UNIVERSITY OF BIRMINGHAM

Applications of data science to investigate risk factors in allogeneic stem cell transplantation

By

David Rhys Greenwood

A thesis submitted to the

University of Birmingham for the degree of

Doctor of Philosophy

Institute of Immunology & Immunotherapy

College of Medical and Dental Sciences

University of Birmingham

August 2019

Abstract

Allogeneic haematopoietic stem cell transplantation (allo-HSCT) from healthy donors can provide curative therapy for a variety of life-threatening haematological malignancies. However, the risk of relapse, graft-vs-host disease (GvHD), infection, and toxicity remain major causes of mortality.

Mature immune cell populations are also infused as a component of the haematopoietic stem cell (HSC) graft. We aimed to investigate how the cellular composition of grafts varied between donors and to study associations of graft composition with recipient prognosis. Moreover, we intended to examine proteins measurable in the serum post-transplant as potential prognostic biomarkers.

Cell populations were identified in graft samples through the application of unsupervised clustering to single-cell mass cytometry (CyTOF) data. Cluster analysis for compositional data was used to define distinct types of graft based on cellular composition. By applying supervised variable selection algorithms to proteomics data, prognostic biomarkers were identified in serum samples taken on the 14^{th} day post-transplant. Lastly, survival and competing risks regression were used to test if graft composition and potential biomarkers were associated with prognosis following allo-HSCT.

The cellular composition of stem cell grafts was shown to be highly variable between donors with HSC representing a small fraction of cellular content. We found evidence that differences in graft composition were associated with significant changes in prognosis beyond that attributable to other established risk factors. Based on cellular composition, we were able to define three distinct types of grafts that were linked with differences in overall survival (OS), GvHD-free relapse-free (GRFS) survival, and transplant-related mortality (TRM). Concerning the contribution of individual cell types, a greater proportion of T cells (relative to HSCs) was predictive of an increased incidence of chronic GvHD. Furthermore, monocytes and a population of DNAM-1^{+} lymphocyte progenitors had a protective effect on the rate of GRFS and incidence of TRM, respectively.

Several serum proteins were identified as potential prognostic biomarkers when measured on the 14^{th} day following transplantation. High concentrations of PD-L1, CCL23, and CDCP1 expression were predictive for lower rates of OS, GRFS, and increased incidence of relapse-related morality, respectively. Moreover, IL-6 concentration was associated with a lower incidence of relapse-related mortality but an increase in deaths from other causes.

These findings reveal that graft composition and proteins expressed in the serum after allo-HSCT are important predictors of clinical prognosis. This work demonstrates the benefit of applied data science for biomarker discovery and highlights therapeutic opportunities for graft modification in the treatment of haematological malignancies with allo-HSCT.

During the development of this thesis, I spent a period of time on secondment to COVID-19 research. A paper published on this topic is integrated as Appendix A.

Acknowledgements

The work in this thesis would not have been possible without the support and expertise of my family, friends, and colleagues.

Thank you to my wonderful parents for encouraging me to follow my ambitions. Your unwavering patience, kindness, and cash made this achievable, and I'll forever be grateful. I'd like to extend my gratitude to my supervisors Professor Paul Moss and Dr Wayne Croft who gave me this opportunity. Thank you for your continued friendship, trust, and support throughout this process. We got there in the end!

To all of my friends and colleagues within the Moss group, thank you for making the past four years so enjoyable. In particular, I'd like to thank Dr Francesca Kinsella and Dr Kriti Verma for letting me contribute to such fascinating projects. Moreover, thank you to Dr. Duncan Murray, Dr Jianmin Zuo, Dr Hayden Pearce, Dr Rachel Bruton for all of your help and support.

I'd like to thank my friend Sims whose cut-rate rent and words of encouragement sent me down this path, cheers for everything. Thanks to Carys and Nick for all of your support over the years. A special thanks to Meadhbh for being so patient and supportive, you made a challenging time far more enjoyable. I'd like to thank Nathan, Will, and Sach for turning getting a PhD into a competitive sport. Thank you to Conor for being a sounding board and a constant source of good tunes and articles. I'd also like to thank my friends Ant, Liz, Bill, Luke, Sam, Bisk, Sally and Tom.

A special thank you to everyone in the Centre for Computational Biology for giving me a place to discuss data science in extreme levels of detail. In particular to Professor Jean-Baptiste Cazier, Professor Christopher Yau, and Dr Anas Rana for sparking my interest in bioinformatics and showing me what can be achieved with a laptop and interesting data! As well as Jessica Mylchreest and Jordan McCormick for going above and beyond throughout the years. I'd also like to thank Professor Krish Nirantharakumar, Dr. Nicki Adderley, and Dr. Tom Taverner for the opportunities to collaborate and learn.

List of publications

Greenwood, D., Taverner, T., Adderley, N. J., Price, M. J., Gokhale, K., Sainsbury, C., Gallier, S., Welch, C., Sapey, E., Murray, D., Fanning, H., Ball, S., Nirantharakumar, K., Croft, W., & Moss, P. (2022). **Machine learning of COVID-19 clinical data identifies population structures with therapeutic potential[1]. *iScience***, *25*(7), 104480. https://doi.org/10.1016/j.isci.2022.104480

Adderley, N. J., Taverner, T., Price, M. J., Sainsbury, C., **Greenwood, D.**, Chandan, J. S., Takwoingi, Y., Haniffa, R., Hosier, I., Welch, C., Parekh, D., Gallier, S., Gokhale, K., Denniston, A. K., Sapey, E., & Nirantharakumar, K. (2022). **Development and external validation of prognostic models for COVID-19 to support risk stratification in secondary care. *BMJ open***, *12*(1), e049506. https://doi.org/10.1136/bmjopen-2021-049506

Parry, H., Bruton, R., Tut, G., Ali, M., Stephens, C., **Greenwood, D**., Faustini, S., Hughes, S., Huissoon, A., Meade, R., Brown, K., Amirthalingam, G., Otter, A., Hallis, B., Richter, A., Zuo, J., & Moss, P. (2021**). Immunogenicity of single vaccination with BNT162b2 or ChAdOx1 nCoV-19 at 5-6 weeks post vaccine in participants aged 80 years or older: an exploratory analysis. *The Lancet***. *Healthy longevity*, *2*(9), e554–e560. https://doi.org/10.1016/S2666-7568(21)00169-0

[1]Published paper integrated into this thesis as **Appendix A**.

Table of contents

Table of Contents

List of figures

List of tables

List of abbreviations

Abbreviation	Definition
adj. HR	Adjusted hazard ratio
adj. HR_{SD}	Adjusted subdistribution hazard ratio
AGvHD	Acute graft-versus-host disease
AIC	Akaike information criterion
AIC selection	Stepwise selection with the Akaike information criterion
allo-HSCT	Allogeneic haematopoietic stem-cell transplantation
ALR	Additive log-ratio transformation
APCs	Antigen-presenting cells
ATG	Anti-thymocyte globulin
AUC	Area under the ROC Curve
B cells	B lymphocytes
BH	Benjamini-Hochberg
BIC	Bayesian information criterion
BM	Bone marrow
bw	Backward
cDCs	Classical DCs
CGvHD	Chronic graft-versus-host disease
CI	Confidence interval
CL	Cell lineage
CLP	Common lymphoid progenitors
CLR	Centered log-ratio transformation
clustCoDa	Cluster analysis for compositional data
CM	Central memory
CMP	Common myeloid progenitors
CMV	Cytomegalovirus
Cox PH	Cox proportional-hazards model (Cox regression)
CS	Cell subset
CSH	Cause-specific hazard model
CXCL10	C-X-C motif chemokine ligand 10
CyA	Cyclosporine
CyTOF	Single cell mass cytometry (cytometry by time of flight)
DAMPs	Damage-associated molecular patterns
DCs	Dendritic cells
DN	Double negative
DNA	Deoxyribonucleic acid
DP	Double positive
DR	Dimension reduction
EASIX	Endothelial Activation and Stress Index
EBMT	The European Group for Blood & Marrow Transplantation
EBV	Epstein-Barr virus

EM	Effector memory
EMRA	Effector memory cells re-expressing CD45RA
EPV	Events per variable ratio
FG	Fine-Gray subdistribution hazard model
fw	Forward
G-CSF	Granulocyte colony stimulating factor
GMM	Gaussian mixture model
GMM clustering	Model-based clustering with the Gaussian mixture model
GRFS	Graft-versus-host disease-free, relapse-free survival
GvHD	Graft-versus-host disease
GvL	Graft-versus-leukemia effect
haploHSCT	Haploidentical haematopoietic stem cell transplantation
HCT-CI	Haematopoietic cell transplantation-specific comorbidity index
HGF	Hepatocyte growth factor
HLA	Human leukocyte antigens
HR	Hazard ratio
HR_{SD}	Subdistribution hazard ratio
HSC	Haematopoietic stem cell
IL15	Interleukin-15
IL6	Interleukin-6
IL8	Interleukin-8
ILR	Isometric log-ratio transformation
int.	Intermediate
IQR	Interquartile range
KM	Kaplan–Meier estimator
LASSO	Least absolute shrinkage and selection operator
Lin-	Lineage negative
Lin+	Lineage positive
Lymph. progenitor	DNAM-1^{+} lymphocyte progenitors
MAC	Myeloablative conditioning
MEM	Marker enrichment modelling
MICE	Multivariate Imputation by Chained Equations
MiRKAT-S	Microbiome regression-based kernel association test
MMF	Mycophenolate mofetil
MMI	Mean marker intensity
MNC	Mononuclear cell
MRD	Measurable/minimal residual disease
MSD	Human leukocyte antigen matched sibling donor
MTX	Methotrexate
MUD	Matched unrelated donors
N.	Number
NK cell	Natural Killer cells
NMAC	Nonmyeloablative conditioning regimen
NPX	Normalized protein expression

NRM	Non-relapse mortality
OMiSA	Optimal microbiome-based survival analysis
OS	Overall survival
PAMPs	Pathogen-associated molecular patterns
PBSC	Peripheral blood stem cell
PBSC graft groups	Peripheral blood stem cell graft compositional clusters
PCA	Principal component analysis
PD-1	Programmed cell death protein 1
pDC	Plasmocytoid DCs
PH	Proportional hazards
PHATE	Potential of heat diffusion for affinity-based transition embedding
PRR	Pattern recognition receptor
PTLD	Post-transplant lymphoproliferative disorders
REG3α	Regenerating islet-derived protein 3-alpha
RIC	Reduced intensity conditioning
RM	Relapse-related mortality
rPAM	Revaluated pre-transplantation assessment of mortality
RSF	Random survival forest
ST2	Suppression of Tumorigenicity 2
T cells	T lymphocytes
T regs	Regulatory T cells
TBI	Total body irradiation
TCR	T cell receptor
Tim-3	T cell immunoglobulin and mucin domain 3
TNFR1	Tumor necrosis factor receptor 1
TRM	Transplant-related mortality
UCB	Umbilical cord blood
UMAP	Uniform manifold approximation and projection
VEGF	Vascular endothelial growth factors
α/β	Alpha/beta
γδ	Gamma delta

1 Introduction

1.1 Overview of the immune system

The immune system is a complex network of organs, cells, molecules, and chemical reactions. This network functions to protect the human body from a diverse range of invaders including harmful substances, microbes, viruses, and cancers.

This protection relies on two activities – recognition and response (1). The immune system must first recognise that a molecule is foreign and not the body's own cell or molecule (self-nonself discrimination), and it must then be able to elicit an appropriate response to neutralise or destroy the invader. Additionally, it needs to recognise changes in the body's cells which could result in cancer.

The immune system in humans consists of two subsystems, innate and adaptive immunity. These complementary mechanisms work in concert - both must function for an effective immune response (1). Innate immunity provides rapid sensing and neutralising of invading molecules and harmful substances. Whereas the adaptive immune system adapts over time to recognise, neutralise and remember specific pathogens.

1.1.1 Innate

Antigens are molecules capable of stimulating an immune response or that are recognised by the products of an immune response (2). Innate immunity is rapid and non-specific, responding to a wide range of different antigens. It also lacks immunologic memory, meaning it cannot recognise and adapt to a previously encountered pathogen.

Innate defences include anatomic (skin, and mucous membranes) and physiologic barriers (temperature and pH). In addition to effector responses by immune cells such as inflammation and phagocytosis (engulfment by immune cells) triggered by the recognition of foreign particles.

An important function of the innate immune response is the release of chemical messengers known as cytokines and chemokines. These are small proteins used for cell-cell communication and to recruit immune cells to the site of infection/inflammation (3). Cytokines also contribute to the development of inflammatory responses such as fever, vasodilation, and vascular permeability (4, 5).

Cells of the innate immune system can recognise conserved structures common to many invading pathogens termed pathogen-associated molecular patterns (PAMPs). Examples include lipopolysaccharides in bacterial cell walls and double-stranded ribonucleic acid produced during viral infection (3). They also react to damage-associated molecular patterns (DAMPs) released by damaged or dying cells (6).

PAMPs and DAMPs bind to pattern-recognition receptors (PRR) on the surface of immune cells. PRRs can initiate phagocytosis, the process of cells engulfing and eliminating large particles such as microorganisms, dead cells, and foreign substances (7). Phagocytosis can also be triggered by a biochemical cascade known as the complement system. The complement system can identify foreign particles and coat them with extracellular proteins (opsonisation). This makes the opsonised particle a target for phagocytosis or may neutralise it directly.

1.1.1.1 Phagocytes

Phagocytosis is performed by a group of immune cells known as phagocytes which include neutrophils, monocytes, macrophages, and dendritic cells (DCs) (1).

During infection or tissue damage, PRR-mediated signalling on monocytes influences a complex cascade of pro- and anti- inflammatory responses (8, 9). Monocytes are recruited to the site of

infection where they are stimulated to differentiate into other phagocytes including DCs and macrophages. Differentiation is the process of cells changing their function and phenotype to enable specialisation into different cell types (10, 11). As well as phagocytosis, the resulting DCs and macrophages play an important role in the interface between the innate and adaptive immune systems.

1.1.1.2 Natural killer cells

Numerous other immune cells are involved in the innate immune response including natural killer (NK) cells, mast cells, basophils, and eosinophils (3).

NK cells have cell-killing or cytotoxic properties and are critical to the destruction of cancerous and virally infected cells. By releasing proteins called perforins and granzymes, NK cells cause lysis in a target cell, inducing programmed cell death (apoptosis) (3).

NK cells can distinguish abnormal from healthy cells, providing an anti-tumour response with limited off-target effects (12, 13). They can inhibit the proliferation and migration of cancerous cells (14), providing an important cytotoxic defence against cancer.

In addition to cytotoxicity, NK cells produce a large number of cytokines that modulate both innate and adaptive immune responses (15).

1.1.1.3 Other innate cells

Mast cells, basophils, and eosinophils are instrumental in acute inflammatory responses including responses to tissue damage, and allergic reactions (16). Mast cells act as “sentinel cells” in connective tissue, producing cytokines in response to infection or injury (3), whereas basophils are similar but typically reside in circulation. Eosinophils facilitate immune response to parasites among other regulatory functions (17).

The innate immune response can prevent most infections at the onset or eliminate them within hours. The targeting of PAMPs, which are structures not usually found in the human body, results in highly efficient self-nonself discrimination. As a result, they are less capable of distinguishing small differences between different invading molecules and are unable to elicit a specific response. This is where the second subsystem, adaptive immunity is critical.

1.1.2 Adaptive immunity

Adaptive immunity provides a second line of defence against microorganisms or foreign molecules which have evaded the initial innate response. It develops days after the infection (1) and adapts to recognise, neutralise and remember specific antigens (antigen-specific).

There are four main characteristics of adaptive immunity: antigen-specific effector responses, diversity of pattern recognition, self-nonself discrimination, and immunologic memory (18).

Antigen specificity allows subtle differences between antigens to elicit a specific response to a pathogen, which can be as small as a single amino acid change (19). This requires a large amount of diversity in the recognition molecules the system uses. This is in contrast with the innate system which provides a general response to many types of pathogens using a non-specific PRR.

The innate response is highly effective at self-nonself discrimination, as it targets molecular patterns that do not naturally occur in the body. In contrast, the diversity of antigens to which the adaptive system can respond increases the risk of self-components initiating an inappropriate immune response. This could lead to autoimmune disease (20). The adaptive response, therefore, requires a mechanism known as central tolerance to ensure it responds primarily to foreign antigens.

The immunologic memory of the adaptive system enables a rapid response if the same or a very similar pathogen is encountered again at a later date. This produces a heightened state of immune reactivity when re-exposed, conferring long-term immunity.

The development of antigen-specific responses, central tolerance, and immunologic memory is mediated by the interplay between two major groups of immune cells, the lymphocytes, and antigen-presenting cells.

1.1.2.1 B cells

B lymphocytes (B cells) are generated and mature in the bone marrow. When released from the marrow they express a unique membrane-bound antigen-receptor known as an antibody molecule, each of which binds to a specific antigen.

Genes encoding the antibody molecules undergo random genetic rearrangement (somatic recombination) during B cell development (20). This enables the recognition of a diverse range of different antigens. However, it can also produce self-reactive antibodies which bind to molecules usually found in the body, which could lead to inappropriate immune responses and autoimmune disease (20).

Negative selection is used to eliminate self-reactive B cells before they are released from the bone marrow. B cells which react to molecules present in the bone marrow undergo apoptosis or receptor editing (20). Whereas B cells tolerant of self-molecules are released into circulation, producing central tolerance.

Lymphocytes that have yet to be exposed to the corresponding antigen are called naïve cells. When a naïve B cell first encounters the antigen that binds its antibody, it begins to divide rapidly, and its progeny undergo differentiation. The activated B cells differentiate into two subpopulations, effector B cells, and memory B cells. Memory B cells are relatively long-lived, sometimes surviving for decades (21), and express the same membrane-bound antibody as the parent cell. Whereas effector B cells secrete soluble antibody molecules into the circulatory system.

Effector B cells are relatively short-lived, but during this time, they can secrete thousands of copies of the antibody molecule (19). The memory B cells remain inactive and rapidly produce effector B cells if the same antigen is encountered at a later date, helping to form an immunologic memory of the pathogen.

Secreted antibodies are an important mediator of the adaptive immune response. They bind to antigens on the surface of pathogens or that are free in circulation. This promotes opsonisation and activates the complement system resulting in phagocytosis or a cytotoxic response from NK cells (22, 23). In this way, the adaptive immune response can trigger innate immune effector mechanisms to neutralise or eliminate an invader.

As well as producing antibodies, B cells contribute to the activation of another important group of cells called T lymphocytes (T cells) via antigen presentation.

1.1.2.2 T cells & antigen presenting cells

T cells also form in the bone marrow but migrate to the thymus to mature. T cells express a unique membrane-bound antigen-receptor called the T cell receptor (TCR). In contrast to the antibody molecules of B cells, TCRs can't bind to antigen alone. Instead, TCRs recognise antigens that are bound to another cell membrane protein called the human leukocyte antigen (HLA).

There are two classes of HLA proteins: HLA class I, which is found of the surface of all nucleated cells in the body; and HLA class II, which is only expressed by certain immune cells including DCs, macrophages, and B cells which are referred to as antigen-presenting cells (APCs) (3).

If a cell is infected by a virus, or an APC has phagocytosed a foreign particle, HLA proteins will present antigen fragments on the surface of the cell. If the HLA-antigen complex is recognised by a TCR, this provides the T cell with an activation signal.

In this way, the development of an adaptive response is dependent on the actions of the innate immune system. As well as neutralising foreign particles, phagocytes such as DCs and macrophages activate T cells through antigen presentation and cytokine signalling. DCs are the most effective APCs (24), however, activated B cells also act as APCs by presenting fragments of antigens that they specifically recognise (3).

HLA proteins were first discovered due to their role in the rejection of foreign tissue in transplantation (1). They are encoded by a genetically diverse (polymorphic) complex of genes, and as a result, different individuals express different variants of the HLA protein. T cells can react strongly against these genetic variants, and the peptides that they present, in a process known as alloreactivity. As discussed later in this thesis (**Section 1.2.1**), this causes serious complications in a transplant setting.

The antigen specificity of T cells is determined by the TCR, which like the antibodies of B cells, are diverse due to gene rearrangement. A central tolerance mechanism is therefore required to eliminate self-reactive T cells. In the thymus, positive selection identifies T cells with a TCR which can bind to MHC proteins, then negative selection eliminates those that are self-reactive (25). Cells retained after this process migrate from the thymus as mature naïve T cells.

Various subpopulations of T cells have been characterised, the two best defined are the T helper cells (CD4+ T cells) and cytotoxic T cells (CD8+ T cells). These are distinguished by the expression of membrane-bound glycoproteins CD4 and CD8, respectively. A key difference is that CD4+ T cells are activated by HLA class II found only on the surface of APCs, whereas CD8+ T cells are activated by HLA class I which is present on all nucleated cells.

There are also T regulatory cells (T regs) that express both CD4 and CD25 and the transcription factor FoxP3. T regs have an important role in suppressing the generation and ongoing activity of immune responses (25). Gamma delta (γδ) T cells are defined by what chains are present in the TCR molecule, they express γδ TCR instead of the αβ TCR expressed by the majority of T cells (26). These cells have various immune functions and are involved in inflammatory responses (27).

Once a naïve T cell has been activated by the appropriate HLA-antigen complex, it causes the cell to proliferate rapidly and to differentiate into various effector T cells and memory T cells with the same antigen affinity. Effector cells are short-lived and migrate to the site of infection to eliminate the pathogen, whereas memory cells are longer-lived than the naïve parent cell.

CD4+ effector T cells are referred to as “helper” cells as they enable the activation of B cells, macrophages, and CD8+ T cells (19). They do not kill cells directly but stimulate the effector responses of other immune cells. By stimulating APCs such as macrophages and B cells, which in turn enable further T cell activity, it provides a self-reinforcing immune response to a pathogen. When activated by APCs, a subset of naïve CD4+ T cells also differentiate into memory cells.

In contrast, naïve CD8+ T cells can recognise almost all nucleated cells through their presentation of HLA class I (3). They play an important role in eliminating cancerous cells, virally infected cells, and foreign cells. CD8+ effector T cells are referred to as cytotoxic T cells because they have cell-killing or

cytotoxic abilities. Like their CD4+ counterparts, a subset of naïve CD8+ T cells will also differentiate into memory cells upon activation.

When naïve lymphocytes such as T cells and B cells first encounter antigen, a primary immune response results in a clonal expansion of cells with the same antigen affinity. The product of this expansion is the effector and memory cells. If the memory cells subsequently encounter the same antigen, it triggers a heightened secondary response targeting that pathogen (18). This is fundamental to the immunologic memory of the adaptive immune system.

An effective immune response therefore depends on a complex interplay between different cellular elements of the adaptive and innate immune systems. Immune cells elicit effector responses which in turn activate more immune cells in mutually reinforcing processes. However, the immune cells which underpin this system have a limited life span, ranging from several weeks to several years (28), and therefore need to be constantly replenished.

