an In-Hospital Mortality Prediction Model for Patients with Suspected Sepsis”. In: *Biomedicines* 10.4 (2022), p. 802.

[297] Sherrianne Ng et al. “Precision medicine for neonatal sepsis”. In: *Frontiers in Molecular Biosciences* 5 (2018), p. 70.

[298] Rebecca A Ward et al. “Harnessing the potential of multiomics studies for precision medicine in infectious disease”. In: *Open Forum Infectious Diseases*. Vol. 8. 11. Oxford University Press US. 2021, ofab483.

[299] Isabelle Guyon and André Elisseeff. “An introduction to variable and feature selection”. In: *Journal of Machine Learning Research* 3.Mar (2003), pp. 1157–1182.

[300] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. “Do we need hundreds of classifiers to solve real world classification problems?” In: *Journal of Machine Learning Research* 15 (2014), pp. 3133–3181. ISSN: 15337928. DOI: 10.1117/1.JRS.11.015020.

[301] David H Wolpert and William G Macready. “No free lunch theorems for optimization”. In: *IEEE Transactions on Evolutionary Computation* 1.1 (1997), pp. 67–82.

[302] Stephen-John Sammut et al. “Multi-omic machine learning predictor of breast cancer therapy response”. In: *Nature* 601.7894 (2022), pp. 623–629.

[303] Kelsey R Dean et al. “Multi-omic biomarker identification and validation for diagnosing warzone-related post-traumatic stress disorder”. In: *Molecular Psychiatry* 25.12 (2020), pp. 3337–3349.

[304] Junru Wu et al. “Multi-omic analysis in injured humans: Patterns align with outcomes and treatment responses”. In: *Cell Reports Medicine* 2.12 (2021), p. 100478.

[305] Won-Young Kim. “Multi-omic approach to identify risk markers specific to COVID-19”. In: *EBioMedicine* 79 (2022).

[306] Katherine A Overmyer et al. “Large-scale multi-omic analysis of COVID-19 severity”. In: *Cell Systems* 12.1 (2021), pp. 23–40.

[307] Bowen Fan et al. “Prediction of recovery from multiple organ dysfunction syndrome in pediatric sepsis patients”. In: *Bioinformatics* 38.Supplement_1 (2022), pp. i101–i108.

[308] Joshua E Lewis and Melissa L Kemp. “Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance”. In: *Nature Communications* 12.1 (2021), pp. 1–14.

[309] Guoxing Tang et al. “Prediction of sepsis in COVID-19 using laboratory indicators”. In: *Frontiers in Cellular and Infection Microbiology* 10 (2021), p. 586054.

[310] Gustavo EAPA Batista, Ana LC Bazzan, Maria Carolina Monard, et al. “Balancing Training Data for Automated Annotation of Keywords: a Case Study.” In: *WOB*. 2003, pp. 10–18.

[311] R. Blagus and L. Lusa. “SMOTE for high-dimensional class-imbalanced data”. In: *BMC Bioinformatics* 14 (Mar. 2013), p. 106.

[312] Haibo He and Yunqian Ma. *Imbalanced Learning*. July 2013. ISBN: 9781118074626.

[313] Jin Huang and C.X. Ling. “Using AUC and accuracy in evaluating learning algorithms”. In: *IEEE Transactions on Knowledge and Data Engineering* 17.3 (2005), pp. 299–310. DOI: 10.1109/TKDE.2005.50.

[314] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Aug. 2022. Chap. 1.4. ISBN: 9788132209065.

[315] D. J. Stekhoven and P. Bühlmann. “MissForest–non-parametric missing value imputation for mixed-type data”. In: *Bioinformatics* 28.1 (Jan. 2012), pp. 112–118.

[316] Ian R White, Patrick Royston, and Angela M Wood. "Multiple imputation using chained equations: issues and guidance for practice". In: *Statistics in Medicine* 30.4 (2011), pp. 377–399.

[317] Michael Mayer. "missRanger: Fast Imputation of Missing Values". In: (2019). R package version 2.1.3.

[318] Kenneth Chi-Yin Wong, Yong Xiang, Liangying Yin, and Hon-Cheong So. "Uncovering Clinical Risk Factors and Predicting Severe COVID-19 Cases Using UK Biobank Data: Machine Learning Approach". In: *JMIR Public Health and Surveillance* 7.9 (2021), e29544.

[319] Sebastian Raschka. "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack". In: *The Journal of Open Source Software* 3.24 (Apr. 2018). DOI: 10.21105/joss.00638. URL: http://joss.theoj.org/papers/10.21105/joss.00638.

[320] Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. "Relief-based feature selection: Introduction and review". In: *Journal of Biomedical Informatics* 85 (2018), pp. 189–203.

[321] Ryan J Urbanowicz, Randal S Olson, Peter Schmitt, Melissa Meeker, and Jason H Moore. "Benchmarking relief-based feature selection methods for bioinformatics data mining". In: *Journal of Biomedical Informatics* 85 (2018), pp. 168–188.

[322] Chris Ding and Hanchuan Peng. "Minimum redundancy feature selection from microarray gene expression data". In: *Journal of Bioinformatics and Computational Biology* 3.02 (2005), pp. 185–205.

[323] Zhenyu Zhao, Radhika Anand, and Mallory Wang. "Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform". In: *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. 2019, pp. 442–452.

[324] Miron B Kursa and Witold R Rudnicki. "Feature selection with the Boruta package". In: *Journal of Statistical Software* 36 (2010), pp. 1–13.

[325] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. "Gene selection for cancer classification using support vector machines". In: *Machine Learning* 46.1 (2002), pp. 389–422.

[326] Xiaohui Lin et al. "Selecting feature subsets based on SVM-RFE and the overlapping ratio with applications in bioinformatics". In: *Molecules* 23.1 (2017), p. 52.

[327] Xuegong Zhang et al. "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data". In: *BMC Bioinformatics* 7.1 (2006), pp. 1–13.

[328] Hector Sanz, Clarissa Valim, Esteban Vegas, Josep M Oller, and Ferran Reverter. "SVM-RFE: selection and visualization of the most relevant features through non-linear kernels". In: *BMC Bioinformatics* 19.1 (2018), pp. 1–18.

[329] Heather Desaire. "How (Not) to Generate a Highly Predictive Biomarker Panel Using Machine Learning". In: *Journal of Proteome Research* (2022).

[330] Miseon Shim, Seung-Hwan Lee, and Han-Jeong Hwang. "Inflated prediction accuracy of neuropsychiatric biomarkers caused by data leakage in feature selection". In: *Scientific Reports* 11.1 (2021), pp. 1–7.

[331] Dana Bazazeh, Raed M Shubair, and Wasim Q Malik. "Biomarker discovery and validation for Parkinson's Disease: A machine learning approach". In: *2016 International Conference on Bio-engineering for Smart Technologies (BioSMART)*. IEEE. 2016, pp. 1–6.

[332] Carly A Bobak, Alexander J Titus, and Jane E Hill. "Comparison of common machine learning models for classification of tuberculosis using transcriptional biomarkers from integrated datasets". In: *Applied Soft Computing* 74 (2019), pp. 264–273.

[333] Sacha Beaumeunier et al. “The Role of Machine Learning in Finding Chimeric RNAs”. In: *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*. IEEE. 2015, pp. 26–30.

[334] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.", 2019.

[335] Ying Xie et al. “Early lung cancer diagnostic biomarker discovery by machine learning methods”. In: *Translational Oncology* 14.1 (2021), p. 100907.

[336] Janez Demšar. “Statistical comparisons of classifiers over multiple data sets”. In: *The Journal of Machine Learning Research* 7 (2006), pp. 1–30.

[337] Kyoham Shin, Jongmin Han, and Seokho Kang. “MI-MOTE: Multiple imputation-based minority oversampling technique for imbalanced and incomplete data classification”. In: *Information Sciences* 575 (2021), pp. 80–89.

[338] Roozbeh Razavi-Far, Maryam Farajzadeh-Zanajni, Boyu Wang, Mehrdad Saif, and Shiladitya Chakrabarti. “Imputation-based ensemble techniques for class imbalance learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 33.5 (2019), pp. 1988–2001.

[339] Anne-Laure Boulesteix. *Ten simple rules for reducing overoptimistic reporting in methodological computational research*. 2015.

[340] Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, and Sathvik Koneru. “Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks”. In: *Medical Imaging 2020: Computer-Aided Diagnosis*. Vol. 11314. SPIE. 2020, pp. 279–284.

[341] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J Casson. “Machine learning algorithm validation with a limited sample size”. In: *PlOS One* 14.11 (2019), e0224365.

[342] Simone A Thair et al. “A single nucleotide polymorphism in NF-κB inducing kinase is associated with mortality in septic shock”. In: *The Journal of Immunology* 186.4 (2011), pp. 2321–2328.

[343] Chamindie Punyadeera et al. “A biomarker panel to discriminate between systemic inflammatory response syndrome and sepsis and sepsis severity”. In: *Journal of Emergencies, Trauma and Shock* 3.1 (2010), p. 26.

[344] Kaushalya Amunugama, Daniel P Pike, and David A Ford. “The lipid biology of sepsis”. In: *Journal of Lipid Research* 62 (2021).

[345] Mathias Bruegel et al. “Sepsis-associated changes of the arachidonic acid metabolism and their diagnostic potential in septic patients”. In: *Critical Care Medicine* 40.5 (2012), pp. 1478–1486.

[346] Federica Sallusto, Danielle Lenig, Reinhold Förster, Martin Lipp, and Antonio Lanzavecchia. “Two subsets of memory T lymphocytes with distinct homing potentials and effector functions”. In: *Nature* 401.6754 (1999), pp. 708–712.

[347] Hassen Kared, Serena Martelli, Tze Pin Ng, Sylvia LF Pender, and Anis Larbi. “CD57 in human natural killer cells and T-lymphocytes”. In: *Cancer Immunology, Immunotherapy* 65.4 (2016), pp. 441–452.

[348] A Bryant, NC Calver, E Toubi, ADB Webster, and J Farrant. “Classification of patients with common variable immunodeficiency by B cell secretion of IgM and IgG in response to anti-IgM and interleukin-2”. In: *Clinical Immunology and Immunopathology* 56.2 (1990), pp. 239–248.

[349] Harald F Langer and Triantafyllos Chavakis. “Leukocyte–endothelial interactions in inflammation”. In: *Journal of Cellular and Molecular Medicine* 13.7 (2009), pp. 1211–1220.

[350] SM Lewis et al. “Expression of CD11c and EMR2 on neutrophils: potential diagnostic biomarkers for sepsis and systemic inflammation”. In: *Clinical & Experimental Immunology* 182.2 (2015), pp. 184–194.

[351] Adam G Laing et al. “A dynamic COVID-19 immune signature includes associations with poor prognosis”. In: *Nature Medicine* 26.10 (2020), pp. 1623–1635.

[352] Yuan Tian et al. “Single-cell immunology of SARS-CoV-2 infection”. In: *Nature Biotechnology* 40.1 (2022), pp. 30–41.

[353] Judea Pearl. “The seven tools of causal inference, with reflections on machine learning”. In: *Communications of the ACM* 62.3 (2019), pp. 54–60.

[354] Sofie Van Gassen, Brice Gaudilliere, Martin S Angst, Yvan Saeys, and Nima Aghaeepour. “CytoNorm: a normalization algorithm for cytometry data”. In: *Cytometry Part A* 97.3 (2020), pp. 268–278.

[355] Masato Ogishi et al. “Multibatch cytometry data integration for optimal immunophenotyping”. In: *The Journal of Immunology* 206.1 (2021), pp. 206–213.