1.1.3 Haematopoiesis

Haematopoietic stem cells (HSC) are of special interest because they are able to differentiate into all other types of blood cells (28), including immune cells, red blood cells (erythrocytes), and platelets (produced by megakaryocytes). They also have the ability to self-renew, sustaining a stable population of HSC. This enables a small population of HSC to maintain the production of new blood cells throughout life, in a process known as haematopoiesis.

At an early stage of haematopoiesis, HSC differentiate along two distinct pathways or lineages of cells. One of these pathways generates lymphocytes and the other myeloid cells. All blood cells can be categorised into one of these two main groups (28, 29), both of which have a common ancestral origin, the HSC.

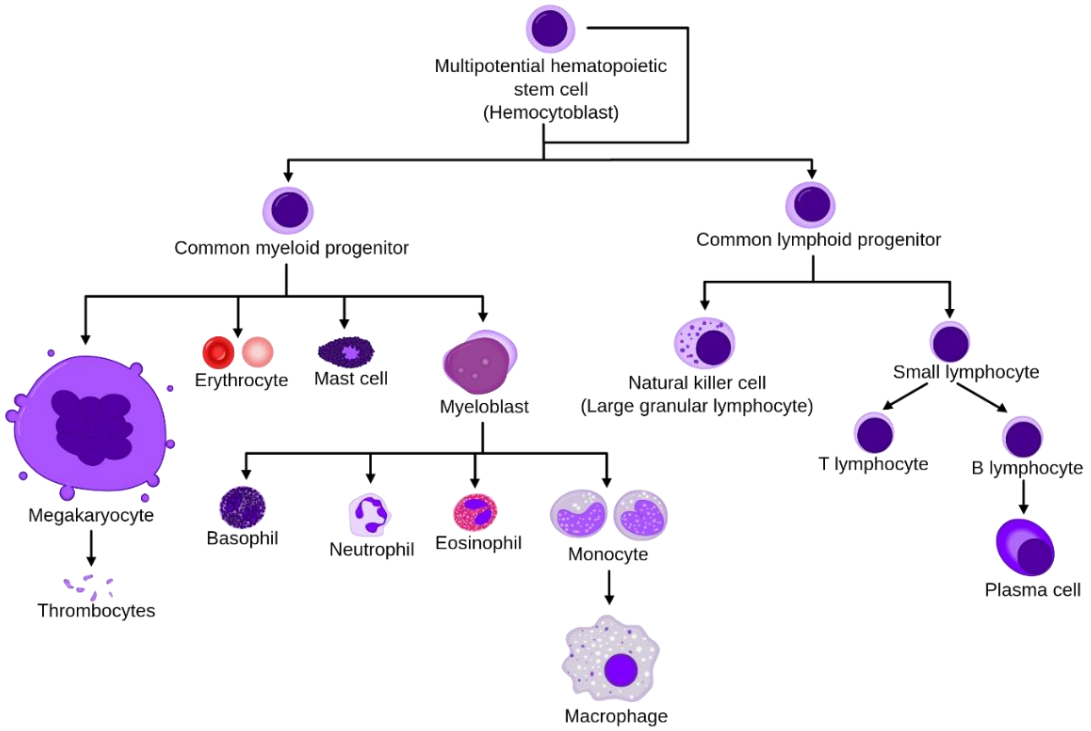


Figure 1.1 Simplified hierarchy of haematopoiesis. Figure by A. Rad and M. Häggström. CC-BY-SA 3.0 license. (30)

During haematopoiesis there is a progressive loss of differentiation potential resulting in the gradual commitment of cells to a particular cell type (31). This process is often described as cells changing from an immature to a mature phenotype (maturation). HSC reside in the bone marrow, which is the primary site of haematopoiesis. However, maturation of lymphocytes also occurs in other organs such as the thymus, lymph nodes, and spleen.

HSC give rise to populations called progenitor cells which have lost their ability to self-renew and are committed to a particular lineage. Common lymphoid progenitors (CLP) generate lymphocytes, whereas common myeloid progenitors (CMP) produce myeloid cells (28). The descendants of CLP and CMP which are referred to as committed precursors are further restricted in their differential pathways.

CLP produce T cells, B cells, and NK cells whereas CMP generate erythrocytes, thrombocytes, and a range of immune cells (monocytes, macrophages, neutrophils, mast cells, basophils, and eosinophils). An exception is DCs which can arise from the progenitors of both lineages.

1.2 Haematopoietic stem-cell transplantation

Transplantation refers to the transfer of cells, tissues, or organs (a graft) from one site to another. A variety of diseases can be cured by implanting a healthy graft from one individual (the donor) to another (the recipient). Transplanting HSC from healthy donors (allogeneic haematopoietic stem-cell transplantation (allo-HSCT)) is an effective treatment for a wide range of haematological malignancies.

Haematological malignancies are cancers of the blood, bone marrow, and other immune organs such as lymph nodes (32). This broad category of diseases includes myeloma, leukaemia and lymphoma. Collectively these were the third largest cause of cancer death in the United Kingdom in 2016 (33), and accounted for 9% of all newly diagnosed cancers in the United States in 2017 (32).

The curative effects of allo-HSCT derive from two mechanisms (34), eradication of the disease through the cytotoxic effects of a conditioning regimen (35), and through the transfer of a healthy immune system capable of eliminating residual cancer cells (36, 37).

Conditioning is the preparative regimen administered before the infusion of the donated cells (38) and can include a combination of chemotherapy and/or radiotherapy and/or immunosuppressant drugs. Conditioning serves several purposes: to eradicate cancerous cells through cytotoxic effects and to suppress the immune system (38).

Conditioning alone is insufficient to control the disease as some malignant cells are capable of surviving even lethal doses of radiotherapy and chemotherapy (36). However, these cells can be eliminated by immunologically active cells from a healthy donor (39). This is known as the graft-versus-leukaemia (GvL) effect, which is mediated by T cells and NK cells (40-42).

The graft also enables reconstitution of the immune system after myeloablation (the depletion of bone-marrow tissue) caused by the conditioning regimen. Experiments showed that a single HSC can repopulate the immune system of a lethally irradiated mouse (43), and transplantation of HSC can protect animals from otherwise lethal doses of total body irradiation (44, 45). Insights such as these led to the development of allo-HSCT to rescue recipients from the toxicity caused by intensive conditioning regimens (35, 39). However, it is now known that the potent GvL effect is also responsible for reducing their cancer burden (39).

Recognition of the GvL effect led to the development of reduced intensity conditioning (RIC) with lower doses of chemoradiotherapy. This has enabled the treatment of older patients (46), and patients with comorbidities that would be unable to tolerate high levels of myeloablative conditioning(47).

However, without additional immune suppression, this may increase the risk of the graft being rejected (36, 47, 48).

Whilst relapse of the original malignancy remains a considerable risk (49), this approach can cure a significant proportion of patients with otherwise incurable disease (50, 51). Reciprocal immune reactions between the recipient's cells and those derived from the donor are fundamental to the success of this therapy by enabling the GvL effect. However, as well as benefits, these immune reactions can have deleterious consequences.

1.2.1 Graft vs host disease

If immune cells derived from the graft react against the recipients' healthy cells, this can lead to graft-vs-host disease (GvHD). GvHD is a significant cause of morbidity and mortality following allo-HSCT (52).

Less frequently, the recipient's residual immune cells, not destroyed in the conditioning process, may react against the graft leading to graft rejection (39). As a result, the conditioning regimen must be sufficiently immunosuppressive to limit the risk of the recipient rejecting the donor cells (53).

These complications are related to genetic variations of the HLA protein expressed by the recipient and donor-derived cells (40). If T cells recognise these variations as foreign antigens, it can result in an alloreactive immune response.

HLA proteins are encoded by a highly polymorphic complex of genes that determines histocompatibility between different individuals (54). If a graft expresses the same set of HLA genes to the recipient, it is said to be histocompatible (HLA-matched) and is less likely to cause alloreactivity.

The highest risk of GvHD and graft rejection occurs in patients that are not HLA-matched with the donor (39). As a result, the use of HLA-typing to select compatible donors has been crucial to the success of allo-HSCT (46). Although GvHD remains a risk for HLA-matched recipients, as cells from the graft can react against "minor" histocompatibility antigens which lie outside the main HLA loci (39, 55).

Notably, the cancer-eradicating GvL effect is strongest in allo-HSCT patients who develop GvHD (39). HLA-mismatched donor T cells have powerful cytotoxic effects which contribute to the GvL effect in addition to NK cell cytotoxicity (56-58).

This indicates that for a successful response to the therapy, a balance between the GvL effect and GvHD is required, both of which are determined by the alloreactive response of donor-derived immune cells. Separating the mechanisms which underpin GvL effects from those of GvHD is a pivotal goal of research into allo-HSCT.

1.2.1.1 HLA-typing

HLA genes are inherited as haplotypes, for example, a sibling has a one-in-four chance of being HLA-matched, and a 50/50 chance of being haploidentical (half-matched). The preferred source of donor cells is an HLA-matched sibling donor (MSD) (59), but these are available for under a third of allo-HSCT recipients(60, 61).

Registries of potential donors can be used to identify matched unrelated donors (MUD) (62). Patients receiving a MUD transplant have comparable outcomes to those infused with cells from an MSD (63-65). However, a match is more likely between individuals of the same ethnicity (66) and the largest registries are biased toward donors from North America and Europe(61).

In the absence of a matched donor, other options include mismatched unrelated donors that are partially incompatible, and haploidentical donors (67). This is generally regarded as suboptimal as

increased HLA mismatch is associated with increased risks of GvHD (61, 68). However, family members can act as haploidentical donors meaning that this can simplify donor selection. Indeed, haploidentical donors can be identified for >95% of allo-HSCT patients, dramatically increasing the pool of potential donors (69).

Recently, immunosuppressive regimes have enabled the effective use of haploidentical allo-HSCT (haploHSCT) (69). In particular, the administration of post-transplant cyclophosphamide (67) and graft manipulation to deplete the number of T cells(70). However, haploHSCT is primarily reserved for situations when there is an urgent need for a transplant to avoid early relapse (71).

Whilst closer HLA-matching decreases GvHD risk, mismatched antigens are also expressed by cancer cells. Therefore, some of the discrepancies in HLA proteins may be important in the GvL effect (72).

The degree of HLA mismatch between recipient and donor is one of the strongest risk factors for the development of GvHD (73, 74) .

1.2.1.2 Acute & chronic graft vs host disease

GvHD occurs when T cells infused with the graft, or derived from the graft, recognise the recipient's healthy cells as foreign. This results in T cell activation and a cytotoxic response attacking the recipient's cells. The two main clinical presentations are acute GvHD (AGvHD) and chronic GvHD (CGvHD).

AGvHD is the leading cause of short-term (day 100 and 1 year) mortality after allo-HSCT (75). AGvHD occurs in approximately 40-60% of recipients within the first 3-6 months post-transplant (75-77). It typically affects the skin, gastrointestinal tract, and or liver and symptoms include a rash, diarrhoea, or abnormal liver function (39) .

Clinical, laboratory, and biopsy assessments of the target organs (72) are used in AGvHD diagnosis. By measuring the extent of involvement of each of the three main target organs, AGvHD severity is graded as I (mild), II (moderate), III (severe), and IV (very severe)(55) . Severe AGvHD is associated with a poor long-term prognosis for allo-HSCT patients (55).

The pathophysiology of AGvHD is thought to occur in a three-phase process: initial tissue damage is caused by conditioning which activates the recipient's APCs by PAMPs and DAMPs (**Section 1.1.1**); activation and proliferation of T cells infused with the graft; a final stage of cytotoxic and inflammatory responses which result in the death of cells (necrosis) (55, 72) .

CGvHD is the leading cause of late death without prior relapse (non-relapse mortality (NRM)) in allo-HSCT patients surviving beyond two years (78, 79) . Occurring in around half of the patients surviving until this time point, CGvHD typically begins between 3 months and 2 years post-transplant (78). It is less likely to occur in children and incidence rates rise significantly with age (80).

Any organ system in the body can be affected by CGvHD, and it is diagnosed based on the clinical severity of symptoms across eight different organs, as well as laboratory tests (79). Overall severity of mild, moderate, and severe is calculated based on the number of affected organs and level of functional impairment (79).

Unlike AGvHD which is primarily mediated by the alloreactive response of T cells infused with the graft, CGvHD has a more complex pathophysiology (55). It is related to the breakdown of central tolerance mechanisms in donor-derived T cells and B cells and dysregulation of the innate immune responses of macrophages, DCs and neutrophils (79). This produces chronic inflammation resulting in scar formation (fibrosis) in the affected organs.

Whilst they have a distinct biology, patients with severe AGvHD are more likely to experience CGvHD (78). This may be due to AGvHD targeting the thymus, the site of T cell maturation (81). This could result in the failure of the central tolerance mechanism, leading to the self-reactive T cells being released without negative selection (78).

The principal approach to GvHD prevention (prophylaxis) beyond the selection of an appropriate HLA-matched donor, is pharmacological immune suppression (82).

A common immune suppressant used for GvHD prophylaxis is cyclosporine (CyA)(76), which interferes with the activation and proliferation of T cells (83). This is used as a standard GvHD prophylaxis in HLA-matched allo-HSCT, often used in combination with methotrexate (MTX) or mycophenolate mofetil (MMF)(84). However, these drugs have toxic side effects and it is unclear if the nonspecific immunosuppression of such therapies also reduces positive GvL effects (40, 85),

As well as negative effects, patients with CGvHD experience lower rates of both relapse and deaths preceded by a relapse (relapse-related mortality (RM)) (78, 86), suggesting the immune reactions mediating CGvHD can have a protective effect against cancer.

1.2.2 Engraftment

A key aim of an allo-HSCT is that the transplanted cells can sustain effective haematopoiesis. For this to occur, the transplanted HSC must first migrate to niches in the bone marrow in a process called engraftment (87). A niche is an essential microenvironment with the correct conditions for HSC to survive and proliferate (88).

A lack of HSC engraftment following allo-HSCT is known as graft failure (89). Although a relatively rare occurrence, a lack of engraftment within the first month following transplantation is associated with a significant risk of mortality (89, 90). Engraftment is usually defined by the first of three consecutive days where the recipients' peripheral blood neutrophil count is above >500 x 10^6 / L (91), showing that haematopoiesis has resumed. Both neutrophils and monocytes show early reconstitution post-transplant (92).

Successful engraftment of HSC can establish durable haematopoietic chimerism in a recipient of allo-HSCT. Chimerism is where cells from two genetically distinct individuals coexist and are tolerant of each other (93, 94) . Chimerism can be 'complete' where only donor HSC are present or 'mixed' where both donor and recipient HSC coexist. This is facilitated by the central tolerance mechanism. Lymphocytes produced in the bone marrow migrate to the thymus and undergo negative selection to remove recipient-reactive and donor-reactive cells (94).

Engraftment of the donor-derived cells can be assessed by chimerism analysis, which is useful for detecting impending graft rejection (95). This involves quantifying the proportion of donor-derived haematopoietic cells relative to those originating from the recipient in the peripheral blood or bone marrow (96).

Graft failure is more likely to occur if there is an HLA disparity between the recipient and donor (87). A variety of other factors affect engraftment including the source of the stem cells and the conditioning regimen (89).

1.2.2.1 Graft source

The cellular material for transplantation was historically derived primarily from the bone marrow (BM), as a result, allo-HSCT are often referred to as "bone marrow" transplants. However, the majority of stem cell grafts are now extracted from the peripheral blood of the donor (PBSC).

Usually, HSC reside mostly in the bone marrow with few in circulation. PBSC collection by apheresis, therefore, requires granulocyte colony-stimulating factor (G-CSF) to mobilise HSC to the peripheral circulation. Alternatively, umbilical cord blood (UCB) can be used as a source of stem cells.

Engraftment occurs fastest in patients infused with PBSC (97) whereas BM and UCB grafts are consistently associated with delayed engraftment (87, 98) . Whilst the incidence of graft failure is similar between BM and PBSC following allo-HSCT with an MSD (99), it is higher when the BM is from an unrelated donor (100). Furthermore, grafts from UCB are associated with a higher risk of graft failure (101).

1.2.2.2 Conditioning regimen

It is thought that engraftment may be limited by the number of niches in the bone marrow (102), resulting in competition between recipient and donor HSC. Therefore, depletion of bone marrow tissue by the conditioning regimen may encourage engraftment by freeing niches for the donated cells to engraft.

Conditioning can include a heterogeneous mix of different chemotherapies and/or total body irradiation (TBI) and/or immunosuppressive drugs. The process has two main components: myeloablation targeting HSC of the recipient, and lymphodepletion which targets their lymphocytes (38).

Chemotherapy functions by eliminating rapidly dividing cells such as malignant cells (103). It does so by preventing cells from dividing or causing programmed cell death (apoptosis) (103). However, this also affects other proliferating cells such as HSC and immune cells. Some chemotherapeutic compounds such as busulfan produce more of a myeloablative effect, whilst others are more lymphodepleting including cyclophosphamide and fludarabine (38).

TBI uses ionizing radiation to kill cells by damaging deoxyribonucleic acid (DNA). When delivered to the entire body, it provides cytotoxic effects and immunosuppression (104). As a result, TBI can both eliminate cancer and free niches in the bone marrow for engraftment. However, TBI is complicated by both short and long-term toxicities (104).

The immunosuppressive drugs used in conditioning may consist of anti-thymocyte globulin (ATG) or alemtuzumab (105). These are antibodies used to eliminate T cells, but can also target other lymphocytes such as B cells (106) and NK cells (107). By suppressing the action of T cells against the graft, they reduce the risk of graft rejection and graft failure (105). These drugs can also persist in the body after transplantation which can reduce the risk of GvHD (106). However, this may also lower the GvL effect and increase the risk of opportunistic infections (36, 108, 109) .

Depending on different patient characteristics and their specific haematological malignancy, a combination of chemotherapy with or without TBI is administered as part of the conditioning regimen.

1.2.2.3 Myeloablative and reduced intensity conditioning

Myeloablative conditioning (MAC) refers to regimens that cause irreversible cytopenia (a reduction in the number of blood cells) that require the transplant of stem cells for the recipient to recover (53). MAC involves high-dose chemotherapy which “ablate” the bone marrow often using a combination of cyclophosphamide and TBI, or busulfan and cyclophosphamide (38).

The profound myeloablative effects of MAC ensure niches are available in the bone marrow to enable HSC engraftment, it suppresses the immune system reducing the chances of graft failure, and the cytotoxic effects help to eliminate the patient’s cancer.

MAC was historically the most common form of conditioning. However, MAC is associated with high regimen-related toxicity and non-relapse mortality (NRM)(38). This prohibits its use with elderly patients or those with a high number of comorbidities. As half of haematological malignancies are diagnosed in those over 65 years of age, this is a severe limitation (110).

This led to efforts to reduce the intensity of the conditioning regimen with lower doses of chemotherapies or TBI (38, 111). Non-myeloablative conditioning (NMAC) and reduced-intensity conditioning (RIC) were introduced to decrease the amount of regimen-related toxicity and NRM.

RIC and NMAC cover a heterogenous group of treatments, often including fludarabine and/or reduced doses of chemotherapy and/or TBI. NMAC results in minimal and reversible levels of cytopenia (111). Whereas RIC is defined as regimens that do not fit the criteria for either MAC or NMAC (111).

RIC and NMAC have enabled effective treatment for a large number of older patients and those otherwise ineligible for allo-HSCT (38). Reducing conditioning intensity is also associated with lower rates of AGvHD and NRM, but similar rates of CGvHD (38, 112).

However, reducing conditioning intensity is not without trade-offs. The host immune system may persist which can result in lower rates of engraftment relative to MAC (47, 87, 113) .

The cancer-eradicating effects of RIC relies primarily on the GvL effect of the transplanted cells rather than the conditioning regimen (114). Higher-intensity conditioning has also been associated with lower relapse incidence (38). As a result, RIC may decrease transplant-related toxicity but increase the probability of relapse (49).

1.2.3 Relapse

Allo-HSCT is often used as a treatment of last resort for patients with haematological malignancies unlikely to be controlled by other less intensive procedures (57). As a result, the haematological malignancies treated with allo-HSCT are amongst the most resistant to treatment, and relapse remains a major cause of mortality (49).

Early relapses could result from insufficient cancer reduction in conditioning, leading to an ineffective GvL effect, or from the GvL effect never becoming established (57). Relapse can also occur after a prolonged period of an effective GvL effect. This can result from a weakening of the immune system, becoming tolerant to cancer-associated antigen, or the disease mutating enabling immune escape (57).

Cancers can evolve by clonal expansion and genetic diversification (somatic mutation). The massive cell death caused by chemotherapy and GvL effects provide a strong selective pressure. Cells capable of evading these therapeutic and immune mechanisms have a selective advantage (115). This is referred to as immune escape or antigenic escape.

In HLA-mismatched all-HSCT, cancerous cells have been found to evade recognition from donor T cells by deleting mismatched HLAs, resulting in relapse (56, 116) . Immune escape by mutation has also been shown to reduce the susceptibility of cancerous cells to NK cell mediated toxicity (42).

The propensity of the particular disease to mutate is therefore critical to relapse risk. The GvL effect also develops over time, so cancers that proliferate more slowly such as chronic leukemias and lymphomas may be more susceptible (57).

As well as the specific haematological malignancy, the disease stage can result in markedly different relapse rates. Generally, the risk is lowest for patients transplanted during complete remission but higher in partial remission or with refractory disease (57).

Immune recovery following transplant is closely associated with protection from relapse (57). The level of chimerism is also predictive of subsequent relapse risk, supporting the importance of successful engraftment of donor cells (117, 118) .

The immune system may fail to control the disease due to immunosuppression administered to control GvHD negatively impacting the GvL effect (119). T cells may also become exhausted due to an abundance of cancer-associated antigen, which could lead to the inhibition of cytotoxic activity by programmed cell death protein 1 (PD-1)(49). Or the opposite, where a paucity of antigen stimulation causes inactivity within memory T cell populations (anergy). All of which weaken the immune response to the disease.

Some of the differences in relapse risk may be attributed to differences in the cellular content of the graft. Donors with a high content of suppressive cells such as T regs are associated with less GvHD but may confer a reduced GvL effect (57, 120). Moreover, the count of NK cells in the graft expressing low levels of a cell surface protein CD56 ($CD56^{dim}$ NK cells) has been related to the rate of relapse, further demonstrating the importance of graft composition to patient outcomes (121).

1.2.4 Graft composition and manipulation

A sufficient number of HSC in the graft is required to overcome the toxicity of the conditioning regimen and facilitate immune reconstitution.

A cell surface protein known as CD34 is expressed by almost all HSC and other immature progenitor cells. The number of CD34-positive ($CD34^{+}$) cells can be used as a proxy for the number of HSC the graft contains, which is an important metric to assess graft quality (122). The a cell dose of $\geq 4 \times 10^6$ CD34+ cells / kg of body weight is generally regarded as adequate for an allo-HSCT (123).

Graft manipulation can be used to optimise the cellular composition and volume of cells in the graft. The graft can be manipulated *ex vivo* to enrich the number of CD34+ cells relative to other cells (CD34+ enrichment). This results in a high number of HSC and relatively few T cells and other immune cells (124). There are conflicting reports on the benefits of CD34+ enrichment. It has consistently been shown that a higher dose of CD34+ cells can improve engraftment following allo-HSCT (125, 126) .

However, the effect on other outcomes is less certain. Whilst some studies indicate CD34+ enrichment reduces CGvHD (127, 128) , others have reported a higher CGvHD with increased CD34+ dose (126, 129) . Similarly for overall survival, some reports suggest increased CD34+ doses are associated with improved survival (126, 130, 131) whilst others indicate no effect on overall survival (132-134) . There is even evidence of negative effects at very high doses, suggesting a complex relationship between CD34+ dose and response (129, 135) .

Grafts can be manipulated to deplete the number of T cells (T cell depletion). T cells are identified by the expression of a cell surface marker CD3, so are described as CD3-positive (CD3+). Negative selection of CD3+ cells from the graft *ex vivo* can specifically deplete T cells while retaining other lymphocytes such as NK cells which have GvL capacity (124).

Ex vivo T cell deletion reduces rates of acute and chronic GvHD(136) and is widely used to prevent GvHD in the setting of halpoHSCT (70).

Alternatively T cells can be depleted in vivo, by administering ATG (109, 137) or alemtuzumab, this has also been shown to reduce the rate of acute and chronic GvHD (136). However, T cell depletion may not improve overall survival rates and may increase the risk of opportunistic infections (109, 138) .

T cells expressing αβ TCR are primarily responsible for the development of GvHD (139, 140). Routines have been developed to negatively select this particular subpopulation of T cells (αβ TCR+ depletion), whilst retaining γδ T cells and other lymphocytes such as NK cells (141). This has been associated with an improved immune reconstitution after haploHSCT (142), reducing the chances of opportunistic infections (143).

Aβ TCR depletion is often combined with the depletion of B cells, identified by the expression of cell surface protein CD19 (CD19+ depletion). B cells can cause a condition called post-transplant lymphoproliferative disorder (PTLD) in patients infected with Epstein–Barr virus (EBV). EBV is a ubiquitous virus that infects 90-99% of adults by mid-adulthood (144). αβ TCR+ / CD19+ depletion effectively prevents PTLD, with the same benefits of αβ TCR+ described above (141).

In addition to HSC, T cells, and B cells for which there are established graft manipulation routines, grafts contain a variety of other mature immune cells.

An increased count of certain DC subpopulations in the graft has been associated with increased relapse incidence and CGvHD incidence (145). Whereas a high count of T-regs has been linked with lower relapse incidence (120). Furthermore, the number of NK cells has been shown to have an impact on relapse incidence following allo-HSCT (121). Specifically, a higher count of NK cells expressing lower levels of CD56 (CD56 dim NK cells) has been linked with lower relapse incidence.

It remains unknown if there is an optimal graft composition (146). The precise composition is likely to be correlated with differences in patient outcomes but is relatively understudied (121). Investigating the impact of graft composition on the outcome of patients following allo-HSCT is a central theme of the project described in Chapter 3 of this thesis.

1.2.5 Clinical endpoints

A successful response to allo-HSCT is survival, without graft failure or ongoing causes of morbidity such as GvHD, relapse, and infection. To study the factors influencing the incidence of these outcomes, clinical endpoints need to be defined.

The gold-standard endpoint in most clinical research is overall survival (OS). It is measured as the time-to-death attributable to any cause. In allo-HSCT research, the time interval typically begins at the time of transplant.

OS is a direct measure of clinical benefit that is unambiguous and easy to measure. However, OS fails to capture ongoing morbidity in a transplant setting which profoundly impacts a patient's quality of life. The composite outcome of GvHD-free, relapse-free survival (GRFS) was developed specifically to study the outcome of allo-HSCT recipients.

GRFS is defined as the absence of grade III-IV AGVHD, chronic GVHD, relapse, and death(147). It is measured as the time-to-event-onset (using the first event) and effectively captures the real recovery of patients without adverse events.

Neither OS nor GRFS provide information on the particular cause of morbidity or mortality. To study the influence of factors on individual outcomes, clinical endpoints can be divided into several competing risks. These, however, introduce additional statistical complexity which is discussed in **Section 1.3.2**.

Death without prior relapse is defined as non-relapse mortality (NRM) (148). When the cause of death is adjudicated specifically to transplant related causes, NRM is referred to as transplant-related mortality (TRM). Major risk factors for NRM include GvHD, infections, and organ failure (52). Whereas

deaths preceded by a relapse or progression of malignancy are defined as relapse-related mortality (RM) (149). As patients can only experience either NRM or RM, these are considered competing risks. Alternatively, the onset of the non-fatal events may be assessed, for example, the time-to-relapse (148).

Given a set of clearly defined clinical endpoints, the factors which influence the outcome of allo-HSCT recipients can be assessed objectively between studies.

1.2.6 Factors influencing patient outcomes

A variety of patient and donor characteristics have been found to influence the outcome of allo-HSCT patients. This is in addition to the therapeutic choices (e.g., graft source, conditioning regimen, GvHD prophylaxis), donor graft characteristics (e.g., HLA-type, graft composition), and patient characteristics (e.g., diagnosis, disease stage) already discussed.

Comorbidity refers to the presence of more than one distinct condition in an individual (150). Patients with better overall functional performance and less comorbidities have better outcomes following allo-HSCT (151, 152). The negative effect of comorbidities on allo-HSCT patients is likely due to the burden of chronic disease and how it interacts with both the treatment and the malignancy (151). Comorbidities increase the risk of toxic effects in response to the treatment, which can negate some of the intended benefits (153).

Compared with younger patients, elderly recipients experience higher rates of transplant-related morbidity and mortality (154). Older patients are more likely to have comorbidities and impaired health (155). Furthermore, they may have malignancies with more adverse disease features which may be subject to less aggressive therapy, due to their susceptibility to toxic side effects (154).

The age of the donor is also influential. Recipients of grafts from older donors have worse rates of OS (156-158). Other than the selection of HLA-matched donors, age is one of the most important factors in donor selection affecting recipient outcomes (159).

The donor's sex and blood type may also be risk factors. Female recipients of grafts from male donors may have an increased risk of GvHD (160, 161). Whereas studies indicate blood type mismatch is associated with complications following transplant, however the precise relationship between GvHD and NRM remains uncertain (162, 163) .

The time from diagnosis to transplantation is also predictive of NRM and relapse incidence (164), with larger intervals generally associated with a worse prognosis for patients with otherwise equal risk (164).

Cytomegalovirus (CMV) is a major cause of morbidity and mortality following allo-HSCT (165). CMV is prevalent globally, with seroprevalence ranging from 40-50% in economically developed countries and up to 95% in developing countries (166). After primary infection, CMV becomes a latent infection that may be reactivated when the immune system is suppressed leading to CMV disease.

The highest risk of CMV reactivation occurs in allo-HSCT patients that are CMV positive at the time of transplant, irrespective of the donors CMV status (167). Furthermore, CMV positive donors are associated with worse OS for CMV negative recipients (168).

The multitude of factors influencing patient outcomes necessitates tools to assess pre-transplant risk. A variety of risk scores or risk indices have been developed to guide clinical decision-making.

1.2.6.1 Risk scores

The European Group for Blood and Marrow Transplantation (EBMT) developed the EBMT risk score to assess the pre-transplant risk of individuals (164). This uses five factors including recipient age, disease stage, the time interval from diagnosis to transplant and donor-recipient gender combination to estimate risk. This provides an additive risk score that can effectively discriminate between high and low-risk patients in a variety of international cohorts (169, 170) .

The haematopoietic cell transplantation-specific comorbidity index (HCT-CI) was developed to estimate pre-transplant risk based on diagnosed comorbidities (171). By accounting for 17 different comorbidities at the time of transplant, the HCT-CI is able to stratify patients into three risk categories.

The discrimination of classification models can be measured using the area under the receiver operating characteristic curve (AUC). This indicates the ability of a model to differentiate between patients who do experience an outcome, from those that don't, at a specified time point. Perfect predictions result in an AUC of 1.0, whereas an 0.5 is equivalent to a random guess.

The EBMT risk score was estimated by the authors to have an AUC of 0.63%when predicting survival 5 years post-transplant (172). Furthermore, a recent external validation estimated an AUC of 0.58 for survival at 2 years post-transplant (173). The HCT-CI was published with an AUC estimate of 0.66 for survival at 2 years post-transplant (171). Whereas external validation studies have indicated a 2-year AUC between 0.55-0.64 (173, 174) .

More recently, scores based on biomarkers such as Endothelial Activation and Stress Index (EASIX) have shown relatively good discrimination at predicting NRM (2 year AUC 0.65) but with worse discrimination for overall survival (2 year AUC 0.60)(173). EASIX utilises laboratory measurements of lactate dehydrogenase, creatinine, and the number of thrombocytes. EASIX scores indicate the level of endothelial dysfunction, which has been related to the development of AGvHD (175).

Another biomarker-based risk score is the revaluated Pre-transplantation Assessment of Mortality (rPAM). rPAM incorporates 8 factors including patient age, donor type, disease risk, conditioning regimen, forced expiratory volume (lung function), carbon monoxide diffusion capacity, serum creatinine level, and serum alanine aminotransferase concentration. Where disease risk is estimated based on chromosomal changes (cytogenetics (176)) and disease stage (177). Using these factors rPAM results in competitive discrimination for overall survival (2 year AUC 0.64) and NRM (2 year AUC 0.66) (173).

These scores have proven useful for stratifying allo-HSCT patients by risk. Despite the wide range of factors included in the various risk indices, a considerable amount of variation in the outcome of patients remains unexplained. As a result, individual risk prediction presents a considerable challenge. The search for additional predictive biomarkers for disease risk is an active area of research, which I aim to contribute to as part of this thesis.

1.2.6.2 Biomarkers

Biomarkers (**bio**logical **markers**) are defined as "objective indications of medical state observed from outside the patient - which can be measured accurately and reproducibly" (178).

Biomarkers can broadly be divided into diagnostic, prognostic, predictive, and response-to-treatment (118, 179). Diagnostic biomarkers are used to detect the onset of a disease. Whereas prognostic biomarkers provide information on the risk of an outcome or disease occurring.

Using GvHD as the example outcome, a diagnostic biomarker would indicate that a patient likely has GvHD, whereas a prognostic biomarker would indicate the risk of GvHD occurring before the onset.

Predictive biomarkers, in contrast, provide information on the risk of an outcome or disease occurring, before the initiation of a particular treatment. Whilst response-to-treatment biomarkers monitor their response after the treatment has been administered.

In the setting of GvHD, a predictive biomarker would indicate the risk of GvHD occurring if a patient is given a particular GvHD prophylaxis. Whereas a response-to-treatment biomarker might indicate the response of the patient to a prophylactic treatment and could be used as a surrogate clinical endpoint to study efficacy.

A wide variety of biomarkers for outcomes following allo-HSCT have been identified. Chen and Zeiser (2020) provide a comprehensive summary of many novel and validated biomarkers for AGvHD, responsiveness to immunosuppressive treatments, NRM, relapse and subsequent death (118).

Once a potential biomarker has been identified in a sufficiently large cohort, it must then be verified or validated in an independent cohort (118). It must then be qualified, which is a regulatory procedure to ensure a biomarker can be relied upon to have a specific interpretation (180).

A large number of molecules in the plasma have been qualified as prognostic biomarkers of increased AGvHD incidence, including interleukin 6 (IL-6), a cytokine involved in the activation of T cells (181, 182); C-X-C motif chemokine ligand 10 (CXCL10) a chemokine involved in immune cell migration (183); a soluble domain of T-cell immunoglobulin and mucin domain 3 (Tim-3), involved in negative regulation of T-cell activation (182, 184).

Whilst not qualified, several studies have associated a soluble form of CD25 with increased occurrence of AGvHD (185, 186) . Soluble CD25 is associated with inflammation and lymphocyte cytotoxicity (187). It is also included in a biomarker panel for increased NRM and lower OS (188) that contains CD25, tumour necrosis factor receptor 1 (TNFR1), interleukin 8 (IL8) and hepatocyte growth factor (HGF). IL-8 is involved in immune cell recruitment and plays a role in (189), TNFR1 has a regulatory role in innate immune responses (190), whereas HGF is a molecule involved in tissue repair (191).

Several other tissue-repair related molecules have been qualified as prognostic biomarkers for NRM. This includes regenerating islet-derived protein 3 alpha (REG3α) and suppression of tumorigenicity 2 (ST2) (191-194). When measured 7 days post-transplant, a combination of ST2 and REG3α can be used to estimate the risk of NRM and GvHD related mortality (195, 196).

In contrast, there is a relative paucity of validated and qualified biomarkers for relapse following allo-HSCT that are measurable around the time of transplantation. Lower peak levels of interleukin 15 (IL-15) at day 14 post-transplantation have been linked with relapse (185). IL-15 is a proinflammatory cytokine that activates T and NK cells (197). A high ratio of two metabolites, stearic acid, and palmitic acid, on day 7 post-transplant is associated with increased relapse, but lower AGvHD (198). However, neither of these biomarkers has been successfully validated.

Several studies have indicated mixed rather than complete chimerism is prognostic of greater relapse incidence, when measured one month or later post-transplants (199-203) .

Pre-transplant measurable/minimal residual disease (MRD) has been associated with increased relapse and worse overall survival following allo-HSCT (204, 205). Pre-transplant MRD, and markers associated with it provide an exciting avenue for exploration. However, the impact of MRD on relapse has mainly been explored at time points beyond a month post-transplantation (206, 207).

The projects in this thesis aim to identify prognostic biomarkers that are measurable either pre-transplant or in the very early period following allo-HSCT.

1.3 Survival and competing event analysis

Survival analysis refers to statistical methods used to investigate the length of time it takes for an event to occur. Using such techniques, researchers can estimate the probability of events occurring over time and assess the impact of variables on the rate at which events occur.

In this thesis, survival analysis is required to estimate the impact of biomarkers on the rate of events following allo-HSCT. However, these techniques are also applicable to non-clinical fields of research (e.g., machine failure in engineering, bankruptcy rates in economics, and customer retention in marketing) (208-210).

The occurrence of a single inevitable event, such as overall survival, can be evaluated using standard survival analysis methods. These include the commonly reported Kaplan-Meier estimator, and the Cox proportional hazards model, which will be referred to as *survival analysis* (211, 212).

When investigating a particular cause of mortality, or non-fatal events, the inference is complicated by the presence of competing risks. Competing risks are present if subjects can experience multiple events (different "causes of failure"), and the onset of one event precludes the observation of other causes of failure.

Allo-HSCT is often used as a canonical example of competing risks, as recipients can experience multiple causes of failure including GvHD and relapse. Mortality from GvHD, will preclude the occurrence of relapse and vice versa.

Survival analysis methodologies which can account for competing risks are therefore essential to make valid inferences in the study of transplant patients. These include estimators of the cumulative incidence, and the Fine-Gray subdistribution hazards models, which will be referred to as *competing risks analysis* (213, 214).

This section introduces some of the fundamental definitions and assumptions upon which survival analysis relies. It explains how the probability of an event can be estimated, and how the impact of variables on the rate of the event can be assessed or controlled for. Followed by an explanation of how these analyses have been adapted for the presence of competing risks.

1.3.1 Survival analysis

1.3.1.1 Censoring

Censoring is when information on time-to-event (e.g., the survival time) is not available for all subjects in a study(215). If this is caused by early dropout from a study, or the lack of an event during the observation period, it is known as right-censored data. For example, in a short study of all-cause mortality some subjects may survive. The survivors are therefore censored at the last time they were observed, as their time of death is unknown.

Right-censored data provides partial information as it encodes how long a subject remained event free, even if their ultimate survival time remains unknown. Other censoring modalities exist (215, 216). However, only right censoring is addressed in this thesis.

Survival analysis methods can account for right-censoring which is essential for valid inference. A key assumption is that the censoring is noninformative. For example, censoring may be informative if subjects who have a negative response to therapy are more likely to drop out of a study early.

1.3.1.2 Survival function

The survival function $S(t)$ describes the probability that a subject will remain event-free beyond a certain point in time, t. Let T be a continuous random variable representing the potential survival time (217). The survival function is defined as

Equation 1

$$S(t) = P(T > t).$$

The survival function can be estimated from observed event (and censoring) times using a non-parametric statistic called the Kaplan-Meier (KM) estimator. This estimator considers time as many discrete intervals, and the product of the survival probabilities in each interval provides the overall estimate.

Let k be the number of events observed in the follow-up period, the event times are ordered such that $t_1 < \ t_2 < \cdots < \ t_k$. The KM estimator is given by the formula

Equation 2

$$\hat{S}(t) = \prod_{i:\, t_i \leq t} \left(1 - \frac{d_i}{n_i}\right).$$

where t_i is a time of an event, d_i is the number of events that occurred by time t_i, and n_i is the number of subjects at risk before time t_i.

A commonly reported summary statistic is the median survival time, the first time at which $S(t) \leq 0.5$, meaning 50% of subjects experienced the event (218). Alternatively, survival probability may be reported for selected times (e.g., 2 years post-transplant).

Two common assumptions are that the $S(0) = 1$ (i.e., no chance of instantaneous events) and that $\lim_{t\to\infty} S(t) = 0$ (i.e., all subjects will experience an event given an infinite time).

1.3.1.3 Hazard function

The hazard function $\lambda(t)$ describes the probability of a subject experiencing an event in the next instance, given they have survived event-free until time t (217). This is closely related to the survival function but instead focuses on the probability of the event occurring rather than non-occurrence.

The hazard function is defined as follows

Equation 3

$$\lambda(t) = \lim_{dt\to 0} \frac{P\{t \leq T < t + dt | T \geq t\}}{dt}$$

where the numerator represents the conditional probability of the event during an interval $[t,t + dt)$, provided they remain event-free at time t, and the denominator is the width of the time interval. Thus, the hazard function $\lambda(t)$ provides a dynamic estimate of the risk of the event over time (219). Whilst the $\lambda(t)$ isn't often directly interpreted, this function forms the basis of proportional hazards models which are commonly reported in survival analysis.

No simple estimator is available for the hazard function $\lambda(t)$ so instead, a quantity called the cumulative hazard rate $\Lambda(t)$ can be assessed. $\Lambda(t)$ is the area under the hazard function $\lambda(t)$, the integral, between times 0 and t. The cumulative hazard $\Lambda(t)$ can be approximated using the non-parametric Nelson-Aalen estimator expressed by the formula

Equation 4

$$\widehat{\Lambda}(t) = \sum_{i:\, t_i \le t} \left(\frac{d_i}{n_i}\right).$$

Where d_i is the number of the events that occurred by time t_i, and n_i is the number of subjects at risk before time t_i. $\Lambda(t)$ can be thought of as the number of events expected per individual by time t, if the event was a repeatable process, and the slope of this curve represents the hazard rate $\lambda(t)$.

There is a close correspondence between the $S(t)$ and $\Lambda(t)$, as illustrated by the KM and Nelson-Aalen estimators relying on the same pieces of information. If any of the $\lambda(t)$, $\Lambda(t)$, or $S(t)$ functions are known the others are automatically determined (220). The relationship between these functions is provided by

Equation 5

$$\Lambda(t) = -\log[S(t)] = \int_0^t \lambda(u)\, du.$$

1.3.1.4 Cox regression

The most widely applied technique to assess the impact of multiple variables on the rate of an event is Cox regression, or the Cox proportional hazards (Cox PH) model (211, 221).

Variables in the model can be variables that are the focus of the study such as candidate biomarkers, as well as confounding covariates which need to be accounted for. In addition to multivariate analyses, the model can incorporate both continuous and discrete variables, and also provides effect size estimates known as hazard ratios (HR).

Cox defined the model in general terms as follows (211). Suppose measurements of variables $z_1, \ldots, z_p$ are available for each individual in a cohort of j subjects. For the j^{th} subject, let the values of z be $z_j = (z_{1j}, \ldots, z_{pj})$. The hazard rate at time t is then given by

Equation 6

$$\lambda(t; z) = exp(z\beta)\, \lambda_0(t).$$

The factor $\lambda_0(t)$ is the baseline hazard function, which represents the underlying hazard for subjects when all covariates are equal to zero (i.e., $z = 0$). This serves as a reference group describing the rate of the event given a standard set of conditions.

This baseline hazard rate is multiplied by $exp(z\beta)$, where β is a vector of unknown regression coefficients estimated for the set of variables z. The factor $exp(z\beta)$, therefore, represents the relative risk, the proportional increase or decrease in the hazard associated with the covariates.

HR are calculated for each covariate by taking the $exp(\hat{\beta})$ estimates from the fitted model. HR correspond to the ratio of the hazard function $\lambda(t)$ between two groups or per unit increase in a continuous variable.

Given the direct relationship between the hazard function and the survival function (**Equation 5**), in the absence of competing risks, the Cox PH model can be used to infer the impact on the rate of survival (217).

A specific relationship is assumed between covariates and the hazard rate, known as the proportional hazards (PH) assumption. This asserts that the effect of covariates on the hazard rate is constant over time. As a result, the HR estimates do not change during the observation period.

1.3.1.5 Follow-up time

The follow-up is the time from the start of the study to the end of the observation period. This needs to be sufficiently long to observe enough of the event of interest. The appropriate follow-up depends on the particular cohort and on the event being studied. In a study of allo-HSCT patients, a follow-up time of 1 year may be sufficient to study AGvHD, an event typically occurs within the first 3-6 months post-transplant, but insufficient to study mortality (77).

As follow-up time may vary between the subjects of a prospective study, the median follow-up time can be used as a summary statistic. The median follow-up time can be estimated using the KM estimator with the outcome indicator reversed (222).

1.3.2 Competing risks analysis

Competing risks are present in studies where subjects can experience multiple events (different "causes of failure"), and the onset of one event precludes the observation of other causes of failure.

For example, one might want to examine the occurrence of relapse following allo-HSCT. In addition to relapse, these subjects are also at risk of non-relapse mortality (NRM). The rate of NRM will impact the number of subjects who survive until relapse, and vice versa. NRM and relapse are therefore competing risks.

In the presence of competing risks, the loss or gain of subjects from the risk set at time t (the number of individuals "at risk" of the event of interest) needs to be accounted for to ensure valid inferences.

1.3.2.1 Cumulative incidence function

The KM and Nelson-Aalen estimators in Equation 2 and Equation 4 can only handle one type of event - individuals either experience the event of interest or are censored. As a result, these methods overestimate the cumulative hazard of events when applied to competing risks (223).

The appropriate method to assess the probability of competing risks is to estimate cumulative incidence functions $CIF(t)$. $CIF(t)$ describe the probability of experiencing the event of interest, before time t, in subjects yet to experience a competing event (217).

For two competing events, j and k, the $CIF(t)$ of event j is estimated in a two-stage process:

1. First, the probability of surviving event-free is estimated at time t_i. This uses the same KM estimator shown in Equation 2, but the event is either of the competing events j and k.
2. Secondly, the Nelson-Aalen estimate of the cumulative hazard of event j is calculated at time t_i with Equation 4.

The $CIF(t)$ for event j is then a function of these two estimates (224). As given by the formula

Equation 7

$$CIF_j(t) = \sum_{i:\, t_i \le t} \left(\frac{j_i}{n_i}\right)\left(\prod_{l:\, t_l \le t_i}\left(1 - \frac{j_l}{n_l}\right)\right)\left(\prod_{l:\, t_l \le t_i}\left(1 - \frac{k_l}{n_l}\right)\right).$$

As a result, $CIF(t)$ estimates factor in both the probability of the event of interest and the probability of competing events.