[356] Quentin Blampey et al. “Interpretable cytometry cell-type annotation with flow-based deep generative models”. In: *arXiv preprint arXiv:2208.05745* (2022).

[357] Etienne Becht et al. “High-throughput single-cell quantification of hundreds of proteins using conventional flow cytometry and machine learning”. In: *Science Advances* 7.39 (Sept. 2021). DOI: 10.1126/sciadv.abg0505. URL: https://doi.org/10.1126/sciadv.abg0505.

[358] Peter Ghazal, Patricia R.S. Rodrigues, Mallinath Chakraborty, Siva Oruganti, and Thomas E. Woolley. “Challenging molecular dogmas in human sepsis using mathematical reasoning”. In: *eBioMedicine* 80 (June 2022), p. 104031. DOI: 10.1016/j.ebiom.2022.104031. URL: https://doi.org/10.1016/j.ebiom.2022.104031.

[359] Joe Alcock. “The emperor has no clothes? searching for dysregulation in sepsis”. In: *Journal of Clinical Medicine* 7.9 (2018), p. 247.

[360] Jennifer G. Wilson et al. “Cytokine profile in plasma of severe COVID-19 does not differ from ARDS and sepsis”. In: *JCI Insight* 5.17 (Sept. 2020). DOI: 10.1172/jci.insight.140289. URL: https://doi.org/10.1172/jci.insight.140289.

A | Appendix

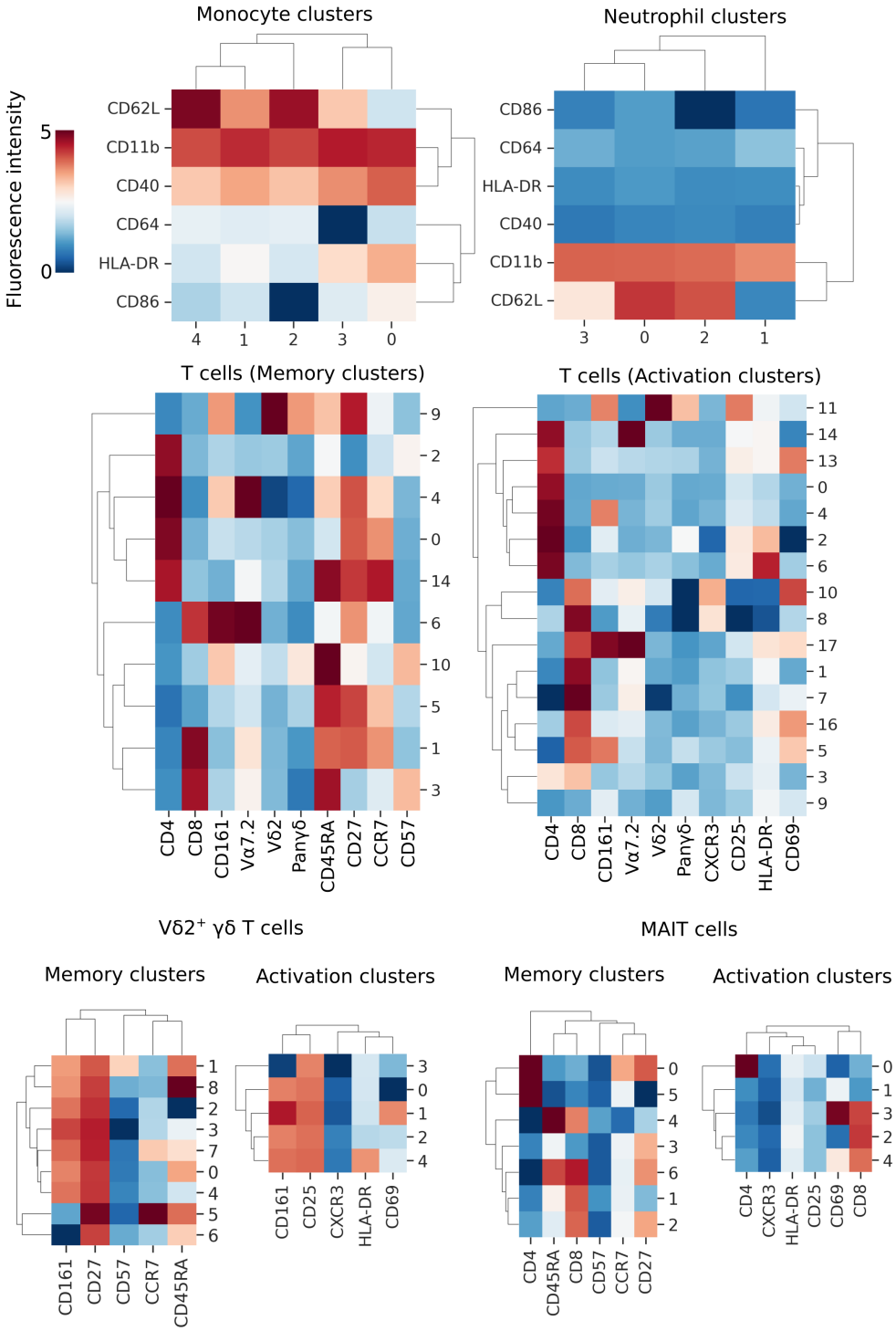

Figure A.1: Summary of clusters identified from flow cytometry analysis of whole blood samples taken from patients with acute severe sepsis as part of the ILTIS study, described in full within Chapter 5.

Table A.1: Description of routine clinical data available for patients diagnosed with sepsis and enrolled on the ILTIS study. The total number of patients with available data is shown (N) along with the average value and interquartile range (IQR).