1.3.2.2 Cause-specific hazards regression

Modified versions of the Cox PH model can be applied in the presence of competing risks, by adapting the definition of the hazard function described in Equation 3.

The first such model is the cause-specific hazards (CSH) model. By censoring subjects who experience competing events, CSH models can estimate covariate effects on the rate of the event of interest but only in individuals who remain event-free. This rate is known as the cause-specific hazard.

The cause specific hazard $\lambda_k^{CS}(t)$ is defined as

Equation 8

$$\lambda_k^{CS}(t) = \lim_{dt \to 0} \frac{P\{t \leq T < t + dt, K = k | T \geq t\}}{dt}$$

where K is a variable indicating the type of event that occurred, and $\lambda_k^{CS}(t)$ is the probability of a subject experiencing the k^{th} event in the next instance, given they have survived event-free until time t (217). Specification of the CSH model is described in detail in Lau et al. 2009 (225).

Separate CSH models are fitted to estimate effect sizes for each competing risk, referred to as cause-specific hazard ratios HR_{cs}. HR_{cs} estimate the difference in the cause-specific hazard of an event between groups, or per 1-unit increase in a continuous variable.

However, effects on the cause-specific hazard do not translate directly into differences in the overall risk of experiencing an event (i.e., the cumulative incidence). As a result, CSH models may be less suitable when investigating the prognostic impact of covariates (217, 225, 226).

Insights such as this motivated the development of models which directly link the effect of covariates to the cumulative incidence function. The most common of which is the subdistribution hazards model, or the Fine-Gray (FG) model.

1.3.2.3 Fine-Gray regression

The FG model is another adaptation of the Cox PH for competing risks which uses Gray's subdistribution hazard technique (214). FG models estimate covariate effects on the rate of the event of interest in individuals who have not experienced the event, or who have experienced a competing event.

The subdistribution hazard function is defined as

Equation 9

$$\lambda_k^{SD}(t) = \lim_{dt \to 0} \frac{P\{t < T \leq t + dt, K = k | T > t \cup (T < t \cap K \neq k)\}}{dt}.$$

where $\lambda_k^{SD}(t)$ is the probability of a subject experiencing the k^{th} event in the next instance, given they have yet to experience the k^{th} event by time t. Note, individuals who have experienced a competing risk are also included in the risk set (214). Specification of the FG model is described in detail in Lau et al 2009 (225).

Separate FG models are fitted to estimate effect sizes for each competing risk, referred to as subdistribution hazard ratios HR_{SD}. These estimate the differences in the subdistribution hazard between groups, or per 1-unit increase in a continuous variable.

As the subdistribution hazard resists an intuitive interpretation, researchers often describe results from FG models in terms of the impact covariates had on the cumulative incidence function $CIF(t)$.

Both the direction and significance of the HR_{SD} equate to the equivalent change in the cumulative incidence of that event. However, the exact magnitude of the effect cannot be stated.

Both the cause-specific hazards models and the FG models therefore account for competing risks but estimate covariate effects on different hazard functions. CSH models are suitable for "studying the etiology of diseases" (225), as they estimate differences in the rate of the event for individuals who remain at risk of the event. Whereas the FG model is purportedly more suitable for the development of predictive/prognostic models (225, 226) or for investigating the impact of covariates on prognosis (217).

Survival and competing risk analysis provide a suite of rigorous statistical tools to test the impact of biomarkers on prognosis following allo-HSCT. Data science refers to the extraction of knowledge from complex data (227, 228). Survival analysis is one example of how data science can be used to facilitate risk stratification, prediction of clinical endpoints, and the identification of prognostic biomarkers in allo-HSCT research.

1.4 Aims

The key aims of this thesis are to:

- Describe the cellular composition of PBSC grafts using single-cell mass cytometry (CyTOF) data.
- Determine if distinct types of PBSC graft can be defined based on variation in cellular composition.
- Investigate associations between PBSC graft composition and prognosis following allo-HSCT.
- Investigate associations between serum proteins post-transplant and prognosis following allo-HSCT.

The clinical endpoints to be studied include AGvHD (grade II-IV), CGvHD, relapse, GRFS, TRM/NRM, RM and OS.

1.5 COVID-19 secondment

The development of this thesis coincided with the onset of the pandemic. As a result, I spent an extended period of time on secondment to COVID-19 research.

I collaborated with colleagues to develop and externally validate prognostic models for COVID-19 (229). This study focused on the use of clinical data collected on the day of admission from three substantial cohorts of COVID-19 patients.

The model development process relied on variable selection algorithms to identify COVID-19 risk factors. Subsequently, several of these algorithms were applied to identify prognostic biomarkers for allo-HSCT recipients in **Chapter 4** of this thesis (**Section 2.2.4.2**).

Moreover, we realised that the methods used to stratify allo-HSCT grafts by risk in **Chapter 3** were equally applicable to these cohorts of COVID-19 patients. In particular, we showed that the use of Uniform Manifold Approximation and Projection (**Section 2.1.7.2**) and model-based clustering (**Section 2.1.6.2**) could provide a way to identify novel risk groups (230).

A paper published on this topic as part of my Ph.D. and is integrated into this thesis as **Appendix A**.

2 Methods

2.1 Allo-HSCT graft composition study

2.1.1 Patient and donor cohorts

Francesca Kinsella, with whom I collaborated on this project, recruited a cohort of allo-HSCT graft donors and recipients. Patients gave written informed consent using the BOOST ethics approved by the regional ethics committee (05/Q2707/175).

In total, 185 consecutive donors were recruited between 2009 and 2017 (donor cohort) (**Table 3.1**). Donor bags were obtained post-stem-cell infusion. Samples were collected from the waste apheresis product for single-cell mass cytometry (**Table 3.2, Table 3.3**).

A prospective cohort of 155 allo-HSCT recipients were recruited between 2009 and 2017 and followed-up for a median of 3.7 years (CI 3.31-4.9). Patients in this cohort were paired with one of the grafts from the donor cohort. Information on the time to the onset of AGvHD (grade II-IV), CGvHD, relapse, and mortality was available.

The information available for the donor cohort included: age, sex, CMV status, relationship to recipient, stem cell source, and CD34+ dose, with some missing observations (**Table 3.1**). The total mononuclear cell dose (MNC) dose, and T cell dose were also available for a subset of donors.

The information available for the patient cohort included: age (at the time of transplant), sex, CMV status, HLA mismatch status, diagnosis, and HCT-CI (**Table 3.4**). The intensity conditioning regimen (MAC / RIC), whether they received TBI (yes / no), and in-vivo T cell depletion (ATG / alemtuzumab) were also known. As was the provision of GvHD prophylaxis which was either cyclosporine or a combination of cyclosporine and methotrexate or mycophenolate mofetil (cyclosporine & other).

For brevity, I will refer to these patient and donor characteristics collectively as the *clinical variables*.

2.1.2 Clinical end points

The primary survival endpoints studied were OS and GRFS. The primary competing risk endpoints investigated were transplant-related mortality (TRM) and RM. The cause of death was adjudicated as TRM or RM by Francesca Kinsella. The onset non-fatal competing risks relapse, AGvHD (grade II-IV) and CGvHD, were analysed as secondary endpoints. The competing risk for relapse was TRM, whereas death (any other cause) was the competing risk for other outcomes. A maximum follow-up time of 2-years was selected, after which all events were censored. Definitions of these endpoints can be found in **Section 1.2.5**.

2.1.3 CyTOF data

Cytometry is a powerful tool with many applications in immunology, haematology, and cancer biology. It enables rapid measurement of multiple physical and chemical characteristics of individual cells (231). This approach has been key to identifying and defining many immune cell subsets (232).

Cytometry by time-of-flight (CyTOF), or single-cell mass cytometry, can detect over 40 cell surface markers per cell (233-235). These devices rely on antibodies conjugated with heavy metal ions, which bind to specific cell surface markers. Ions are then detected using time-of-flight mass spectrometry which can indicate the presence or absence of particular markers on individual cells (236). This information can then be used to infer the phenotypic characteristics of each cell (237).

The 185 graft samples were subjected to CyTOF by Francesca Kinsella. CyTOF was conducted using a panel of 37 markers which included most major immune lineage-defining markers in combination with numerous phenotypic and functional markers (**Table 3.2**).

The markers were divided into *lineage markers*, these are relatively stable features consistently associated with particular cell types. In contrast, the remaining markers primarily give information on transient cell states, *cell state markers*. Whilst in practice there is overlap between these two groups of markers, the lineage markers were used when defining the cellular populations.

CyTOF data acquisition and pre-processing were conducted by Francesca Kinsella, whereas I conducted the data analysis. The raw CyTOF data was exported to an FCS file per graft. In pre-processing, FCS files were de-barcoded and consecutive gates were applied to remove non-biological events, doublets and debris.

2.1.4 Data pre-processing

Analysis was conducted using R version 4.0.2. (238). FCS files were imported and manipulated using the R packages "flowCore" (239) and "data.table" (240).

The CyTOF data contained 4,441,335 events in total, with a median of 20,467 events per FCS file (IQR 9633 – 32725). The raw marker intensities in CyTOF data have strongly skewed distributions, with varying ranges of expression. The arcsin transform with a cofactor of five was applied, in line with standard practice (234, 241). Principal component analysis was applied to visualise FCS files by batch (242). Visualisation of graft centroids – the average marker expression across all markers – did not indicate strong batch effects (**Figure C.1**).

2.1.5 Cluster analysis for CyTOF data

For each event in CyTOF data, the intensity of marker expression describes the phenotypic characteristics of a single cell detected by the mass spectrometer (event = cell). With this information, events can be classified into groups of similar cells or assigned to a known cell subset.

Historically, the gold standard method to identify cell subsets has been "manual gating" in which a series of 2-D bi-axial plots are compared to manually divide cells using prior knowledge. This hypothesis driven approach can be applied to any sample to accurately classify cells into known cell subsets. However, the sequence of markers on which to gate cells, the cut-off for each marker and the expected phenotypes are determined by the researcher.

As a result, manual gating is highly subjective and hard to reproduce between experiments. The time and labour required for each sample means this approach does not scale to larger cohorts. In practice, not all combinations of markers will be compared so information is lost and rare or unexpected cell populations may be overlooked (243).

Cluster analysis can overcome some of the limitations of manual gating. Cluster analysis encompasses wide variety of algorithms designed to partition data into groups of similar observations (clusters). These methods segment observations based on patterns in the data without the use of prior labels (244). This has a remarkably wide range of applications in medical science, engineering, the humanities, and day to day life (244).

Performance comparisons have demonstrated the ability of the PhenoGraph and FlowSOM clustering algorithms to identify cell subsets in CyTOF data which closely match those of expert manual gating (234, 245-247). Moreover, they have extremely fast run times so are suitable for interactive analysis (246). Liu et al. 2019 provide a comparison framework to guide the choice of clustering method when analysing CyTOF data (247).

2.1.5.1 PhenoGraph algorithm

PhenoGraph takes as input a matrix of N individual cells and partitions the data into subpopulations corresponding to the cellular phenotypes present in the data.

To do this, the algorithm relies on graphs (or networks), mathematical structures which model pairwise relationships between observations. A nearest-neighbour graph represents each observation as a node, with edges connecting the most similar observations – the k-nearest neighbours. Methods to construct such graphs are amongst the most influential in data mining (248).

Briefly, the PhenoGraph algorithm has two main stages: 1. Construct a nearest-neighbours graph where N cells are represented as nodes with edges connecting the most similar cells (the k-nearest neighbours). 2. Partition the graph into distinct clusters using Louvain community detection, such that cells with a similar phenotype are clustered together (245, 249).

Community detection is extremely efficient with very large data sets since this approach was developed for the analysis of large mobile phone and social networks (249, 250). The main computational bottleneck is the construction of the nearest-neighbour graph, which is overcome by efficient algorithms such as using approximate nearest neighbours (251, 252).

Community detection makes no assumptions about the number, size or distribution of clusters (253). As a result, PhenoGraph infers the number of clusters directly from the data. The inability to pre-determine the number of clusters can, however, result in a large number of clusters, which may not all be biologically meaningful (234). The resulting clusters also need to be labelled manually and can be hard to interpret if phenotypes do not conform to expected cell subsets.

2.1.5.2 FlowSOM algorithm

FlowSOM utilises the self-organising map algorithm to partition N individual cells into subpopulations (254, 255).

The FlowSOM workflow has three main computational steps: 1. Construct a self-organising map with a grid of 100 nodes (by default) representing patterns of expression in randomly sampled cells. N cells are then assigned to the most similar node. 2. Build a minimal spanning tree graph to visualise nodes (256, 257). 3. Perform consensus hierarchical clustering to determine a final cluster assignment (258).

FlowSOM has a fast runtime and in-built visualisation tools which make it easier to determine the phenotype of detected clusters. One potential draw-back is that FlowSom requires the number of clusters to be predetermined which is difficult in exploratory experiments where the diversity and number of cell types is unknown (247). FlowSom provides the option to estimate the number of clusters from the data, however, this has been shown to negatively impact the performance (247).

2.1.5.3 Implementation

PhenoGraph was utilised in this study to avoid pre-determination of the number of clusters. PhenoGraph clustering was implemented using the "Rphenoannoy" R package (245, 259). The only user defined parameter, k, corresponds to the number of nearest neighbours in the first stage of the algorithm. PhenoGraph is reportedly robust to a wide range of k values, therefore the default value of 30 was utilised (245).

To facilitate immediate comparison across samples, PhenoGraph was run on all samples pooled into a single matrix of 4,441,335 cells. Phenotypic similarity was defined using only the 23 lineage markers. As a result, PhenoGraph received a 4,441,335 x 23 matrix as input with k set to 30 nearest neighbours.

The output, a 4,441,335 x 1 matrix of cluster assignments, was appended to the input matrix for visualisation and onward analysis.

The number of cells assigned to each cluster was summarised along with the number of grafts in which clusters were observed. Clusters observed in less than 50% of grafts were considered likely to be

artifacts in the structure of the data, or present in insufficient numbers for subsequent analysis, so were excluded.

2.1.6 Cluster analysis for compositional data: identifying distinct PBSC graft groups

One of the aims of this thesis was to assess if variation in graft composition between donors constituted distinct types of PBSC graft.

To enable a comparison of cellular composition between grafts, cell counts from the CyTOF data were normalised to proportions. The total cell count of cells N from each graft was split into D distinct cell lineages (parts) and transformed into a composition by dividing by the total

Equation 10

$$(x_1, \dots, x_D) = \frac{N_1, \dots, N_D}{N_1 + \cdots + N_D}.$$

Cluster analysis was applied to estimate an unknown number of groups based on differences in graft composition. A complication was that the cell type abundance data generated by CyTOF are compositional in nature (260), as are many types of count based omic data (261-263).

2.1.6.1 Compositional data

Compositional data are defined as data that represent the proportions of some whole (264). Examples include proportions, percentages, and probabilities – these non-negative numbers provide relative information rather than absolute values.

Compositional data pose statistical challenges because their sum is constrained to a constant value c, such as 1 for proportions. This "constant-sum" constraint results in complicated correlations between variables and makes compositional data incompatible with classical statistical methods such as t-tests, ANOVAs and multivariate analyses including clustering and regression (263, 265, 266).

This led to the development of compositional data analysis by Aitchinson (264). The main idea of this approach is the use of log-ratio transformations to eliminate the constant-sum constraint, thereby enabling the use of the standard suite of statistical tools (267, 268).

One such transformation is the additive log-ratio (ALR) (264). The ALR of a composition, x, with D parts, is given by the formula

Equation 11

$$ALR(x) = [y_1, \dots, y_{D-1}] = \left[\log\left(\frac{x_1}{x_D}\right), \dots, \log\left(\frac{x_{D-1}}{x_D}\right)\right].$$

One part of the composition is selected as the denominator, or *reference*, resulting in $D - 1$ log-ratios.

After the transformation of the compositional data into log-ratios, the analysis precedes in the usual fashion, taking into account the interpretation of effects in terms of ratios. For example, the log-ratios can be included as predictors in a regression model and additional non-compositional variables can be included to adjust for confounding effects (269).

Other types of log-ratio transformations have been proposed, such as the centred log-ratio (CLR) and the isometric log-ratio (ILR). Each of these has its advantages and disadvantages (268, 270). The CLR and ILR were developed to transform compositions into log-ratios whilst preserving the original geometry of the data, the isometry. This is particularly useful for distance-sensitive analyses such as dimension reduction (263, 271) and clustering (263, 272, 273). A disadvantage of CLR and ILR is that

the interpretation of individual variables is complicated, as they involve the ratio between several parts of the composition.

A variety of standard clustering algorithms have shown to be effective with log-ratio transformed compositional data, in particular using the ILR transformation. These include the widely used *k*-means (273) algorithm and model-based clustering with the Gaussian mixture model (GMM clustering) (272, 274). The GMM clustering algorithm was utilised in this study as it has been shown to outperform *k*-means on compositional data (272, 275).

2.1.6.2 GMM clustering

GMM clustering is a popular algorithm for clustering continuous data, that has been extended for use with compositional data (272, 274) . GMM clustering is based on a finite mixture model that presumes that data is generated from a mixture of K distinct mixture components, one for each cluster (276).

The GMM clustering algorithm makes the following assumptions: (1) the number of clusters, K, is known in advance. (2) Each data point has a particular probability of being associated with each cluster. (3) The clusters follow a multivariate Gaussian distribution.

Given these assumptions, model parameters describing the Gaussian distributions need to be estimated along with the associated probabilities for each data point.

The number of clusters, K, needs to be specified. The GMM algorithm benefits from a formalised statistical framework which allows for the appropriate number of to be inferred from the data using the Bayesian information criterion (BIC) (276, 277). BIC compares the number of parameters estimated in the model k and the number of events n, with the maximum value of the log-likelihood function $\hat{L}$, using the formula

Equation 12

$$BIC = -2\log(\hat{L}) + \log(n) \cdot k.$$

The lower the BIC value, the better the perceived quality of the model, provided each model was fitted to the same set of observations. Alternative approaches to K selection include the silhouette score (described in **Section 2.1.6.4**) and the GAP statistic (278), which compares intra-cluster variation with the expected intra-cluster variation of an appropriate null reference distribution.

2.1.6.3 Implementation

Cluster analysis for compositional data (clustCoDa) was conducted using the R package "robCompositions" (279) (272). Only grafts from donors paired with a recipient in the patient cohort were included (N = 155).

Compositions were calculated for each graft using the formula in **Equation 10**. Compositions were divided into D parts corresponding to the number of general cell lineages (e.g., T cells, NK cells, monocytes, … etc). The characterisation of clusters into these cell lineages is described in **Section 2.1.8**.

As input clustCoDa received a 155 x D matrix of compositions and was run with the default clustering algorithm and transformation, GMM clustering and ILR transformation, respectively (274).

clustCoDa was run iteratively with K ranging from 2-10 and the GMM with the lowest BIC value was selected. Using the final selected GMM, each graft was assigned to the cluster with the highest probability of generating the observed composition.

The output from clustCoDa was a 155 x 1 matrix of compositional cluster assignments. To avoid confusion with cell populations, these compositional clusters will be referred to as *PBSC graft groups*.

2.1.6.4 Cluster evaluation criteria

Clustering algorithms may partition a data set whether or not distinct groups are present, it is therefore useful to assess the validity of the resulting clusters.

External evaluation metrics compare the accuracy of cluster assignments with a pre-determined set of classifications (244). Common examples include the Rand index (280), adjusted Rand index (281), and normalised mutual information (282). These metrics were recently applied to test the precision of PhenoGraph clustering compared to manual gated cell populations (247). However, external evaluation is only suitable when the cluster labels are known in advance.

In contrast, internal validation metrics don't require any external information to assess clustering quality, instead they rely solely on the data used as input for clustering. In general, these utilise two measures to assess clustering quality, cohesion and separation. Cohesion quantifies the similarity of observations within a cluster (an intra-cluster metric). Whereas separation measures the dissimilarity of observations between clusters (an inter-cluster metric).

A proximity function such as a distance metric is required to define the similarity between pairs of observations. Clustering is considered good quality when there is high separation between clusters and high cohesion within clusters (283). A variety of metrics have been developed which combine cohesion and separation into a single measure. Examples include the Silhouette width (284-286), Dunn index, and the Davies-Bouldin index, each of which is described in detail in Zhang et al. 2023 (287).

The value of the Silhouette width ranges from -1 to 1, with positive values indicating a better cluster partitions and negative values indicating the clusters are mixed together (288). As well as for internal validation, comparing the Silhouette width can be used to determine K in model fitting. One of the main draw backs of calculating the Silhouette width is the high computational complexity, rendering this approach impractical for very large data sets.

Silhouette widths were calculated for the PBSC graft groups using the R package "robCompositions" (279), and positive silhouettes were interpreted as a valid cluster configuration. Bar plots and box plots were generated to compare donor factors and graft composition between each of the identified PBSC graft groups. Plots were generated using the R package "ggplot2" (289).

Whilst Silhouette can assess the validity of clusters, it could be argued that clusters should be evaluated based on the usefulness of the result (290, 291). For example, by testing if the PBSC graft groups can be used to stratify recipients by risk. Therefore, the prognosis of recipients was compared by PBSC graft group (**Section 2.1.10**).

2.1.7 Visualisation using dimension reduction

Dimension reduction (DR) provides an intuitive way to visualise complex data sets in 2-D or 3-D. DR transforms high-dimensional data into a smaller number of latent features, referred to as an embedding. This lower-dimensional representation of the original data is optimised to retain as much of the statistical information as possible (292).

DR was applied to the CyTOF data for several purposes: to visually identify distinct cell subpopulations, to confirm if subpopulations were delineated effectively by PhenoGraph, and to compare the relative intensity of markers expression between subpopulations.

Three DR algorithms were utilised, t-distributed stochastic neighbour embedding (t-SNE) (293, 294), uniform manifold approximation and projection (UMAP) (295), and potential of heat diffusion for affinity-based transition embedding (PHATE) (296, 297).

2.1.7.1 t-SNE

t-SNE, published in 2008, is the most widely used method for visualising single-cell analyses such as CyTOF (298, 299). Whilst not suitable for general purpose dimension reduction, it produces intuitive visualisations of single cell data.

t-SNE was implemented using the "Rtsne" R package (300). The computational and memory requirements of t-SNE scale quadratically with the number of data points (293) making it inefficient for visualising very large-scale data sets (301). Therefore, t-SNE was run on a subsample of events to reduce the computational time, using the 23 lineage markers as the input dimensions.

t-SNE received a 1,000,000 x 23 matrix as input, and was run with the default parameters, the maximum number of iterations was increased to 30,000 to ensure convergence.

2.1.7.2 UMAP

UMAP, a relatively new algorithm published in 2018, has been widely adopted for visualising single-cell analyses (299). This has been used to gain insight into the differentiation of hemopoietic cells (302) and how they interact with transcription factors (303), and to assess the impact of T cell receptor repertoires on checkpoint-blockade therapy (304).