Variable	Category	N	Average [IQR]
APTT	Coagulation screen	72	32.61 [27.39 - 33.89]
Alanine transaminase	Liver function test	71	112.5 [18.0 - 70.0]
Albumin	Bone profile	72	23.23 [17.0 - 28.0]
Alkaline phosphatase	Bone profile	72	119.03 [66.0 - 139.0]
Amylase	Amylase	20	88.58 [24.25 - 92.88]
Base excess	Blood Gas Venous	10	2.1 [1.02 - 2.82]
Base excess ecf	Blood Gas Arterial	38	2.36 [0.9 - 2.9]
Basophil count	Full blood count	69	0.06 [0.0 - 0.1]
Bilirubin	Liver function test	68	26.37 [9.0 - 29.5]
Bilirubin	Blood Gas Arterial	13	19.5 [2.33 - 22.0]
C-reactive protein (CRP)	C-reactive protein	71	215.27 [131.0 - 289.25]
Calcium	Bone profile	64	2.06 [1.93 - 2.17]
Calcium (adjusted)	Bone profile	68	2.31 [2.19 - 2.4]
Calcium (ionised)	Blood Gas Arterial	69	1.12 [1.08 - 1.19]
Carboxyhaemoglobin	Blood Gas Arterial	71	1.04 [0.77 - 1.2]
Chloride	Blood Gas Arterial	71	106.62 [103.0 - 109.83]
Clauss fibrinogen level	Coagulation screen	72	6.46 [4.41 - 7.84]
Creatine kinase	Creatine kinase	7	952.57 [60.5 - 482.0]
Creatinine	Estimated GFR	77	176.86 [67.5 - 192.25]
Eosinophil count	Full blood count	69	0.16 [0.1 - 0.2]
Estimated GFR	Estimated GFR	54	39.72 [21.88 - 55.5]
Free T4	Thyroid function test	6	12.88 [11.5 - 13.42]
Globulin	Bone profile	71	32.96 [28.0 - 37.5]
Glucose	Blood Gas Arterial	71	8.88 [6.52 - 9.75]
Haematocrit (Hct)	Full blood count	77	0.34 [0.28 - 0.38]
Haemoglobin (Hb)	Full blood count	77	113.64 [94.0 - 126.0]

Table A.1 continued from previous page

High Sensitivity Troponin I	hs-Troponin I time not stated	21	265.91 [13.25 - 164.62]
Inspired oxygen	Blood Gas Arterial	68	38.01 [24.31 - 45.76]
International normalised ratio	International normalised ratio	5	3.03 [1.6 - 5.0]
Lactate	Lactate	18	2.79 [1.23 - 3.71]
Lactate	Blood Gas Arterial	71	2.18 [1.18 - 2.82]
Lactate dehydrogenase	Lactate dehydrogenase	5	608.3 [405.0 - 797.5]
Lymphocyte count	Full blood count	69	1.0 [0.63 - 1.3]
Magnesium	Magnesium	68	0.83 [0.65 - 0.94]
Mean cell haemoglobin (MCH)	Full blood count	77	30.28 [28.57 - 32.05]
Mean cell volume (MCV)	Full blood count	77	90.75 [86.5 - 94.0]
Methaemoglobin	Blood Gas Arterial	71	1.15 [0.93 - 1.34]
Monocyte count	Full blood count	69	0.98 [0.65 - 1.2]
Neutrophil count	Full blood count	69	12.28 [8.0 - 15.9]
Nucleated red blood cell (NRBC) count	Full blood count	23	0.05 [0.0 - 0.1]
Phosphate	Bone profile	70	1.39 [0.93 - 1.66]
Platelet (PLT) count	Full blood count	72	234.31 [134.0 - 273.54]
Potassium	Urea and electrolytes	77	4.43 [4.0 - 4.7]
Potassium	Blood Gas Arterial	71	4.3 [3.88 - 4.4]
Previous CRP	Previous CRP	9	123.67 [2.0 - 199.0]
Protein	Bone profile	71	56.88 [49.08 - 64.0]
Prothrombin time (PT)	Coagulation screen	72	16.54 [12.92 - 17.8]
Red blood cell (RBC) count	Full blood count	72	3.72 [3.25 - 4.1]
Red cell distribution width (RDW)	Full blood count	77	14.71 [12.73 - 15.7]
SO2	Blood Gas Arterial	70	93.11 [91.74 - 96.88]

Table A.1 continued from previous page

Sodium	Urea and electrolytes	77	137.98 [134.0 - 140.5]
Sodium	Blood Gas Arterial	71	136.65 [133.03 - 139.5]
Standard bicarbonate (Arterial)	Blood Gas Arterial	69	22.31 [19.44 - 25.37]
TSH	Thyroid function test	6	0.8 [0.41 - 1.11]
Temperature	Blood Gas Arterial	69	37.22 [36.9 - 37.59]
Total Hb (calculated)	Blood Gas Arterial	29	88.43 [82.5 - 93.2]
Urea	Urea and electrolytes	77	12.58 [6.14 - 14.86]
White blood cell (WBC) count	Full blood count	72	15.16 [10.09 - 19.29]
pCO2	Blood Gas Arterial	69	5.53 [4.59 - 6.26]
pCO2	Blood Gas Venous	36	5.93 [5.07 - 6.35]
pH (Arterial)	Blood Gas Arterial	70	7.35 [7.29 - 7.42]
pO2 (Arterial)	Blood Gas Arterial	69	12.25 [10.4 - 13.39]
pO2	Blood Gas Venous	35	5.82 [4.5 - 6.69]

Table A.2: Description of all variables considered as potential features for machine learning models. Where a variable was later removed, a reason for exclusion is given.