As well as for visualisation, UMAP can be used for general-purpose dimension reduction. This has led to its use as a pre-processing step in the development of prognostic models (305) (**Appendix A**), and to quantify minimal residual disease in haematological malignancies (306).

UMAP was implemented using the "UWOT" R package (307). The number of nearest neighbours, k, is a parameter that controls the balance between local and global data structure, with a larger value assessing a larger neighbourhood of observations.

UMAP was run on all events using the 23 lineage markers as the input dimensions. UMAP received a 4,441,335 x 23 matrix as input and was run with default parameters. This was repeated with values of k from 15 – 50, and k = 30 was selected as it produced a favourable visualisation.

2.1.7.3 PHATE

UMAP and t-SNE are optimised primarily to retain the local data structure. As a result, they can fail to preserve meaningful distances between observations on the embedding (308). For example, data that has a continuous or hierarchical structure (e.g., the continuous differentiation of cell types) might appear as distinct "clusters" on a t-SNE or UMAP embedding.

PHATE, a recently published visualisation algorithm from 2020 (296, 297), was developed to visualise such data structures. PHATE reduces dimensionality to two or three dimensions whilst seeking to preserve meaningful distances between observations (308).

As a result, PHATE captures trajectory structures between observations as well as preserving the intrinsic geometry (297). This helps to identify branching or endpoints in the data structure, for example, cell differentiation (296). This has found use in studies of differentiation pathways in haematopoietic stem cells (309) and T cells (310).

PHATE was implemented using the "phateR" R package (311). Like UMAP, PHATE has a parameter, knn, which controls the number of nearest neighbours.

PHATE was run on all events using the 23 lineage markers as the input dimensions. PHATE received a 4,441,335 x 23 matrix as input and was run with the default parameters. This was repeated with values of knn from 5 – 50, and k = 15 was selected as it produced a favourable visualisation.

2.1.7.4 2-D visualisation

The embeddings output by t-SNE, UMAP and PHATE consisted of n x 2 matrices, with n equal to the number of cells. 2-D plots were generated from these embeddings using the R package "ggplot2"(289). These were utilised to infer the presence of subpopulations of cells, to provide an intuitive visualisation of the clustering results, and to help characterise the phenotype associated with each cluster (see next heading).

2.1.8 Characterisation of cluster phenotypes

To be interpretable, the defining features of clusters need to be characterised. This presents a challenge when applying cluster analysis to CyTOF data. The phenotype associated with each cluster was assessed systematically, using a combination of four complimentary characterisation methods: mean marker intensity, marker enrichment modelling, dimension reduction and marker intensity visualisation.

2.1.8.1 Mean marker intensity

The defining characteristics of a cell population are often assumed to be the markers that are most differentially or highly expressed in each cluster (243). These can be identified by comparing statistics such as the average marker intensity (243).

Each cluster was represented by its mean marker intensity, a 37-dimensional vector computed by the mean of each marker across all cells in that cluster. A threshold at which to consider a cluster positive for a marker was determined by visualising the distribution of marker intensities across randomly sampled cells using density plots generated by the R package "ggplot2"(289).

Mean marker intensities indicate the defining characteristics of a cluster thereby facilitating phenotype identification. However, this approach does not take into account variability within and between clusters and relies on subjective thresholds.

2.1.8.2 Marker enrichment modelling

Statistical metrics such as marker enrichment modelling (MEM) (312) use effect size estimates to infer which markers are enriched or downregulated relative to other clusters. MEM calculates the level of marker enrichment by cluster compared to a reference group (e.g., all other clusters), whilst taking into account the within and between cluster variability.

MEM was conducted using Cohen's *d* (313, 314) given by the formula

Equation 13

$$Cohen's\ d = \frac{\bar{X}_1 - \overline{X_2}}{\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}}$$

where the numerator is the difference between the means of two groups with n cells, and the denominator is the pooled standard deviation. Cohen's D corresponds to small (*d* = ±0.2), medium (*d* = ±0.5), and large (*d* ≥ ±0.8) effects (313, 314).

Cohen's D was calculated for each marker by cluster using the R package "esci" (315). Effects ≥ ±0.5 were considered evidence of either marker enrichment or downregulation. These were then visualised as a heatmap using the R package "ggplot2" (289).

2.1.8.3 Dimension reduction

DR is a powerful tool to interrogate the association of clusters with marker expression by enabling the visualisation of high-dimensional data. t-SNE (293, 294) and UMAP (295) are both well suited to this task. Provided a dataset contains distinct sub-populations, t-SNE and UMAP visualisation will produce isolated "clusters" that often correspond with labels assigned by a clustering algorithm (316). Note, these algorithms help to visualise clusters but do not explicitly partition observations into groups.

When visualised in 2-D each cell can be coloured by factors such as marker intensity, cluster assignment, and suspected cell type. If areas of high marker expression colocalise with the distribution of clusters on the embedding, it can reveal relative marker enrichment. Additionally, clusters which colocalise are likely to have similar defining characteristics which provides clues as to the identity of clusters that are difficult to ascertain by other methods.

2.1.8.4 Marker intensity visualisation

Typically, learning the phenotype or cell type associated with a cluster relies on human experts (243, 312). Violin plots and bi-axial plots can be used to compare the expression of markers corresponding to well-characterised types of cell types to determine the likely phenotype.

This is time-consuming when the number of clusters is prohibitively large, and biases results toward expected phenotypes. Therefore, marker intensity visualisation was utilised to confirm phenotypes indicated by the other characterisation methods.

Violin plots and bi-axial plots were generated using the R package "ggplot2" (289). The intensity of marker expression in each cluster of cells was compared to a comparison group which consisted of cells sampled at random from all other clusters.

2.1.8.5 Assigning clusters to a cell lineage and cell subset

The results of each cluster characterisation methods were compared. If a consensus was reached by the various methods, it provided additional confidence that the assigned phenotype was a reasonable description of the clusters defining characteristics.

PHATE visualisation (**Section 2.1.7.3**) of cell trajectories can indicate differential pathways. PHATE was utilised to ensure the characterisation of cluster phenotypes corresponded with the biological understanding of how the cell types observed differentiate.

Clusters were characterised into cell types hierarchically. At the highest level, each cluster was classified into general cell lineages (CL). These included: B lymphocytes (B cells), dendritic cells (DCs), natural killer cells (NK cells), monocytes and T lymphocytes (T cells), and hemopoietic stem cells (HSCs).

The markers used to define each of these lineages were CD19 (B cells), HLA-DR (DCs), CD56 (NK cells), CD14 (monocytes) and CD3 (T cells). Clusters positive for at least one of these markers were considered lineage positive (Lin^+). Cells that were lineage negative (Lin^-) but positive for CD34 were labelled as hemopoietic stem cells (HSCs), whereas if no lineage could be assigned, they were labelled unclear.

Once assigned to a general CL, clusters were further classified into finer-grained cell subsets (CS) based on lineage-specific characteristics and the phenotypes identified in the data. The cell type of each cluster was therefore defined by its general CL and a more detailed CS.

2.1.9 Summarising cellular composition of PBSC grafts

Using the results of the preceding analysis, the cellular composition of PBSC grafts could be described. The proportion of cells assigned to each CL was calculated to summarise the total cellular composition of the CyTOF data. The total MNC, T cell, and CD34+ doses were also compared for a subset of grafts for which the data was available.

Cells counts were standardised to proportions using **Equation 10** to enable comparison between graft donations. Median and interquartile ranges (IQR) were used to summarise average graft composition and how much this varied between donors.

2.1.10 Prognosis by PBSC graft group

Survival and competing risks analysis were utilised to test if the PBSC graft groups identified by clustCoDa (**Section 2.1.6**) were associated with significant differences in patient prognosis.

Survival endpoints were studied with KM estimates and Cox PH models using the R package "Survival" (317). Competing events were analysed with CIF estimates and FG models using the R packages "cmprsk" (214, 318) and "riskRegression" (319).

The log-rank test (320) was used to compare KM estimates between groups, whereas Gray's test (213) was used for CIF estimates. The KM and CIF estimates were visualised with the R package "survminer" (321). Cox PH and FG models were fitted to estimate univariate hazard ratios (HR) and subdistribution hazard ratios (HR_{SD}) between groups, respectively.

To test if differences in prognosis observed by PBSC graft group were apparent after adjusting for potential confounders, a multivariate analysis was conducted.

2.1.10.1 Missing data imputation

Multivariate analysis was complicated by the presence of missing observations in the clinical variables (**Table 3.1, Table 3.4**).

Missing observations are a pervasive problem in prospective studies and multiple imputation has become the established tool to minimise the impact of this issue (148, 322-324). The missing observations are imputed (assigned a value) stochastically m times. Each of the m complete data sets is analysed and the m results are combined using Rubin's rules (325, 326).

This relies on the assumption that the distribution of the observed and missing values have similar values (missing completely at random), or if there are systematic differences, that these can be explained by other observed values (missing at random) (327). An example of missing at random might be if HCT-CI risk scores are less likely to be recorded for younger patients with fewer comorbidities. Provided age is in the imputation model, this sort of missingness can be accounted for.

Given these assumptions are met multiple imputation has many benefits. It allows for the analysis of all available observations to preserve sample size, and the use of standard complete data methods. However, if systematic differences are present that can't be explained by the available information (missing not at random), multiple imputation can result in biased estimates and incorrect inferences. Moreover, including variables in the imputation model that are unavailable at the time of model evaluation could lead to information leakage.

Missing observations in the clinical variables were quantified and imputed from the observed data, under the assumption it was missing at random or missing completely at random (326). Multivariate imputation by chained equations (MICE) (328, 329) was applied to produce 5 imputations of the missing observations using the R package "mice" (329). Predictive mean matching was used for continuous variables, logistic regression for binary variables, and polytomous regression for ordinal variables (329, 330).

In each case, all clinical variables were included in the imputation model. Variables that were not observable at the time of model evaluation (e.g., survival status at follow-up) were excluded to avoid information leakage.

2.1.10.2 Survival and competing risks regression for PBSC graft groups

Models were fitted to each imputation and coefficients were pooled into a single estimate by applying Rubin's rules (325). All multivariate models were adjusted for the clinical variables.

Variables for which the proportional hazards (PH) assumption was not satisfied were excluded; a PH test based on Schoenfeld residuals was used for Cox PH models (331) and a PH test based on the cumulative sum of residuals was used for Fine-Gray models (332).

The resulting models were interpreted to assess if recipients of grafts from alternative PBSC graft groups had significant differences in prognosis, adjusting for patient & donor characteristics.

2.1.11 Prognosis by PBSC graft composition

The relationships between individual lineages and patient outcomes were investigated to address the third aim of this thesis – to assess if variation in graft composition was associated with patient outcomes.

Compositional data analysis has been utilised for a variety of regression models including linear, logistic, and Cox regression (333-335), and is widely adopted in the study of microbiomes for which data is often compositional in nature (262).

A formulation of Cox PH regression compatible with compositional data was utilised to test the effect of composition on survival outcomes (268, 335). This was extended to the Fine-Gray model to study competing events.

2.1.11.1 Survival and competing risks regression for compositional data

Compositions were calculated for each graft using the formula in **Equation 10**, dividing grafts into D parts (e.g., T cells, NK cells, monocytes, … etc). Compositions were then transformed with the ALR transformation shown in **Equation 11** (264).

The proportion of HSCs was used as the reference part for the ALR transformation. It has been demonstrated that the same inference should result irrespective of the choice of reference (264), Aitchison (1986, Chapter 5). The impact of the choice of reference part was assessed by calculating a statistic called the Procrustes correlation (**Table C.1**) (262). This tests how well the transformed data represents the exact log-ratio geometry.

ALR transformation and Cox PH model fitting were implemented using the R package "epicoda" (334, 336). FG model fitting was implemented using a customised function based on code from the R packages "epicoda" (334, 336), "cmprsk" (214, 318) and "riskRegression"(319) which is available via GitHub (337).

For the survival endpoints, multivariate Cox PH models were fitted with ALR transformed compositional covariates. For the competing risk endpoints, multivariate FG models were fitted with

ALR transformed compositional covariates. All multivariate models were adjusted for the clinical variables, and missing observations were handled by multiple imputation, in the manner described in **Section 2.1.10.1**.

The resulting models were used to assess the difference in prognosis associated with an increase in one part of the composition (e.g., T cells) relative to the proportion of HSC. A likelihood-ratio test was applied to test the significance of the entire composition. These comparisons were used to test the association of graft composition with prognosis, thereby addressing the third aim of the thesis.

2.2 Post-transplant serum protein study

2.2.1 Patient cohort and proteomics data

Kriti Verma, with whom I collaborated on this project along with Wayne Croft, recruited a prospective cohort of 154 allo-HSCT recipients between 2009 and 2017. On the 14th day post-transplant, Kriti Verma collected serum samples and conducted proteomics analysis.

Protein concentration was measured using the Olink inflammation panel (Olink Proteomics AB, Uppsala, Sweden) (338) which includes 92 protein targets. Assay readout is measured as Normalized Protein eXpression (NPX), arbitrary units on a log2 scale.

A subset of 145 individuals within the cohort were followed-up for a median of 3 years (CI 1.98-3.9). Information on the time to the onset of AGvHD (grade II-IV), CGvHD, relapse, and mortality was available.

Data available on patient & donor characteristics included: the age of the patient and donor, sex of the recipient and donor (male/female), CD34+ cell dose, diagnosis (lymphoid or myeloid), HLA mismatch status (match or mismatch), HCT-CI score (0, 1, 2, or 3+), total body irradiation (TBI, yes or no), reduced intensity conditioning (RIC, yes/no), and GvHD prophylaxis (ciclosporin or ciclosporin & mycophenolate or methotrexate (combination)). For brevity, I will refer to these patient and donor characteristics collectively as the *clinical variables*.

2.2.2 Differential expression analysis

Wayne Croft applied differential protein expression analysis by Mann-Whitney U test with the Benjamini Hochberg (BH) method to control for false discovery rate (339).

Differential expression analysis compared patients who experienced an outcome (AGvHD, CGvHD, relapse and mortality) at any time during the follow-up period with those that were censored. This identified a subset of 19 proteins that were differentially expressed between outcome groups. This subset of 19 proteins were considered candidate prognostic biomarkers for allo-HSCT recipients. I will refer to these collectively as the *protein variables*.

Using the clinical and protein variables, I addressed the fourth aim of this thesis – to test the association between post-transplant serum proteins and recipient prognosis.

2.2.3 Clinical endpoints

The primary survival endpoints analysed were OS and GRFS. The primary competing risk endpoints investigated were NRM and RM. The onset non-fatal competing risks relapse, AGvHD (grade II-IV) and CGvHD, were analysed as secondary endpoints. The competing risk for relapse and RM was NRM, whereas death (any other cause) was the competing risk for other outcomes. Individuals that did not experience an event were censored at the final follow-up.

2.2.4 Survival and competing risks analysis

Survival and competing risks analysis was utilised to test if the protein variables were associated with significant differences in patient prognosis. Serum NPX data were imported and analysed using R version 4.0.2. (238). NPX values were standardised to have a mean of 0 and SD of 1. Individuals with missing outcome information were excluded (N. excluded = 9).

Survival endpoints were studied with KM estimates and CoxPH models using the R package "Survival" (317). Competing events were analysed with CIF estimates and FG models using the R packages "cmprsk" (214, 318) and "riskRegression" (319).

2.2.4.1 Missing data

Missing observations were present in the clinical variables. Multiple imputation was utilised to impute missing values as described in **Section 2.1.10.1** under the assumption that the data mechanism was missing at random or missing completely at random. All clinical and protein variables were included in the imputation models.

2.2.4.2 Variable selection

The number of clinical and protein variables was large relative to the number of observations, therefore the data dimension needed to be reduced prior to multivariate modelling. The most important strategy is to use clinical knowledge and previous literature to guide the choice of variables (340). In an exploratory analysis such as this study, where there is limited prior knowledge, variable selection algorithms can be used to guide choice.

Variable selection is a machine learning technique to select relevant variables that accurately describe a given problem, whilst discarding irrelevant or redundant features with a minimum reduction in performance (341). Only a few variable selection algorithms have been extended to survival data, these include the least absolute shrinkage and selection operator (LASSO) (342-344); stepwise selection with the Akaike information criterion (AIC selection) (345, 346); and random survival forests (RSF) (347, 348) .

Each of these methods has been utilised in biomarker discovery in allo-HSCT research (349-351) and applied to Olink proteomics data (352, 353). They are also suitable for variable selection in the presence of competing risks using Gray's subdistribution hazards method (354-356).

A general introduction to variable selection and the function of each of the aforementioned algorithms is included in **Appendix B** of this thesis.

LASSO was implemented using the R packages 'mass' and 'crrp' (354, 357). For stepwise AIC selection, the R package "pec" was utilised (358). Whereas RSF was implemented using the 'randomForestSRC' R package (347, 359). The default parameters were used for each algorithm.

For each of the primary clinical endpoints, variable selection was conducted using LASSO, AIC selection, and RSF. For comparison, univariate filtering was also conducted with a significance threshold of $P < 0.1$. This resulted in several plausible models for each clinical endpoint.

2.2.4.3 Information criteria & events-per-variable ratio

To evaluate the results of the different variable selection algorithms, and to determine which to use for final model development, information criteria and the events-per-variable (EPV) ratios were compared.

As the number of variables increases, the apparent fit of a regression model will also increase when measured by the likelihood function. Information criteria were developed to penalize the apparent

model fit based on model complexity. BIC (277) compares the number of parameters estimated in the model k (equal to the number of variables) and the number of events n, with the maximum value of the log likelihood function $\hat{L}$, using the formula shown in **Equation 12**.

The lower the BIC value, the better the perceived quality of the model, provided each model was fitted to the same set of event data. The difference, Δ, in an information criterion, indicates the level of evidence in favour of the model with the lowest value. The difference in BIC is defined as $\Delta BIC = BIC_i - BIC_{min}$ where BIC_i is the value for a given model M_i, and BIC_{min} is the lowest observed value across all models.

Suggested ΔBIC thresholds correspond to weak ($\Delta BIC \leq 2$), positive ($2 < \Delta BIC \leq 6$), strong ($6 < \Delta BIC \leq 10$) or very strong ($\Delta BIC > 10$) evidence in favour of the model with the lower BIC (277, 360).

The number of variables to include may be guided by calculating the EPV ratio. This is the ratio between the number of events and the number of variables included in the model, as given by

$$EPV = \frac{n}{k}.$$

Whilst subject to debate, it is often cited that an EPV ratio above 10 is desirable to limit model overfitting to the observed data (345). Whereas recent simulation studies have argued this could be relaxed to an EPV of 5-9 without substantially increasing the risk of bias (361). The BIC and EPV ratio should be weighed against the inclusion of clinically relevant variables, so that confounding effects can be accounted for.

Models were fitted to each set of variable selection results. ΔBIC and EPV ratios were calculated to determine which set should be utilised for final model development. Due to the presence of missing data, BIC was calculated for each imputation and median values were reported.

2.2.4.4 Model fitting

Models were fitted to each imputation of the missing observations and coefficients were pooled into a single estimate by applying Rubin's rules (325).

Multivariate Cox PH models were fitted using variables retained after variable selection, adjusting for age, HCT-CI, and clinical variables with a univariate association (P<0.1). These additional clinical variables were forced into the model, irrespective of variable selection. Multivariate FG models were fitted using variables retained after variable selection for either NRM or RM. Additional clinical variables were not included due to the prohibitively low EPV ratios.

Variables that significantly (P < 0.05) violated the proportional hazards (PH) assumption were excluded; a PH test based on Schoenfeld residuals was used for CoxPH models (331) and a PH test based on the cumulative sum of residuals was used for FG models (332).

The resulting models were used to assess the difference in prognosis associated with an increase in the concentration of the selected serum proteins. Thus, addressing the fourth aim of the thesis.

3 Allo-HSCT graft composition study

3.1 Introduction

Haematopoietic stem cells (HSCs) are an essential functional constituent of peripheral blood stem cell (PBSC) grafts. Allogeneic transplantation of HSCs enables haematopoiesis, the formation and turnover of blood cells, in patients with dysfunctional or depleted bone barrow. This can provide a curative therapy for a variety of life-threatening haematological disorders.

In addition to HSCs, a multitude of mature immune cells are also infused during allogeneic haematopoietic stem cell transplantation (allo-HSCT). These include T cells, NK cells, B cells, dendritic cells and monocytes. The cellular composition of PBSC grafts has been shown to have a profound effect on patient outcomes (121, 139, 362). However, the optimal composition remains unknown (146) and the impact of individual subsets remains under-studied.

In collaboration with Francesca Kinsella, this chapter studies the cellular composition of PBSC grafts and how this relates to patient prognosis following allo-HSCT. The main objectives were to describe the cellular composition of PBSC grafts using single-cell mass cytometry (CyTOF). To define groups of PBSC grafts according to cellular composition, and to test the relationship between graft composition and prognosis. The clinical endpoints investigated are overall survival (OS), GvHD-free relapse-free survival (GRFS), relapse, transplant-related mortality (TRM), as well as acute and chronic GvHD (AGvHD, CGvHD).

To do this, CyTOF analysis was conducted on 185 PBSC graft donations (**Section 3.2**), and clinical and time-to-event data was collected for 155 recipient patients (**Section 3.3**). PhenoGraph clustering was utilised to identify clusters of cells from the CyTOF data (**Section 3.4**) which were characterised based on the expression of immune lineage-defining markers (**Section 3.5**). The cellular composition of grafts was described and compared between donors (**Section 3.6**). Cluster analysis for compositional data was conducted to identify distinct groups of PBSC grafts based on differences in the cell lineage composition (**Section 3.7**). Detailed cell subset classifications were then determined for each cluster and the composition of PBSC grafts groups were contrasted (**Section 3.8**). Lastly, statistical models tested the association of PBSC graft groups and graft composition with patient prognosis (**Section 3.9**).

Note, acquisition of the CyTOF data was conducted by Francesca Kinsella. I conducted the rest of the analysis described in this chapter under the supervision of Francesca Kinsella, Wayne Croft and Paul Moss.

3.2 PBSC graft data

The information available for each donor is summarised in **Table 3.1.** Variables included: age, sex, donor CMV (DCMV) status, relationship to recipient (related (RD) or unrelated (UD)), and stem cell source. The majority stem-bags were from male (71%), unrelated donors (63%) extracted from the peripheral blood (97%). For simplicity these are referred to as PBSC grafts, however, 4 were derived from bone marrow and 1 from cord blood. There was an even split of donors by CMV status (50% positive) and the average donor age was 39 (IQR 27-50). Of the donors, 46% of were not the same sex as the recipient (sex mismatch). In the donor information, between 5 and 6% of values were missing, except for donor age which was missing for 25% of donors.

Graft samples were subjected to CyTOF across 8 batches (**Appendix Figure C.1**) using a panel of 37 markers which included most major immune lineage defining markers in combination with numerous phenotypic and functional markers (**Table 3.2**). Markers were divided into two groups *lineage markers* and *cell state markers*. Lineage markers are relatively stable features consistently associated with particular cell types, whereas cell state markers primarily give information on transient cell states.

The raw CyTOF data was exported to an FCS file per graft. In pre-processing, FCS files were de-barcoded and consecutive gates were applied to remove non-biological events, doublets and debris. CyTOF data acquisition and pre-processing was conducted by Francesca Kinsella, whereas I conducted the data analysis. Following pre-processing, a total of 4,441,335 non-duplicate CyTOF events were observed, with a median of 20,467 events per FCS file (IQR 9633 – 32725) (**Table 3.3**).

Table 3.1 PBSC graft donor information.

Variable	Group	Overall	Missing (%)
N. stem bags		185	
Donor age (mean [IQR])		39 [27, 50]	24.3
Donor sex (%)	Male	124 (71)	5.9
	Female	50 (29)	
DCMV status (%)	Negative	87 (50)	5.9
	Positive	87 (50)	
Donor relationship (%)	RD	65 (37)	5.4
	UD	110 (63)	
Stem cell source (%)	PBSC	170 (97)	5.4
	BM	4 (2)	
	CORD	1 (1)	
Sex mismatch (%)	No	93 (54)	7
	Yes	79 (46)	
Patient sex (%)	Male	103 (59)	5.4
	Female	72 (41)	
CD34+ dose (cells x10^6; median [IQR])		4.91 [4.27, 5.79]	13
MNC dose (cells x10^8; median [IQR])		4.32 [3.20, 6.39]	53
T-cell dose (cells x10^6; median [IQR])		159.93 [124.15, 216.02]	46.5

Table 3.2 CyTOF panel.

Mass and tag	Specificity	Marker group
141Pr	CD3	Lineage
145Nd	CD4	Lineage
168Er	CD8a	Lineage
158Gd	CD10	Lineage
165Ho	CD11b Mac-1	Lineage
147Sm	CD11c	Lineage
159Tb	CD14	Lineage
209Bi	CD16	Lineage
142Ce	CD19	Lineage
149Sm	CD25 IL-2R	Lineage
155Gd	CD27	Lineage
144Nd	CD31 PECAM-1	Lineage
163Dy	CD33	Lineage
166Er	CD34	Lineage
160Gd	CD39	Cell state
89Y	CD45	Lineage
170Er	CD45RA	Lineage
176Yb	CD56 NCAM	Lineage
151Eu	CD70	Cell state
156Gd	CD86	Cell state
164Dy	CD96	Cell state
161Dy	CD117 c-kit	Lineage
143Nd	CD127 IL-7Ra	Lineage
173Yb	CD141	Cell state
169Tm	CD159a NKG2A	Cell state
175Lu	CD184 CXCR4	Cell state
167Er	CD197 CCR7	Lineage
171Yb	CD226	Cell state
172Yb	CD273 PD-L2	Cell state
148Nd	CD274 PD-L1	Cell state
174Yb	CD279 PD-1	Cell state
162Dy	CD335 NKp46	Cell state
150Nd	HLA-DR DP	Lineage
146Nd	IgD	Lineage
153Eu	NKG2C	Cell state
152Sm	TCRgd	Lineage
154Sm	TIGIT	Cell state

Table 3.3 CyTOF events per FCS file. The number of CyTOF events (N. events) per FCS file. The number of events corresponds to the number cells detected in each graft sample.

Statistic	N. events
Minimum	122
1st Quartile	9633
Median	20467
Mean	24007
3rd Quartile	32725
Maximum	175083

3.3 Allo-HSCT recipient cohort

A subset of the PBSC grafts (N =155) were paired with a patient in the recipient cohort (**Table 3.4**). Information on these patients included: recipient age (at time of transplant), sex, CMV status, HLA mismatch status, and haematopoietic cell transplantation-specific comorbidity index (HCT-CI). The type of conditioning regimen including total body irradiation (TBI), myeloablative conditioning (MAC) or reduced intensity conditioning (RIC) was known. As was the provision of GvHD prophylaxis which was either cyclosporine, or a combination of cyclosporine and methotrexate or mycophenolate mofetil (cyclosporine & other).

The majority of patients (93%) underwent T-cell–depleted allo-HSCT with either Alemtuzumab or anti-thymocyte globulin (ATG). Patient diagnoses were summarised into three broad groups: Myeloid (acute myeloid leukaemia (N 70), myelodysplastic syndromes (N 20), myelofibrosis (N 4), chronic myeloid/myelomonocytic leukaemia (N 11)), Lymphoid (B-cell non-Hodgkin lymphoma (N 19), B-cell acute lymphoblastic leukaemia (N 16), chronic lymphocytic leukaemia (N 6), myeloma (N 1), Hodgkin lymphoma (N 8), T-cell non-Hodgkin lymphoma (N 12), T-cell acute lymphoblastic leukaemia (N 3), and Others (other malignancies (N 3), aplastic anaemia (N 7)).

Patients were followed up for a median of 3.7 years (CI 3.31-4.9, IQR 2.74-5.75). A maximum follow-up time of 2-years was selected. During this 2-year period, 67% of recipients experienced an event (AGvHD (grade II-IV), CGvHD, relapse or death). Of the recipients, 26% were diagnosed with AGvHD, 26% with CGvHD and 20% with relapse of their residual disease. Overall, 38% of patients did not survive – with 22% of patients experiencing transplant-related mortality (TRM), whereas 16% of patients experienced relapse-related mortality (RM).

Table 3.4 Allo-HSCT patient information.

Variable	Group	Overall	Missing (%)
N. patients		155	
Age (median [IQR])		53.00 [44.00, 61.00]	0
Patient sex (%)	Male	91 (59)	0
	Female	64 (41)	
Diagnosis (%)	Myeloid	90 (58)	0
	Lymphoid	57 (37)	
	Others	8 (5)	
Patient CMV status (%)	Negative	40 (26)	0
	Positive	115 (74)	
HLA mismatch (%)	No	141 (93)	2.6
	Yes	10 (7)	
T-cell depletion (%)	None	11 (7)	0
	Alemtuzumab	102 (66)	
	ATG	42 (27)	
TBI (%)	No	139 (90)	0
	Yes	16 (10)	
GvHD Prophylaxis (%)	Cyclosporine	112 (72)	0
	Cyclosporine & Other	43 (28)	
MAC/RIC (%)	MAC	48 (31)	0
	RIC	107 (69)	
HCT-CI (%)	0	85 (55)	0.6

	1	19 (12)	
	2	19 (12)	
	3+	31 (20)	
OS	Censored	96 (62)	0
	Event	59 (38)	
GRFS	Censored	51 (33)	0
	Event	104 (67)	
AGvHD (Grade>2)	Censored	115 (74)	0
	Event	40 (26)	
CGvHD	Censored	115 (74)	0
	Event	40 (26)	
Relapse	Censored	124 (80)	0
	Event	31 (20)	
Competing risks	Censored	96 (62)	0
	TRM	34 (22)	
	RM	25 (16)	

3.4 Cluster analysis for CyTOF data

CyTOF events (event = cell) were partitioned into cell populations by the PhenoGraph clustering algorithm using the lineage markers as input (**Table 3.2**). An introduction to this method can be found in **Section 2.1.5.1**.

The PhenoGraph algorithm identified 39 cell populations (clusters) within the CyTOF data. The number of cells assigned to each cluster are detailed in **Table 3.5**, along with the number of PBSC grafts in which each cluster was observed.

The median number of cells per cluster was 76,673 (range 44 - 374,077). The majority of clusters, 32 out of 39, were observed at least 90% of the PBSC grafts. Some small clusters (< 5000 cells) including 20, 21, 23 and 33 were observed in over 88% of PBSC grafts. However, other small clusters (< 5000 cells) such as 9, 26, 28, and 34 were observed in less than 50% of samples, so may not be biologically relevant (artifacts in the structure of the data) or may be present in insufficient numbers to include in downstream analysis.

Table 3.5 Distribution of CyTOF events by cluster. Count of cells (events in the CyTOF data) assigned to each cluster by PhenoGraph, with the count of PBSC grafts in which the cluster was observed.

	Cells		PBSC grafts	
Cluster	**N**	**Proportion**	**N**	**Proportion**
1	76673	0.02	185	1.00
2	251417	0.06	185	1.00
3	13498	<0.01	180	0.97
4	37797	0.01	183	0.99
5	327877	0.07	185	1.00
6	177214	0.04	184	0.99
7	165873	0.04	182	0.98
8	27397	0.01	131	0.71
9	152	<0.01	21	0.11
10	159085	0.04	185	1.00
11	144854	0.03	185	1.00
12	25453	0.01	155	0.84
13	267396	0.06	184	0.99
14	131806	0.03	185	1.00
15	134709	0.03	185	1.00
16	21654	<0.01	183	0.99
17	8394	<0.01	177	0.96
18	226262	0.05	184	0.99
19	142822	0.03	185	1.00
20	1959	<0.01	162	0.88
21	2691	<0.01	169	0.91
22	58167	0.01	185	1.00
23	4038	<0.00	168	0.91
24	242513	0.05	185	1.00
25	374077	0.08	185	1.00
26	159	<0.01	62	0.34
27	251728	0.06	185	1.00
28	44	<0.01	26	0.14
29	265224	0.06	185	1.00
30	233504	0.05	181	0.98
31	213644	0.05	184	0.99
32	42216	0.01	184	0.99
33	4345	<0.01	172	0.93
34	1115	<0.01	48	0.26
35	63608	0.01	185	1.00
36	13047	<0.01	181	0.98
37	168645	0.04	185	1.00
38	54433	0.01	185	1.00
39	105845	0.02	184	0.99

3.5 Characterisation of cluster phenotypes: cell lineages

Due to the unsupervised nature of the PhenoGraph clustering algorithm, the cell subset or phenotype of each cluster is unknown, and the defining features need to be characterised.

The phenotype associated with each cluster was assessed systematically, using a combination of four complimentary characterisation methods: mean marker intensity (**Section 453.5.1**), marker enrichment modelling (**Section 3.5.2**), dimension reduction (**Section 3.5.3**), and visualisation of marker expression (**Section 3.5.4**). I point the reader towards **Section 2.1.8** of the methods chapter for an introduction to each of these methods.

The final lineage classifications and composition of the PBSC grafts are summarised in **Section 3.6**. Each cluster was initially classified into general cell lineages (CL). These included: B lymphocytes (B cells), dendritic cells (DCs), natural killer cells (NK cells), monocytes and T lymphocytes (T cells). Once assigned to a lineage, clusters were subsequently classified into detailed cell subsets (CS), this is discussed later in this chapter (**Section 3.8**).

The markers used to define each lineage were CD19 (B cells), HLA-DR (DCs), CD56 (NK cells), CD14 (monocytes) and CD3 (T cells). Cells that were lineage negative (Lin^-) but positive for CD34 were labelled as hemopoietic stem cells (HSCs), whereas if no lineage could be assigned, they were labelled unclear.

Note, clusters consistently positive or negative for a marker are labelled $^+$ or $^-$, respectively. If the intensity of expression was exceptionally high or low compared to cells of the same lineage, they were labelled bright ($^{++}$) or dim (^{dim}).

3.5.1 Mean marker intensity

The distribution of marker intensity was visualised for the main lineage defining markers (**Figure 3.1**). CD45, which is expressed variably across a diverse range of cell subsets, was included on the x-axis to aid delineation of potential groups.

For the T cell, monocyte, B cell and HSC lineage markers there was a clear separation of randomly sampled cells into positive and negative groups. For the NK cell and DC lineage markers, there was a less clear distinction between positive and negative cells. The mean intensity thresholds determined for these markers were: 2 (CD3), 1.5 (CD14), 1 (CD19) and 2 (CD34), 0.5 (CD56) and 1 (HLA-DR).

Mean marker intensity (MMI) was then calculated for each cluster of cells and compared to the respective thresholds (**Table 3.6**). Clusters which exceeded a threshold were considered positive for that marker and assigned to the appropriate cell lineage (CL by MMI) (**Table 3.6**).

Clusters which were HLA-DR$^+$ CD14$^+$ were labelled as monocytes whereas those that were HLA-DR$^+$ CD19$^+$ were labelled as B cells. Clusters that were positive for an implausible combination of lineage markers or were lineage negative (Lin$^-$) were labelled as being unclear.

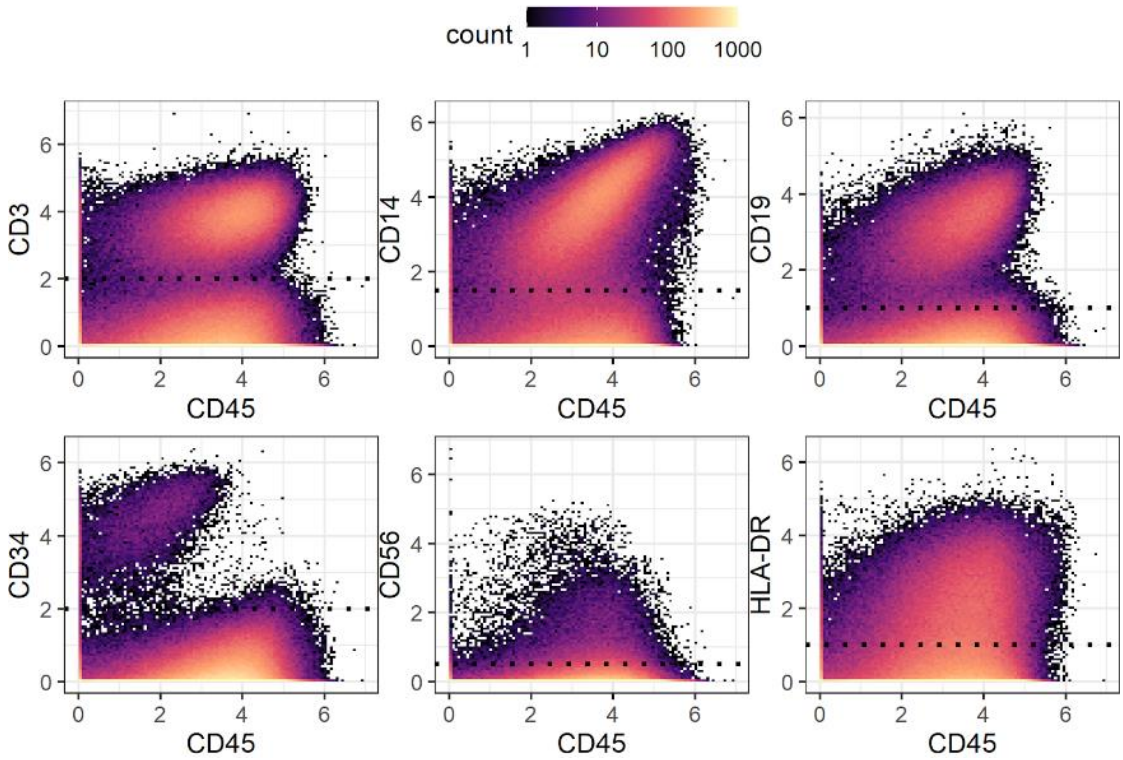


Figure 3.1 Mean marker intensity thresholds. Bi-axial plots showing marker intensity by cell count across 500,000 randomly sampled cells. Dashed lines indicate the threshold determined for each lineage marker.

Table 3.6 Mean marker intensity by cluster and cell lineage classification. The cell lineage assigned to each cluster based on mean marker intensity (CL by MMI). Mean marker intensity values which exceeded the respective threshold values are highlighted in green.

Cluster	CD3	CD19	CD56	CD14	CD34	HLA-DR	CL by MMI
1	0.21	0.12	0.03	0.7	0.11	1.15	DC
2	0.23	0.11	0.03	0.4	0.18	0.83	Unclear
3	0.22	3.66	0.03	0.1	0.22	1.86	B
4	0.26	0.1	0.1	0.21	0.38	2.27	DC
5	0.33	0.09	0.03	3.13	0.09	0.27	M
6	0.22	0.1	0.03	2.69	0.13	1.73	M
7	0.37	0.13	0.06	4.6	0.5	0.33	M
8	0.48	0.1	0.08	3.65	0.17	0.25	M
9	0.94	1.89	0.02	3.21	0.12	0.8	Unclear
10	3.46	0.11	0.07	0.15	0.2	0.08	T
11	3.4	0.13	0.1	0.17	0.06	0.08	T
12	1.36	0.12	0.05	0.24	0.21	0.05	Unclear
13	0.32	0.13	0.06	4.31	0.29	1.99	M
14	3.65	0.1	0.2	0.12	0.21	0.09	T
15	0.2	0.08	0.92	0.13	0.16	0.08	NK
16	3.63	0.11	0.03	0.1	0.05	0.21	T
17	3.62	0.21	0.07	2.82	0.24	0.58	Unclear
18	0.23	2.91	0.04	0.1	0.18	1.29	B
19	3.55	0.11	0.03	0.14	0.34	0.05	T
20	3.51	3.01	0.07	0.16	0.3	2.01	Unclear
21	0.21	0.09	0.13	0.17	0.14	0.14	Unclear
22	0.17	0.1	0.03	0.1	0.26	1.94	DC
23	0.39	2.73	0.06	3.15	0.3	2.17	Unclear
24	3.64	0.16	0.03	0.18	0.07	0.07	T
25	0.19	3.37	0.05	0.1	0.24	2.37	B
26	0.48	0.18	0.03	3.78	4.33	1.27	Unclear
27	0.31	0.11	0.04	3.2	0.14	0.2	M
28	3.93	0.18	0.01	0.19	4.69	0.85	T, HSC
29	4.04	0.15	0.03	0.13	0.18	0.06	T
30	0.29	0.11	0.06	4.27	0.25	0.21	M
31	0.21	0.08	0.02	2.22	0.06	0.23	M
32	0.17	0.08	0.03	0.33	0.37	0.1	Unclear
33	3.61	0.14	0.04	0.16	0.2	0.08	T
34	4.51	0.17	2.34	0.14	0.16	0.09	T, NK
35	4	0.13	0.22	0.14	0.08	0.08	T
36	3.28	0.1	0.02	0.13	0.05	0.2	T
37	0.14	2.28	0.02	0.11	0.12	1.76	B
38	0.27	0.16	0.07	0.8	0.14	0.3	Unclear
39	0.23	0.09	0.02	0.1	4.09	0.67	HSC
Threshold:	2	1	0.5	1.5	2	1	

3.5.2 Marker enrichment modelling

Marker Enrichment Modelling (MEM) was used to infer enrichment or downregulation of markers relative to other clusters (**Figure 3.2**). An introduction to this methodology can be found in **Section 2.1.8.2**.

Comparing the MEM results to known cell subsets, a putative cell lineage was determined for each cluster (CL by MEM) (**Table 3.7**). An additional lineage label was required to classify Cluster 34 which was enriched for both CD56 and CD3 so was labelled as natural killer T (NKT) cells (363).

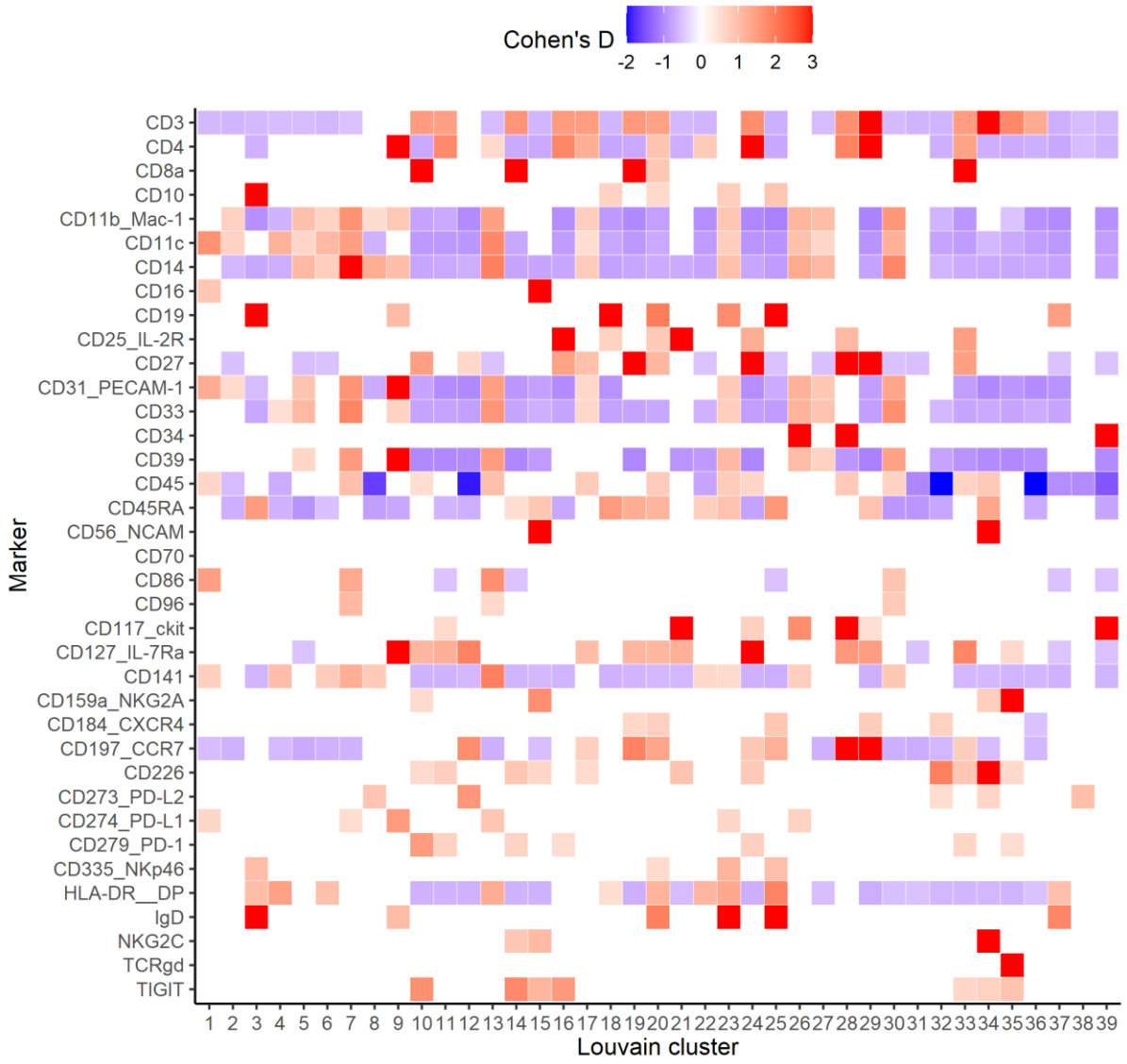


Figure 3.2 Marker enrichment estimates by cluster. Positive and negative marker enrichment (Cohen's d) by cluster. Small effects (d < ±0.5) were excluded, and estimates were limited to ±3 to aid visual interpretation.

There was a broad correspondence between the CL labels assigned by MEM and by mean marker intensity. There were a few exceptions; the lineages of clusters 22 & 31 were unclear based on marker enrichment but were labelled as DCs and monocytes, respectively, when comparing MMI. Clusters 26 & 28 had an ambiguous phenotype by MEM but only contained 159 and 44 cells, respectively. They were also observed in relatively few grafts so may not be biologically relevant.

Cluster 2 was enriched for mixture of both monocyte and DC markers (CD11c & CD11b) (364, 365). However, the average expression of CD14 and HLA-DR were relatively low so a classification could not be determined by either method. As was the case for Cluster 17 which was enriched for an implausible combination of CD3 and CD14.