Feature name	Category	Reason for exclusion
Age (Years)	Demographics	
Gender	Demographics	
Body Mass Index (BMI)	Physiology	
APACHE-II Score	Severity score	
Days ventilated	Interventions	
Renal Replacement Therapy (RTT)	Interventions	
Emergency/Trauma	Medical history	
Actual bicarbonate (aHCO3)	Blood gas analysis (Arterial)	
Actual bicarbonate (aHCO3)	Blood gas analysis	Only available for one patient
ASAMTS13 protease assay	Clinical laboratory measurement - ADAMST	Only available for one patient
Alanine Transaminase (ALT)	Clinical laboratory measurement - Liver function tests	Imputation OOB error >= 1.0
Albumin	Clinical laboratory measurement - Liver function tests	
Alkaline Phosphatase	Clinical laboratory measurement - Liver function tests	
Amikacin	Clinical laboratory measurement - Amikacin	NMAR and only available for one patient
Ammonia	Clinical laboratory measurement - Ammonia	Only available for one patient
Amylase	Clinical laboratory measurement - Amylase	More than 40% of patients with missing data

Table A.2 continued from previous page

Activated partial thromboplastin clotting time	Clinical laboratory measurement - Coagulation screen	
Aspartate Transaminase (AST)	Clinical laboratory measurement - Aspartate Transaminase	Only available for two patients
Base excess of extracellular fluid	Blood gas analysis (Arterial)	More than 40% of patients with missing data
Base excess of extracellular fluid	Blood gas analysis (Venous)	Imputation OOB error >= 1.0
Base excess	Blood gas analysis (Venous)	Imputation OOB error >= 1.0
Basophil count	Clinical laboratory measurement - Full blood count	
Bicarbonate	Clinical laboratory measurement - Bicarbonate	
Bilirubin	Blood gas analysis (Arterial)	Imputation OOB error >= 1.0
Bilirubin	Blood gas analysis (Venous)	Only available in seven patients and overlaps with liver function tests
Bilirubin	Clinical laboratory measurement - Liver function tests	
C-Reactive Protein	Clinical laboratory measurement - C-Reactive Protein	
Albumin adjusted calcium	Clinical laboratory measurement - Calcium	

Table A.2 continued from previous page

Albumin adjusted calcium	Clinical laboratory measurement - Liver function tests	Merged with calcium request
Ionized Calcium	Blood gas analysis (Arterial)	
Ionized Calcium	Blood gas analysis (Venous)	More than 40% of patients with missing data
Calcium	Clinical laboratory measurement - Calcium	
Carboxyhaemoglobin	Blood gas analysis (Arterial)	
Carboxyhaemoglobin	Blood gas analysis (Venous)	More than 40% of patients with missing data
Chloride	Blood gas analysis (Arterial)	
Chloride	Blood gas analysis (Venous)	More than 40% of patients with missing data
Chloride	Clinical laboratory measurement - Chloride	Only available for one patient and overlaps with blood gas analysis
Cholesterol	Clinical laboratory measurement - Cholesterol	
Clauss Fibrinogen Level	Clinical laboratory measurement - Coagulation screen	
Cortisol	Clinical laboratory measurement - Cortisol	NMAR and only available for one patient
Creatine Kinase	Clinical laboratory measurement - Creatine Kinase	Only available for seven patients

Table A.2 continued from previous page

Creatinine ratio (Urine)	Clinical laboratory measurement - Protein	Only available for one patient
Creatinine (Urine)	Clinical laboratory measurement - Protein	Only available for one patient
Creatinine	Clinical laboratory measurement - Electrolyte profile	Merged with urea and electrolytes request
Creatinine	Clinical laboratory measurement - Estimated GFR	Merged with urea and electrolytes request
Creatinine	Clinical laboratory measurement - Urea and Electrolytes	
Digoxin	Clinical laboratory measurement - Digoxin	NMAR and only available for one patient
Eosinophil count	Clinical laboratory measurement - Full blood count	
Estimated GFR	Clinical laboratory measurement - Electrolyte profile	Merged with Estigated GFR request
Estimated GFR	Clinical laboratory measurement - Estimated GFR	
Factor Vii Level	Clinical laboratory measurement - Factor Vii	Only available in two patients
Ferritin	Clinical laboratory measurement - Ferritin	Only available in one patient
Folate	Clinical laboratory measurement - Folate	Only available in one patient
Free T4	Clinical laboratory measurement - Thyroid function test	Only available in two patients

Table A.2 continued from previous page

G-Glutamyl Transferase	Clinical laboratory measurement - G-Glutamyl Transferase	Only available in one patient
Gentamicin	Clinical laboratory measurement - Gentamicin	NMAR and only available for five patients
Globulin	Clinical laboratory measurement - Globulin	
Glucose (Random)	Clinical laboratory measurement - Glucose (Random)	Only available in four patients
Glucose	Blood gas analysis (Arterial)	
Glucose	Blood gas analysis (Venous)	More than 40% of patients with missing data
Haematocrit (Hct)	Clinical laboratory measurement - Full blood count	
Haemoglobin (Hb)	Clinical laboratory measurement - Full blood count	
Haemoglobin (Hb)	Blood gas analysis	Only available for one patient
HDL Cholesterol	Clinical laboratory measurement - Lipid profile	Only available for one patient
HDL Ratio	Clinical laboratory measurement - Lipid profile	Only available for one patient
High Sensitivity Troponin I	Clinical laboratory measurement - High Sensitivity Troponin I	Imputation OOB error >= 1.0
Inspired Oxygen	Blood gas analysis (Arterial)	

Table A.2 continued from previous page

Inspired Oxygen	Blood gas analysis (Venous)	More than 40% of patients with missing data
Lactate Dehydrogenase	Clinical laboratory measurement - Lactate Dehydrogenase	Only available for five patients
Lactate	Clinical laboratory measurement - Lactate	
Lactate	Blood gas analysis (Arterial)	
Lactate	Blood gas analysis (Venous)	More than 40% of patients with missing data
LDL Cholesterol	Clinical laboratory measurement - Lipid profile	Only available for one patient
Lymphocyte count	Clinical laboratory measurement - Full blood count	
Magnesium	Clinical laboratory measurement - Liver function tests	Merged with magnesium request
Magnesium	Clinical laboratory measurement - Magnesium	
Mean cell Haemoglobin	Clinical laboratory measurement - Full blood count	
Mean cell volume	Clinical laboratory measurement - Full blood count	
Methaemoglobin	Blood gas analysis	Only available for one patient and overlaps with arterial blood gas analysis