3.5.3 Dimension reduction

Dimension reduction was applied to visualise cellular populations and determine phenotypic characteristics. Cells in the CyTOF data were projected onto a 2 dimensional (2-D) embedding output by UMAP and t-SNE. An introduction to each of the dimension reduction algorithms applied can be found in **Section 2.1.7.**

This revealed distinct areas of the UMAP embedding with high levels of CD3, CD19 or CD14 expression (**Figure 3.3**). A separate population of $CD34^+$ cells was also visible. However, CD34 expression was visible sporadically across the embedding. High levels of HLA-DR expression often coincided with cells expressing CD14 or CD19. Cells expressing CD56 were concentrated in the middle of the embedding, but some $CD56^+$ cells were distributed in the $CD3^+$ areas.

Two distinct subpopulations of $CD3^+$ cells were clearly visible, one that was largely $CD3^+$ $CD4^+$ and another that was $CD3^+$ $CD8^+$. This illustrates that UMAP can group together complex phenotypic subgroups based solely on structure within the data without any prior knowledge.

Cells were less concentrated in the t-SNE visualisation, making it easier to identify smaller populations with relatively high marker expression (**Appendix Figure C.2**). However, the CD3+ CD4+ and CD3+ CD8+ cells were distributed across several disjointed areas demonstrating how less of the global structure is retrained by t-SNE compared with UMAP.

A lineage classification was determined for each cluster by comparing the cluster distribution to areas of high marker expression (**Table 3.7**). Classification by dimension reduction corresponded closely with the labels assigned by MMI and MEM analysis. Some unclassified clusters were shown by dimension reduction to contain T cells (Cluster 12), monocytes (Cluster 9) or a mixture of monocytes and DCs (Cluster 2).

An additional method, potential of heat diffusion for affinity-based transition embedding (PHATE), was also applied (296). PHATE captures trajectory structures between observations, this helps to identify branching or endpoints in the data structure such as cell differentiation (296, 297). PHATE analysis was conducted after characterisation of cluster phenotypes so is discussed later in this chapter (**Section 3.6**).

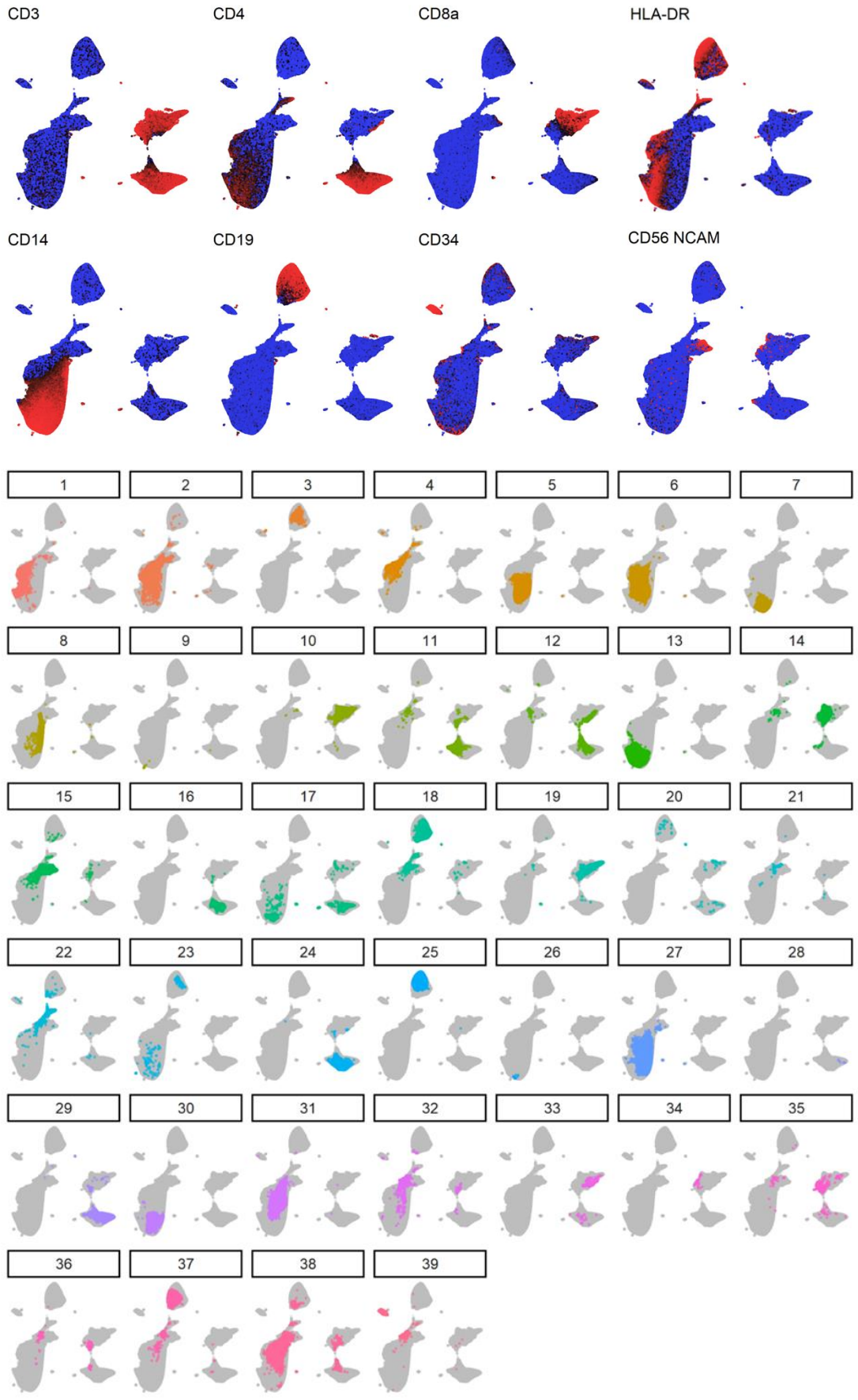


Figure 3.3 UMAP visualisation of cells by marker expression intensity and cluster.

Table 3.7 Dimension reduction cell subset classification. The general cell lineage of clusters by marker enrichment modelling (CL by MEM), by mean marker intensity (CL by MMI), and by dimension reduction (CL by UMAP/t-SNE), the count of cells in the CyTOF data by cluster, and the count of PBSC grafts in which the cluster was observed.

Cluster	CL by MMI	CL by MEM	CL by UMAP / t-SNE	Cells (N)	Stem-bags (N)
1	DC	DC	DC	76,673	185
2	Unclear	M / DC	M / DC	251,417	185
3	B	B	B	13,498	180
4	DC	DC	DC	37,797	183
5	M	M	M	327,877	185
6	M	M / DC	M / DC	177,214	184
7	M	M / DC	M	165,873	182
8	M	M	M	27,397	131
9	Unclear	Unclear	M	152	21
10	T	T	T	159,085	185
11	T	T	T	144,854	185
12	Unclear	Unclear	T	25,453	155
13	M	M	M	267,396	184
14	T	T	T	131,806	185
15	NK	NK	NK	134,709	185
16	T	T	T	21,654	183
17	Unclear	Unclear	Unclear	8,394	177
18	B	B	B	226,262	184
19	T	T	T	142,822	185
20	Unclear	Unclear	Unclear	1,959	162
21	Unclear	Unclear	Unclear	2,691	169
22	DC	Unclear	DC	58,167	185
23	Unclear	B	B	4,038	168
24	T	T	T	242,513	185
25	B	B	B	374,077	185
26	M, HSC	Unclear	Unclear	159	62
27	M	M	M	251,728	185
28	T, HSC	Unclear	T	44	26
29	T	T	T	265,224	185
30	M	M	M	233,504	181
31	M	Unclear	M	213,644	184
32	Unclear	Unclear	Unclear	42,216	184
33	T	T	T	4,345	172
34	T, NK	NKT	NKT	1,115	48
35	T	T	T	63,608	185
36	T	T	T	13,047	181
37	B	B	B	168,645	185
38	Unclear	Unclear	Unclear	54,433	185
39	HSC	HSC	HSC	105,845	184

3.5.4 Labelling unclassified clusters

For the majority of clusters, a clear lineage could be assigned by comparing the results of MMI, MEM and dimension reduction. To classify the remaining clusters marker intensity was visualised as bi-axial density plots and MEM results were reviewed infer which markers influenced cluster assignment.

Cluster 32 represented ~1% of cells but did not conform to any of the expected cell lineages. These cells were Lin^- $CD34_{dim}$ DNAM-1^+ $CXCR4^+$ $CD45RA^-$ PD-$L2^+$ (**Figure 3.4**) which broadly corresponds to the description of a population of committed lymphoid precursors (366). These cells typically reside in the bone marrow but are recruited to the peripheral blood by G-SCF mobilisation and chronic infection (367, 368). The similarities were considered sufficient to assign a DNAM-1^+ lymphoid progenitor cell (lymph. progenitor) label to these cells.

Clusters 6, 7 and 31 cells were $CD14^+$ and Cluster 38 was $CD14_{dim}$ so were assigned to the monocyte lineage (**Appendix Figure C.3**). Cluster 22 was confirmed to contain HLA-DR^+ $CD14^-$ cells so were likely DCs (**Appendix Figure C.4**). Cluster 2 cells were $CD11b^+$ $CD11c^+$ but appeared to contain two separate sub-populations, one that was $CD14^-$ HLA-DR^+ $CD11b^+$ $CD11c^+$ (DCs) and another was $CD14_{dim}$ HLA-DR^+ $CD11b^+$ $CD11c^+$ (monocytes) (**Appendix Figure C.5**). This was interpreted as the cluster containing a mixture of both monocytes and DCs (mixed M / DC).

Cluster 12 was $CD3_{dim}$ but otherwise lineage negative (Lin^-), resulting in a T cell classification (**Appendix Figure C.6**). Whereas clusters 9, 26 and 28 were observed in less than 35% of stem-bags so were labelled as having low numbers (Low N.).

Cluster 21, which contained 2,691 cells, was Lin^- $CD34^-$ and strongly enriched for CD117 (c-Kit). This populations al co-localised with lymphoid cells on the UMAP embedding (**Appendix Figure C.7**). These cells were labelled common lymphoid progenitors (CLP) based on the expression of c-Kit and lack of myeloid markers (369, 370).

Two small clusters were found to include $CD3^+$ doublets which were not removed in pre-processing (**Appendix Figure C.8**). Cluster 17 contained 8,394 $CD3^+$ cells, a number of which were $CD3^+$ $CD14^+$ doublets. This could also indicate technical artifacts; however, T cell-monocyte complexes associated with immune perturbations have been reported in flow-cytometric analysis (371). Due to the consistent CD3+ expression these were simply labelled as T cells. Cluster 20, which contained 1,959 cells, was positive for a biologically implausible combination of CD3 and CD19. As this likely represents a technical artifact of the CyTOF analysis this cluster was labelled as spurious.

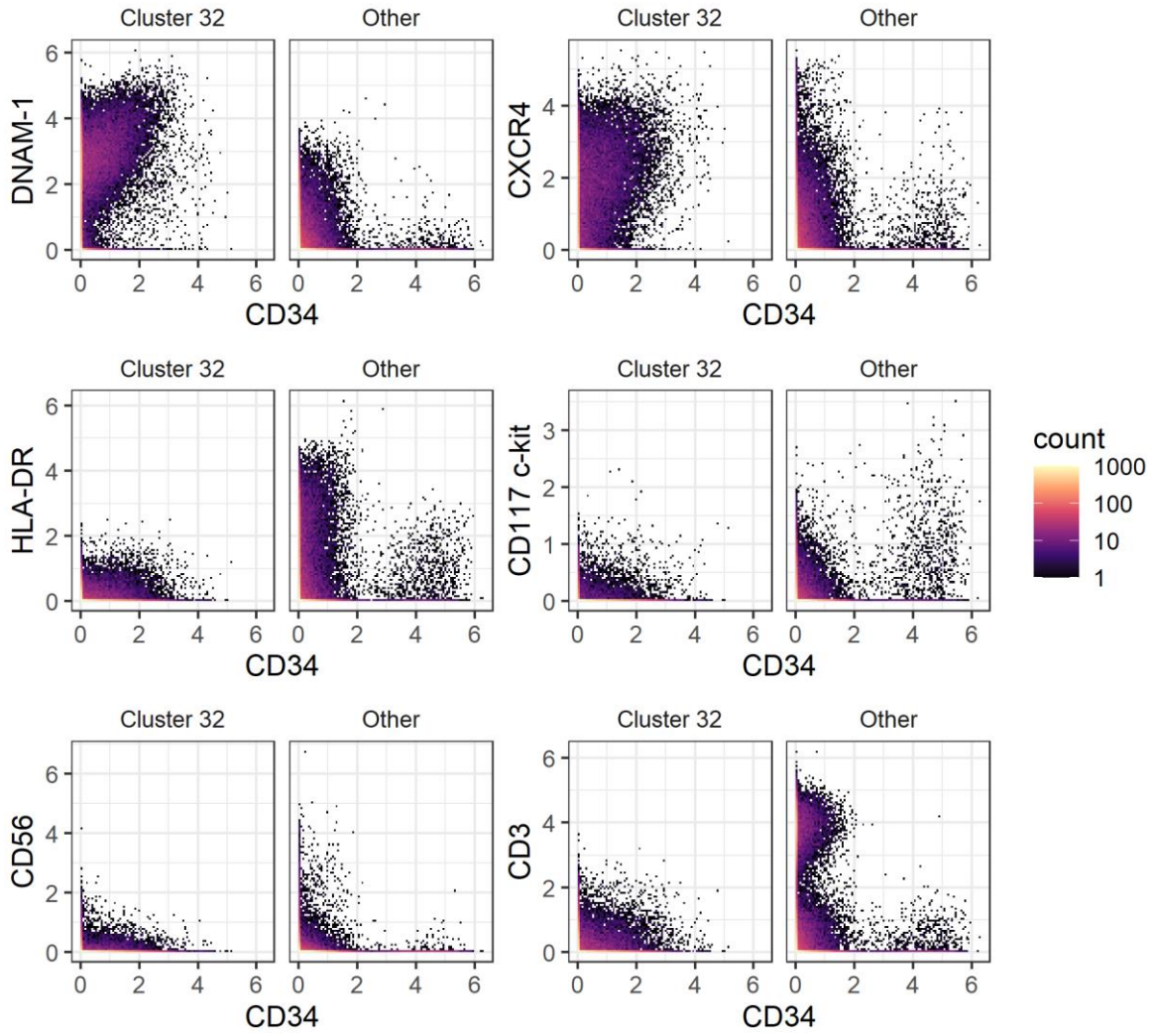


Figure 3.4 Marker intensity in cells from Cluster 32 – DNAM-1^{+} Lymphoid Progenitors. Bi-axial plots of marker intensity by cell count across cells assigned to Cluster 32 compared to 500,000 cells sampled at random (Other).

3.6 Description of PBSC graft composition

Summarising the results of each cluster characterisation method (**Section 3.5**)**,** a general cell lineage classification was determined for each cluster (**Table 3.8**).

Table 3.8 Cell lineage assigned by cluster. The general cell lineage of clusters by all methods (CL).

C	CL
1	DC
2	Mixed M / DC
3	B
4	DC
5	M
6	M
7	M
8	M
9	Low N
10	T
11	T
12	T
13	M
14	T
15	NK
16	T
17	T
18	B
19	T
20	Spurious
21	CLP
22	DC
23	B
24	T
25	B
26	Low N
27	M
28	Low N
29	T
30	M
31	M
32	Lymph. progenitor
33	T
34	NKT
35	T
36	T
37	B
38	M
39	HSC

The majority of cells within the CyTOF data were monocytes (~39%), T cells (~28%) and B cells (~18%) (**Table 3.9**). A smaller proportion of cells belonged to the DC (~4%), NK cell (~3%) and HSC (~2%) compartments, and a further ~6% of cells showed a mixture of monocyte or DC like traits. Approximately 1% of cells were DNAM-1^{+} lymphoid progenitor cells (Lin^{-} CD34$_{dim}$ DNAM-1^{+} CXCR4^{+} CD45RA^{-}). CLP and NKT cells each constituted < 0.1% of observed events in the CyTOF data.

Table 3.9 Composition of CyTOF data by cell lineage (CL).

CL	Events/ Cells (N)	Perc. (%)
M	1,719,066	38.71%
T	1,222,805	27.53%
B	786,520	17.71%
Mixed M/DC	251,417	5.66%
DC	172,637	3.89%
NK	134,709	3.03%
HSC	105,845	2.38%
Lymph. progenitors	42,216	0.95%
CLP	2,850	0.06%
Spurious	1,959	0.04%
NKT	1,115	0.03%
Low N	355	<0.01%

Clusters assigned to the same lineage appeared to colocalise when visualised on the UMAP embedding (**Figure 3.5**). This increases confidence that the lineage classifications were reasonable. One limitation of UMAP is that cell lineages appear to be completely distinct, when in reality cells exist in a continuum of cellular differentiation. PHATE, a recently developed visualisation method, is better at representing differentiation pathways between cell types (296). When PHATE was applied to the CyTOF data, it illustrated the continuous differentiation of HSCs into each of the distinct cell lineages. The location of the lymph. progenitor cells in between the HSC and NK/T cell subsets provides further evidence that this represents an intermediate stage of differentiation.

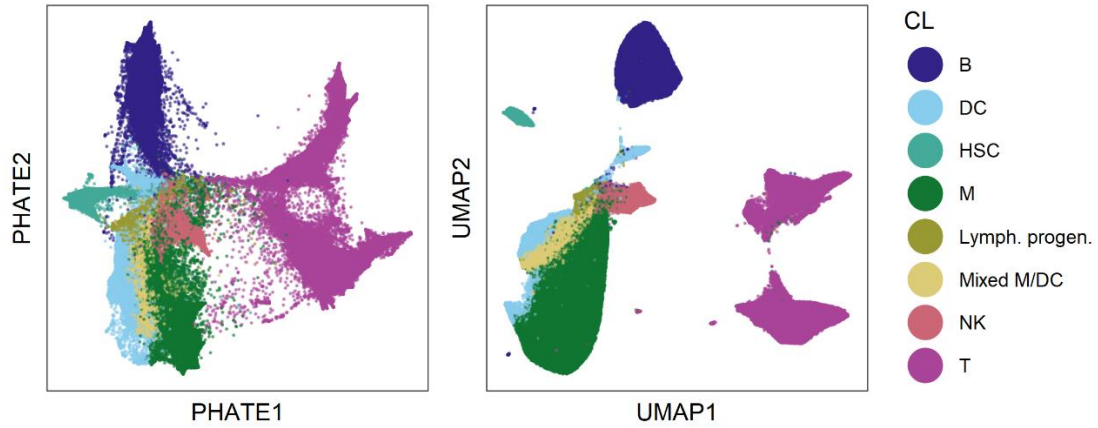


Figure 3.5 UMAP and PHATE visualisation of cell lineages (CL) across all grafts. Clusters labelled as NKT, CLP, Low N and spurious are not shown.

As the CyTOF data was generated from an aliquoted sample of the PBSC grafts, the cell counts are unrepresentative of what was infused into each recipient. However, the relative proportion of cell types in the CyTOF data is biologically meaningful. Graft cellular composition was therefore compared between donations as summarised in **Table 3.10.**

Based on the median, the majority of cells in each graft were monocytes (36%; IQR 20 - 51%), T cells (28%; IQR 17 - 39%), and B cells (15%; IQR 10 - 23%). A smaller proportion of cells in the grafts belonged to the DC (3%; IQR 2 - 5%), NK cell (2%; 1 - 4%), and HSC (2%; 1-4%) compartments. Whereas Lymph. progenitors contributed a median of 1% (IQR 0 – 2%) of cells detected in grafts. The wide interquartile ranges indicate that there was considerable variation in the cellular composition between donors.

Substantial heterogeneity was also apparent in the measured cell concentrations of mononuclear cells (MNC), T cells and CD34+ cells for a subset of grafts (**Table 3.11**).

Table 3.10 Composition of PBSC grafts by cell lineage (CL). Percentage of graft cells summarised by the median, mean, interquartile range (IQR) and range.

CL	Median	Mean	IQR
M	35.72%	35.88%	(19.85 - 50.81%)
T	27.74%	29.03%	(17.42 - 38.81%)
B	15.47%	17.97%	(10.04 - 22.74%)
DC	3.17%	3.79%	(2.12 - 4.55%)
NK	2.13%	3.89%	(1.28 - 4.47%)
HSC	2.08%	2.60%	(0.97 - 3.67%)
Mixed M/DC	1.77%	5.09%	(0.64 - 4.33%)
Lymph. progenitors	0.64%	1.62%	(0.34 - 1.66%)
CLP	0.04%	0.06%	(0.02 - 0.07%)
Spurious	0.03%	0.05%	(0.01 - 0.07%)
NKT	0.00%	0.01%	(0 - 0%)
Low N	0.00%	0.01%	(0 - 0.01%)

Table 3.11 Measured cellular doses. Measured dose of mononuclear cells (MNC), T cells and CD34+ cells infused by recipient. Concentrations are summarised by the median, mean, IQR and number of observations (N. obvs.).

CL	Unit	Median	Mean	IQR	N. obvs.
MNC dose	x 10^8 / kg	4.32	4.97	(3.200 - 6.392)	86
T cell dose	x 10^6 / kg	159.93	178.62	(124.15 - 216.02)	99
CD34+ dose	x 10^6 / kg	4.91	5.03	(4.27 - 5.79)	160

3.7 Identifying distinct PBSC graft groups

Cellular composition varied substantially between PBSC grafts and as a result the recipient patients were infused with vastly different doses of each lineage. This raises serval important questions; does this variation constitute distinct types of PBSC graft? Do donor factors correlate with differences in graft composition, and ultimately might this impact a patient's response to the therapy?

To address these questions, cluster analysis was conducted to infer the presence or absence of distinct groups of PBSC grafts based on differences in cellular composition. Graft compositions were divided into parts corresponding to the general cell lineages (M, T, B, DC, NK, Mixed M/DC and Lymph. progenitors and HSC). Differences in donor factors, cellular composition and patient outcome were then contrasted between the resulting groups of PBSC grafts.

As described in **Section 2.1.6.1**, the cell counts generated by CyTOF are compositional in nature (260). Compositional data are incompatible with classical statistical methods due to the constant-sum constraint (265)(372). A log-ratio approach was utilised in this study to transform the composition into unconstrained variables prior to cluster analysis.

3.7.1 Cluster analysis for compositional data

Cluster analysis for compositional data (clustCoDa) was conducted using an isometric log-ratio transformation and Gaussian Mixture Model (GMM) clustering. An introduction to this methodology can be found in **Section 2.1.6.2**.

The number clusters present in the data was inferred using the Bayesian information criterion (BIC) (276). BIC model comparison indicated the use of a GMM with $K = 3$ (**Figure 3.6**). A second local minimum was apparent with $K = 5$. However, this resulted in prohibitively small clusters (N < 10), therefore, the final GMM was fitted with 3 mixture components (clusters).