Table A.2 continued from previous page

Methaemoglobin	Blood gas analysis (Arterial)	
Methaemoglobin	Blood gas analysis (Venous)	
Monocyte count	Clinical laboratory measurement - Full blood count	
Neutrophil count	Clinical laboratory measurement - Full blood count	
Non-HDL Cholesterol	Clinical laboratory measurement - Lipid profile	Only available for one patient
Nucleated red blood cell count	Clinical laboratory measurement - Full blood count	
Oxyhaemoglobin	Blood gas analysis (Arterial)	Only available for two patients
pCO2	Blood gas analysis (Arterial)	
pCO2	Blood gas analysis	Only available for one patient and overlaps with arterial blood gas analysis
pCO2	Blood gas analysis (Venous)	
pH	Blood gas analysis (Arterial)	
pH	Blood gas analysis	Only available for one patient and overlaps with arterial blood gas analysis

Table A.2 continued from previous page

Phosphate	Clinical laboratory measurement - Liver function tests	Merged with phosphate request
Phosphate	Clinical laboratory measurement - Phosphate	
Platelet count	Clinical laboratory measurement - Full blood count	
pO2	Blood gas analysis (Arterial)	
pO2	Blood gas analysis	Only available for one patient and overlaps with arterial blood gas analysis
pO2	Blood gas analysis (Venous)	Imputation OOB error >= 1.0
Potassium	Blood gas analysis (Arterial)	
Potassium	Blood gas analysis (Venous)	
Potassium	Clinical laboratory measurement - Electrolyte profile	Imputation OOB error >= 1.0
Potassium	Clinical laboratory measurement - Urea and Electrolytes	
Procalcitonin	Clinical laboratory measurement - Procalcitonin	Only available for one patient
Protein (Urine)	Clinical laboratory measurement - Protein	Only available for one patient

Table A.2 continued from previous page

Protein	Clinical laboratory measurement - Liver function tests	
Prothrombin time	Clinical laboratory measurement - Coagulation screen	
Red blood cell count	Clinical laboratory measurement - Full blood count	
Red cell distribution width	Clinical laboratory measurement - Full blood count	
Reptilase clotting time	Clinical laboratory measurement - Reptilase clotting time	Only available for one patient
Reticulocytes	Clinical laboratory measurement - Reticulocyte count	Only available for one patient
SO2	Blood gas analysis (Arterial)	
SO2	Blood gas analysis	Only available for one patient and overlaps with arterial blood gas analysis
Sodium	Blood gas analysis (Arterial)	
Sodium	Blood gas analysis (Venous)	
Sodium	Clinical laboratory measurement - Electrolyte profile	Imputation OOB error >= 1.0

Table A.2 continued from previous page

Sodium	Clinical laboratory measurement - Urea and Electrolytes	
Standard bicarbonate	Blood gas analysis (Arterial)	
Standard bicarbonate	Blood gas analysis	
Temperature	Blood gas analysis (Arterial)	
Temperature	Blood gas analysis (Venous)	
Thrombin time	Clinical laboratory measurement - Thrombin time	Only available for two patients
Total CO2	Blood gas analysis	Only available for one patient
Total Hb calculated	Blood gas analysis (Arterial)	
Total Hb calculated	Blood gas analysis (Venous)	Imputation OOB error >= 1.0
Triglyceride	Clinical laboratory measurement - Lipid profile	Only available for one patient
Triglyceride	Clinical laboratory measurement - Triglyceride	Only available for one patient
TSH (Thyroid-stimulating hormone)	Clinical laboratory measurement - Thyroid function test	Only available for six patients
Urate	Clinical laboratory measurement - Urate	Only available for three patients
Urea	Clinical laboratory measurement - Electrolyte profile	Merged with urea and electrolytes request

Table A.2 continued from previous page

Urea	Clinical laboratory measurement - Liver function tests	Merged with urea and electrolytes request
Urea	Clinical laboratory measurement - Urea	Merged with urea and electrolytes request
Urea	Clinical laboratory measurement - Urea and Electrolytes	
Urine volume	Clinical laboratory measurement - Creatinine clearance	Only available for one patient
Vancomycin	Clinical laboratory measurement - Vancomycin	NMAR and only available for two patients
Vitamin B12	Clinical laboratory measurement - Vitamin B	Only available for one patient
White blood cell count	Clinical laboratory measurement - Full blood count	
CCL-5 plasma concentration	Luminex	
CXCL10 plasma concentration	Luminex	
IL-4 plasma concentration	Luminex	
Lactoferrin plasma concentration	Luminex	
MMP-8 plasma concentration	Luminex	
MMP-9 plasma concentration	Luminex	
PD-L1 plasma concentration	Luminex	
VEGF plasma concentration	Luminex	
IL-6 plasma concentration	ELISA	
CCL2 plasma concentration	Luminex	

Table A.2 continued from previous page

CXCL13 plasma concentration	Luminex
FLT3-L plasma concentration	Luminex
G-CSF plasma concentration	Luminex
IL-10 plasma concentration	Luminex
IL-15 plasma concentration	Luminex
IL-1α plasma concentration	Luminex
CXCL8 plasma concentration	Luminex
OSM plasma concentration	Luminex
Procalcitonin plasma concentration	Luminex
Ferritin plasma concentration	Luminex
IFNγ plasma concentration	ELISA
TNFα plasma concentration	ELISA
T cells (% of PBMCs)	Flow cytometry - Major subsets
Monocytes (% of Leukocytes)	Flow cytometry - Major subsets
Neutrophils (% of Leukocytes)	Flow cytometry - Major subsets
$CD4^+$ $CD8^-$ T cells (% of T cells)	Flow cytometry - T cell subsets
$CD4^-$ $CD8^+$ T cells (% of T cells)	Flow cytometry - T cell subsets
V$\delta 2^+$ $\gamma\delta$ T cells (% of T cells)	Flow cytometry - T cell subsets
MAIT cells (% of T cells)	Flow cytometry - T cell subsets
T cells memory cluster 1 (% of T cells)	Flow cytometry - Memory T cell clusters

Table A.2 continued from previous page

T cells memory cluster 3 (% of T cells)	Flow cytometry - Memory T cell clusters	
T cells memory cluster 14 (% of T cells)	Flow cytometry - Memory T cell clusters	
T cells memory cluster 2 (% of T cells)	Flow cytometry - Memory T cell clusters	
T cells memory cluster 4 (% of T cells)	Flow cytometry - Memory T cell clusters	Imputation OOB error >= 1.0
T cells memory cluster 0 (% of T cells)	Flow cytometry - Memory T cell clusters	
T cells activation cluster 1 (% of T cells)	Flow cytometry - Activated T cell clusters	
T cells activation cluster 12 (% of T cells)	Flow cytometry - Activated T cell clusters	
T cells activation cluster 5 (% of T cells)	Flow cytometry - Activated T cell clusters	
T cells activation cluster 14 (% of T cells)	Flow cytometry - Activated T cell clusters	
T cells activation cluster 16 (% of T cells)	Flow cytometry - Activated T cell clusters	
T cells activation cluster 2 (% of T cells)	Flow cytometry - Activated T cell clusters	
T cells activation cluster 7 (% of T cells)	Flow cytometry - Activated T cell clusters	
T cells activation cluster 6 (% of T cells)	Flow cytometry - Activated T cell clusters	
T cells activation cluster 4 (% of T cells)	Flow cytometry - Activated T cell clusters	
T cells activation cluster 10 (% of T cells)	Flow cytometry - Activated T cell clusters	

Table A.2 continued from previous page

T cells activation cluster 0 (% of T cells)	Flow cytometry - Activated T cell clusters	
T cells activation cluster 13 (% of T cells)	Flow cytometry - Activated T cell clusters	
T cells activation cluster 8 (% of T cells)	Flow cytometry - Activated T cell clusters	
V$\delta 2^+$ $\gamma\delta$ T cells memory cluster 3 (% of V$\delta 2^+$ $\gamma\delta$ T cells)	Flow cytometry - Memory V$\delta 2^+$ $\gamma\delta$ T cell clusters	Imputation OOB error >= 1.0
V$\delta 2^+$ $\gamma\delta$ T cells memory cluster 0 (% of V$\delta 2^+$ $\gamma\delta$ T cells)	Flow cytometry - Memory V$\delta 2^+$ $\gamma\delta$ T cell clusters	
V$\delta 2^+$ $\gamma\delta$ T cells memory cluster 1 (% of V$\delta 2^+$ $\gamma\delta$ T cells)	Flow cytometry - Memory V$\delta 2^+$ $\gamma\delta$ T cell clusters	
V$\delta 2^+$ $\gamma\delta$ T cells memory cluster 4 (% of V$\delta 2^+$ $\gamma\delta$ T cells)	Flow cytometry - Memory V$\delta 2^+$ $\gamma\delta$ T cell clusters	
V$\delta 2^+$ $\gamma\delta$ T cells memory cluster 2 (% of V$\delta 2^+$ $\gamma\delta$ T cells)	Flow cytometry - Memory V$\delta 2^+$ $\gamma\delta$ T cell clusters	
V$\delta 2^+$ $\gamma\delta$ T cells activation cluster 1 (% of V$\delta 2^+$ $\gamma\delta$ T cells)	Flow cytometry - Activated V$\delta 2^+$ $\gamma\delta$ T cell clusters	
V$\delta 2^+$ $\gamma\delta$ T cells activation cluster 3 (% of V$\delta 2^+$ $\gamma\delta$ T cells)	Flow cytometry - Activated V$\delta 2^+$ $\gamma\delta$ T cell clusters	
V$\delta 2^+$ $\gamma\delta$ T cells activation cluster 2 (% of V$\delta 2^+$ $\gamma\delta$ T cells)	Flow cytometry - Activated V$\delta 2^+$ $\gamma\delta$ T cell clusters	

Table A.2 continued from previous page

Vδ2$^+$ $\gamma\delta$ T cells activation cluster 4 (% of Vδ2$^+$ $\gamma\delta$ T cells)	Flow cytometry - Activated Vδ2$^+$ $\gamma\delta$ T cell clusters
Vδ2$^+$ $\gamma\delta$ T cells activation cluster 0 (% of Vδ2$^+$ $\gamma\delta$ T cells)	Flow cytometry - Activated Vδ2$^+$ $\gamma\delta$ T cell clusters
MAIT cells memory cluster 1 (% of MAIT cells)	Flow cytometry - Memory MAIT cell clusters
MAIT cells memory cluster 3 (% of MAIT cells)	Flow cytometry - Memory MAIT cell clusters
MAIT cells memory cluster 5 (% of MAIT cells)	Flow cytometry - Memory MAIT cell clusters
MAIT cells memory cluster 2 (% of MAIT cells)	Flow cytometry - Memory MAIT cell clusters
MAIT cells memory cluster 6 (% of MAIT cells)	Flow cytometry - Memory MAIT cell clusters
MAIT cells memory cluster 4 (% of MAIT cells)	Flow cytometry - Memory MAIT cell clusters
MAIT cells memory cluster 0 (% of MAIT cells)	Flow cytometry - Memory MAIT cell clusters
MAIT cells activation cluster 1 (% of MAIT cells)	Flow cytometry - Activated MAIT cell clusters
MAIT cells activation cluster 3 (% of MAIT cells)	Flow cytometry - Activated MAIT cell clusters
MAIT cells activation cluster 2 (% of MAIT cells)	Flow cytometry - Activated MAIT cell clusters
MAIT cells activation cluster 4 (% of MAIT cells)	Flow cytometry - Activated MAIT cell clusters
MAIT cells activation cluster 0 (% of MAIT cells)	Flow cytometry - Activated MAIT cell clusters

Table A.2 continued from previous page

Monocyte cluster 1 (% of Monocytes)	Flow cytometry - Monocyte clusters
Monocyte cluster 3 (% of Monocytes)	Flow cytometry - Monocyte clusters
Monocyte cluster 2 (% of Monocytes)	Flow cytometry - Monocyte clusters
Monocyte cluster 4 (% of Monocytes)	Flow cytometry - Monocyte clusters
Monocyte cluster 0 (% of Monocytes)	Flow cytometry - Monocyte clusters
Neutrophil cluster 2 (% of Neutrophils)	Flow cytometry - Neutrophil clusters
Neutrophil cluster 3 (% of Neutrophils)	Flow cytometry - Neutrophil clusters
Neutrophil cluster 0 (% of Neutrophils)	Flow cytometry - Neutrophil clusters
Neutrophil cluster 1 (% of Neutrophils)	Flow cytometry - Neutrophil clusters
Monocytes HLA-DR MFI	Flow cytometry - Monocyte activation marker MFI
Monocytes CD86 MFI	Flow cytometry - Monocyte activation marker MFI
Monocytes CD40 MFI	Flow cytometry - Monocyte activation marker MFI
Monocytes CD64 MFI	Flow cytometry - Monocyte activation marker MFI

Table A.2 continued from previous page

Monocytes CD62L MFI	Flow cytometry - Monocyte activation marker MFI
Neutrophils HLA-DR MFI	Flow cytometry - Neutrophil activation marker MFI
Neutrophils CD86 MFI	Flow cytometry - Neutrophil activation marker MFI
Neutrophils CD40 MFI	Flow cytometry - Neutrophil activation marker MFI
Neutrophils CD64 MFI	Flow cytometry - Neutrophil activation marker MFI
Neutrophils CD62L MFI	Flow cytometry - Neutrophil activation marker MFI
$CD8^+$ T cells CXCR3 MFI	Flow cytometry - $CD8^+$ T cell activation marker MFI
$CD8^+$ T cells CD161 MFI	Flow cytometry - $CD8^+$ T cell activation marker MFI
$CD8^+$ T cells HLA-DR MFI	Flow cytometry - $CD8^+$ T cell activation marker MFI
$CD8^+$ T cells CD69 MFI	Flow cytometry - $CD8^+$ T cell activation marker MFI
$CD8^+$ T cells CD25 MFI	Flow cytometry - $CD8^+$ T cell activation marker MFI
$CD4^+$ T cells CXCR3 MFI	Flow cytometry - $CD4^+$ T cell activation marker MFI

Table A.2 continued from previous page

$CD4^+$ T cells CD161 MFI	Flow cytometry - $CD4^+$ T cell activation marker MFI
$CD4^+$ T cells HLA-DR MFI	Flow cytometry - $CD4^+$ T cell activation marker MFI
$CD4^+$ T cells CD69 MFI	Flow cytometry - $CD4^+$ T cell activation marker MFI
$CD4^+$ T cells CD25 MFI	Flow cytometry - $CD4^+$ T cell activation marker MFI
MAIT cells CXCR3 MFI	Flow cytometry - MAIT cell activation marker MFI
MAIT cells HLA-DR MFI	Flow cytometry - MAIT cell activation marker MFI
MAIT cells CD69 MFI	Flow cytometry - MAIT cell activation marker MFI
MAIT cells CD25 MFI	Flow cytometry - MAIT cell activation marker MFI
V$\delta 2^+$ $\gamma\delta$ T cells CXCR3 MFI	Flow cytometry - V$\delta 2^+$ $\gamma\delta$ T cell activation marker MFI
V$\delta 2^+$ $\gamma\delta$ T cells CD161 MFI	Flow cytometry - V$\delta 2^+$ $\gamma\delta$ T cell activation marker MFI
V$\delta 2^+$ $\gamma\delta$ T cells HLA-DR MFI	Flow cytometry - V$\delta 2^+$ $\gamma\delta$ T cell activation marker MFI
V$\delta 2^+$ $\gamma\delta$ T cells CD69 MFI	Flow cytometry - V$\delta 2^+$ $\gamma\delta$ T cell activation marker MFI
V$\delta 2^+$ $\gamma\delta$ T cells CD25 MFI	Flow cytometry - V$\delta 2^+$ $\gamma\delta$ T cell activation marker MFI

Table A.2 continued from previous page

C4 carnitine plasma concentration	Lipids
C6 carnitine plasma concentration	Lipids
C8 carnitine plasma concentration	Lipids
C10 carnitine plasma concentration	Lipids
C12 carnitine plasma concentration	Lipids
C2 carnitine plasma concentration	Lipids
C14 carnitine plasma concentration	Lipids
C16 carnitine plasma concentration	Lipids
C18 carnitine plasma concentration	Lipids
C3 carnitine plasma concentration	Lipids
C18:1 carnitine plasma concentration	Lipids
C12-2OH/3OH plasma concentration	Lipids
C22:6 plasma concentration	Lipids
C18:2 plasma concentration	Lipids
C18:3 plasma concentration	Lipids
C20:5 plasma concentration	Lipids
C18:1 plasma concentration	Lipids
C8:0 plasma concentration	Lipids

Table A.2 continued from previous page

C10:0 plasma concentration	Lipids
C12:0 plasma concentration	Lipids
C20:4 plasma concentration	Lipids
C14:0 plasma concentration	Lipids
C16:0 plasma concentration	Lipids
C18:0 plasma concentration	Lipids