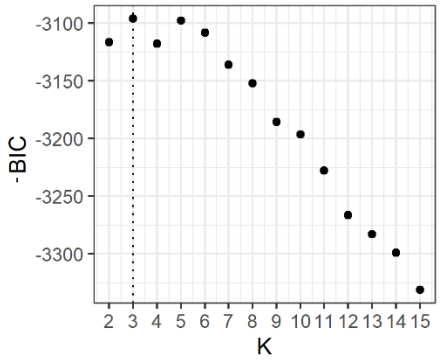


Figure 3.6 Model comparison by BIC to select number of clusters. Value of the (negative) Bayesian information criterion (BIC) by the number of mixture components (K) used in fitting the clustering model. The dashed line represents the number of components which maximised (negative) BIC.

The final fitted GMM was used to classify each of the 155 PBSC grafts into one of 3 clusters, which will be referred to as *PBSC graft groups* (**Table 3.12**). The Silhouette width was calculated for each PBSC graft group (284-286) (**Table 3.12**). All PBSC graft groups had a positive silhouette, indicating they constituted valid clusters.

Overall, cluster analysis for compositional data inferred the presence of three distinct PBSC grafts groups within the CyTOF data. One large group of grafts, A (N = 80), and two smaller groups, B (N = 41) and C (N = 34). Each of which were defined based on differences in cellular composition.

Table 3.12 Number (N.) of grafts by PBSC graft group and silhouette width.

PBSC graft group	N.	Silhouette
A	80	0.25
B	41	0.15
C	34	0.17

3.7.2 Prognosis by PBSC graft group

UMAP was applied to the log-transformed data to visualise the PBSC graft groups (**Figure 3.7A**). Each point represents an individual PBSC graft coloured by the group assigned by the GMM. Group C formed a distinct cluster, whereas the distributions of groups A and B were adjacent suggesting greater similarity in the composition of these two groups.

To assess if donor factors differed between the PBSC graft groups, the proportion of CMV positive, female and unrelated donors was compared along with the donor age distribution (**Figure 3.7B**). Group A had the youngest donors (median 34; IQR 24 – 48) and the lowest percentage of CMV positive donors (43%). Group B had the oldest donors (median 41; IQR 29 - 51), a greater percentage of which were female (39%) but relatively few that were unrelated to the recipients (51%). Whereas Group C had the most CMV positive (62%) unrelated donors (71%), and the lowest percentage of female donors (23%).

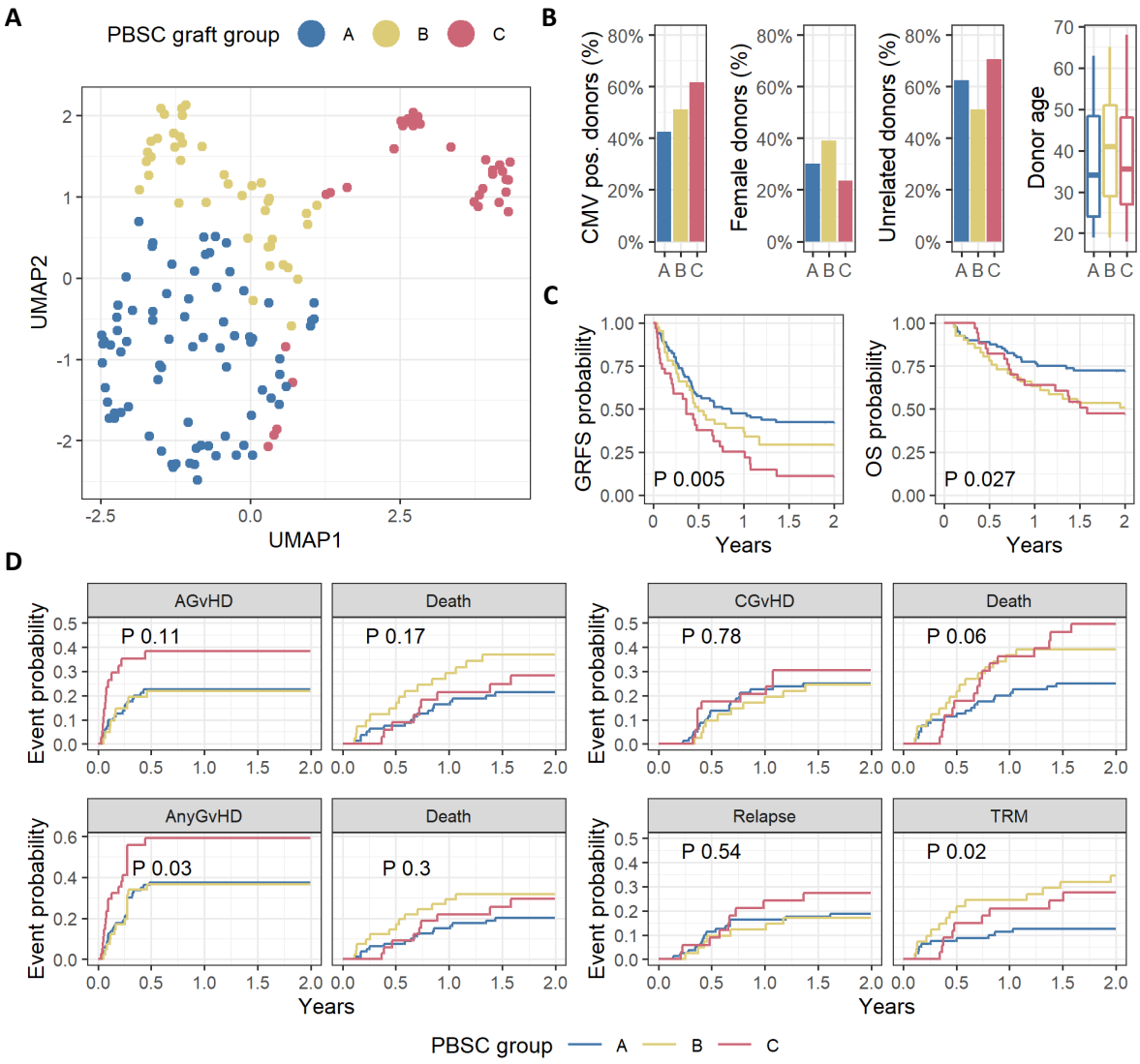


Figure 3.7 Donor factors and prognosis by PBSC graft group. (A) UMAP visualisation of PBSC graft groups. (B) Donor factors by PBSC graft group. (C) Probability of GvHD-free, relapse-free survival (GRFS) and overall survival (OS) by PBSC graft group based on the Kaplan-Meier estimator. (D) Cumulative incidence of AGvHD,

CGvHD, GvHD of any kind (Any GvHD), death from other causes (death), relapse and transplant-related mortality (TRM).

The GvHD-free, relapse-free survival (GRFS) and overall survival (OS) rates were estimated for recipients stratified by PBSC graft group (**Figure 3.7C**). The Kaplan-Meier estimator indicated significant differences in the GRFS rate (p = 0.005) and OS rate (p = 0.027) between groups.

Patients receiving a PBSC graft from Group A had a 48% (CI 38 - 60) probability of surviving event-free at one-year post transplant. This is compared to 37% (CI 24 - 55) for grafts from Group B, and 25% (CI 14 - 45) for grafts from Group C. By two years post-transplant, the probability of GRFS had more than halved to 11% (CI 4 - 31) for recipients of Group C grafts. At the same time point, the GRFS probability was 42% (CI 0.33 – 0.55) in Group A, and 29% (CI 0.18 – 0.47) in Group B. Recipients of Group A also had the highest probability of OS 2 years post-transplant at 72% (CI 63-83), compared to 51% (CI 38 - 69) for Group B and 48% (CI 33 - 69) for Group C.

The cumulative incidence of AGvHD and CGvHD were estimated but no significant differences were observed between groups (CGvHD P 0.78; AGvHD P 0.11) (**Figure 3.7D**). Due to the relatively low number of events within each group, the comparison was repeated with both events combined into a single composite outcome, GvHD of any kind (GvHD). This suggested the incidence of GvHD was significantly different between PBSC graft groups (P 0.03). Group C had a 59% (CI 42 – 76) probability of GvHD one-year post transplant compared to 38% (CI 27 – 48) and 37% (CI 22 – 51) in groups A and B, respectively.

Comparing the incidence of relapse and transplant-related mortality (TRM) as competing risks suggested no difference in relapse incidence (P 0.54) but a significant difference in TRM (P 0.02) (**Figure 3.7D**). Two years post-transplant the probability of TRM in group A was 12% (CI 5 - 20) but more than double in groups B (35%; CI 20 – 49) and C (28%; CI 12 – 43). A similar pattern emerged when the incidence of RM (death preceded by a relapse) was compared as the competing risk for TRM, with no difference in the cumulative incidence functions (P 0.71) (**Figure 3.8**).

Together these findings indicate recipients of a graft from Group A had an improved probability both OS and GRFS compared with groups C and B as well as a lower rate of TRM. However, no difference in the relapse or relapse-related mortality were observed between groups. Group C had a higher incidence of both TRM and GvHD of any kind, whereas Group B only had a higher incidence of TRM.

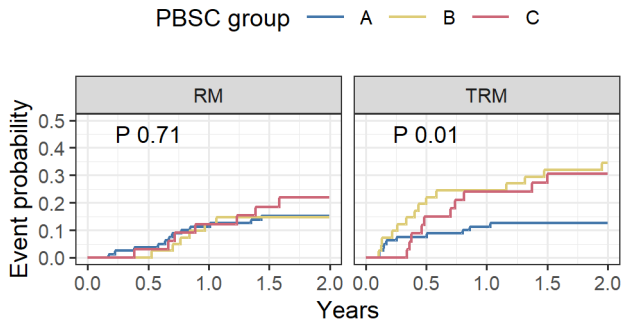


Figure 3.8 Cumulative incidence of relapse mortality (RM) by PBSC graft group. The competing risk transplant-related mortality (TRM) is shown for comparative purposes.

3.7.3 Lineage composition of PBSC graft groups

The assignment of PBSC grafts into groups was conditional on cellular composition. Differences in the proportion of cells by lineage were therefore expected. However, the GMM does not indicate which differences were important to cluster assignment. To infer the defining compositional characteristics of each cluster, the PHATE embedding (**Figure 3.9A**) and lineage proportions (**Figure 3.9B**) were visualised stratified by PBSC graft group and the distributions were compared.

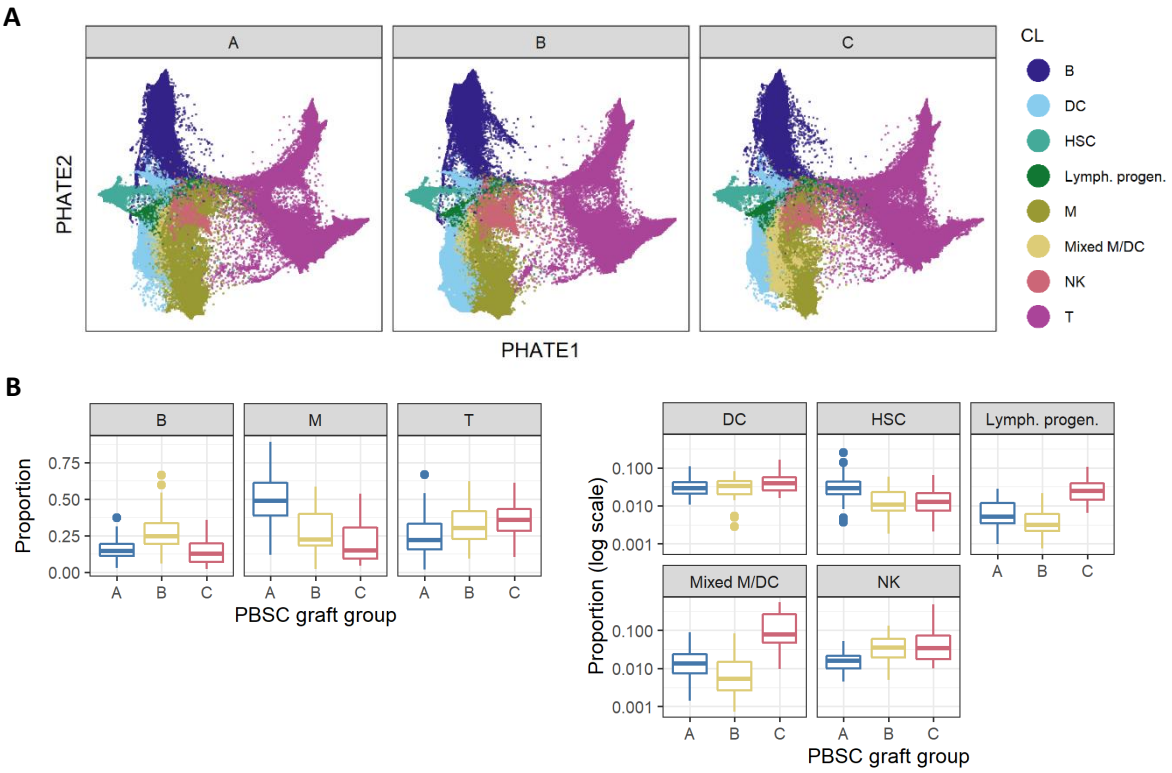


Figure 3.9 Lineage composition by PBSC graft group. (A) PHATE visualisation of cell lineages by PBSC graft group. (B) The proportion of cells in each PBSC graft group by lineage, boxes represent the median and IQR, whiskers extend to the lowest/highest datum within 1.5 IQR of the lower/upper quartile, • indicate outliers.

The proportion of T cells and monocytes showed the greatest variability between the three groups. Grafts in Group A had the highest proportion of monocytes (median 0.488; IQR 0.387 - 0.609) and lowest proportion of T cells (median 0.219; IQR 0.156 - 0.332). The opposite was the case for Group C which had the lowest proportion of monocytes (median 0.147; IQR 0.091 - 0.303) but greatest proportion of T cells (median 0.358; IQR 0.284 - 0.431). In contrast, Group B was in between the other groups for both monocytes (median 0.225; IQR 0.184 - 0.399) and T cells (median 0.301; IQR 0.226 - 0.419) by proportion of cells.

Group A contained a relatively low proportion of NK cells (median 0.016; IQR 0.010 - 0.021) compared with both groups B (median 0.035; IQR 0.019 - 0.059) and C (median 0.034; IQR 0.017 - 0.074), whilst containing the greatest proportion of HSCs (median 0.029; IQR 0.020 - 0.043). The proportion of HSCs was similar in groups B (median 0.011; IQR 0.007 - 0.023) and C (median 0.013; 0.007 - 0.022). However, Group C had a relative abundance of lymphoid progenitor cells (median 0.025; IQR 0.014 - 0.039) and cells with a mixed monocyte and DC like phenotype (median 0.078; IQR 0.047 - 0.263).

The proportion of DC cells was comparable between groups but slightly higher in Group C (median 0.039; 0.025 - 0.056) compared with A (median 0.029; IQR 0.021 - 0.042) and B (median 0.033; IQR

0.020 - 0.044). Notably, there was a substantial proportion of B cells in Group B (median 0.245; IQR 0.193 - 0.333) but relatively few in Groups A (median 0.144; IQR 0.110 - 0.194) and C (median 0.125; IQR 0.070 - 0.197).

These comparisons indicate that the pro-survival Group A was characterised by a relative abundance of monocytes and HSCs, but a lower proportion of T and NK cells. Group B, with a high TRM incidence, was associated with a greater proportion of lymphoid cells including T cells, B cells and NK cells. Whilst Group C, which had a high TRM and GvHD incidence, had the greatest proportion of T cells but the lowest proportion of monocytes.

3.8 Characterisation of cluster phenotypes: cell subsets

The composition of grafts was shown to been highly variable when compared by cell lineage (**Section 3.6**). This variability was sufficient to define three distinct PBSC graft groups with cluster analysis (**Section 3.7**). However, the phenotypic and functional diversity of cells within each lineage was not taken into account. Therefore, clusters were further characterised into finer-grained cell subsets (CS) based on lineage-specific characteristics and compared between PBSC graft groups.

3.8.1 T cells

3.8.1.1 Classification of T cell subsets

The CD3+ T cell clusters (cluster numbers 10, 11, 12, 14, 16, 17, 19, 24, 28, 29, 33, 35 and 36) were characterised into T cell subsets.

Alpha-beta (α/β) and gamma-delta (γδ) T cells were determined by the expression of the γδ T-cell receptor (TCR). One cluster, Cluster 35, was distinctly γδ TCR^+ (**Appendix Figure C.9**). The remaining γδ TCR^- clusters were divided into $CD4^+$, $CD8^+$, double positive or double negative cell subsets. All but two clusters were broadly $CD4^+$ or $CD8^+$ (**Figure 3.10**). Cluster 33 was double positive (DP) whereas Cluster 36 was double negative (DN).

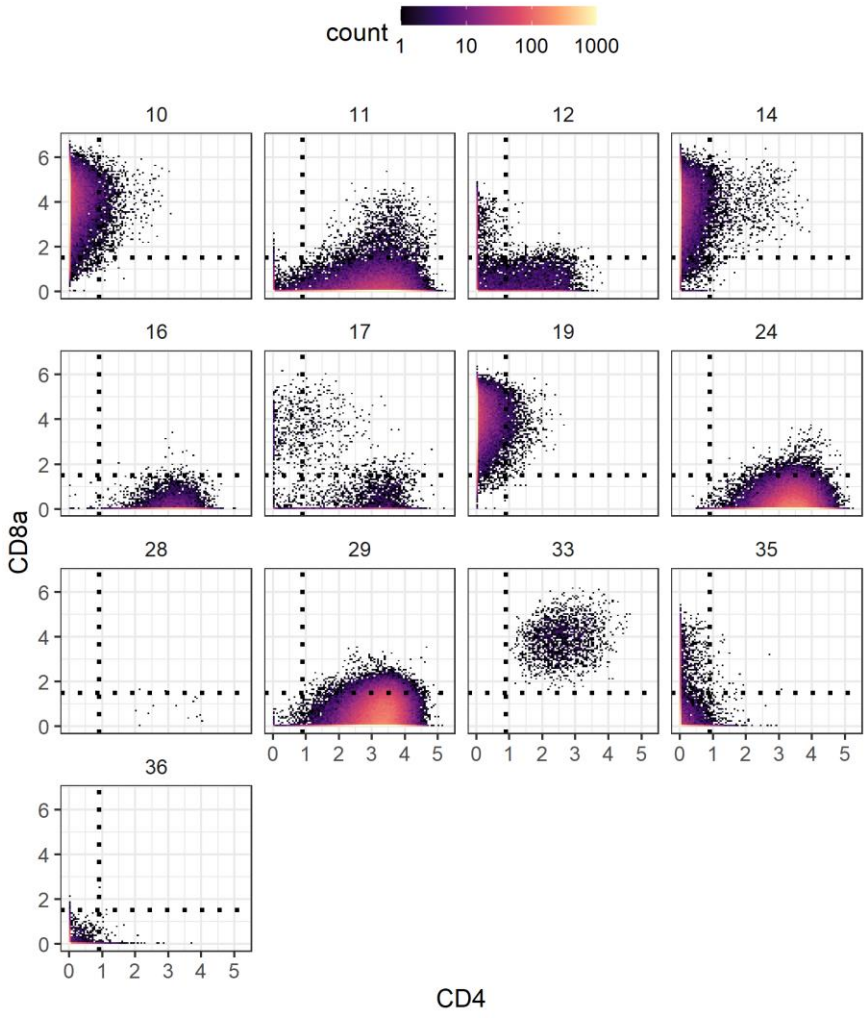


Figure 3.10 T cell clusters - CD8 & CD4 expression. Bi-axial plots showing marker intensity by cell count across clusters identified as T cells. Dashed lines indicate a qualitative threshold to aid visual interpretation.

The $CD4^+$ clusters were labelled as regulatory (T-reg) or non-regulatory or T cells based on the expression of IL-7RA and CD25. Cluster 16 was $CD25^+$ $IL\text{-}7RA^-$ indicating a T-reg phenotype (**Appendix Figure C.10**). A small number of T-regs were also observed within Cluster 29.

The memory status of $CD8^+$ and non-regulatory $CD4^+$ T cell clusters was determined by the proportion of cells expressing CD45RA and CCR7 (**Figure 3.11**). An intensity threshold of 2 was selected qualitatively for each marker. $CD45RA^-$ $CCR7^-$ cells were labelled effector memory (EM), $CD45RA^+$ $CCR7^-$ cells as effector memory cells re-expressing CD45RA (EMRA), $CD45RA^-$ $CCR7^+$ cells were considered central memory (CM) and $CD45RA^+$ $CCR7^+$ were labelled naïve (**Appendix Figure C.11**).

Cluster 11 contained mostly EM cells. Cells in Cluster 12 had a CM-like phenotype and clusters 19 & 29 were mostly Naïve T cells. Cluster 14 contained a mixture of EM and EMRA cells, the same was true for Cluster 10, however, a small proportion of CM cells were also observed. Cluster 24 consisted of a mixture of both EM and CM cells. Whereas Cluster 17, which contained a number of $CD3^+$ $CD14^+$ doublets, did not have a definable memory status. Lastly, Cluster 28 was not labelled due to the low number of cells (N = 44).

Clustering with the PhenoGraph algorithm appeared to have successfully divided the main T cell subsets ($CD3^+$ $\gamma\delta TCR^-$ $CD4^+$; $CD3^+$ $\gamma\delta TCR^-$ $CD8^+$, $CD3^+$ $\gamma\delta TCR^+$) and was able to separate T-regs from other $CD4^+$ cells. Additionally, two small groups of $CD3^+$ DP and $CD3^+$ DN cells were identified. Whilst Naïve and memory T cells were divided, PhenoGraph did not accurately separate memory phenotypes.

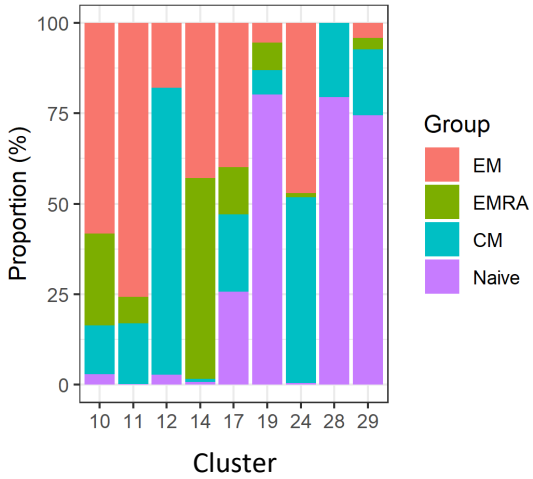


Figure 3.11 Proportion of cells by memory status in CD8+ and non-regulatory CD4+ T cell clusters. The proportion of T cells characterised as naïve, effector memory (EM), effector memory cells re-expressing CD45RA (EMRA), and central memory (CM) by cluster.

3.8.1.2 Composition of T cell compartment

The cell subset classification and defining phenotypic characteristics of each T cell cluster are described in **Table 3.13**. As a proportion of cells assigned to the T cell lineage, 57.9% were CD4+, 35.5% were CD8+ and 5.2% were TCR γδ+. This suggests a CD4/CD8 ratio of 1.6:1 across all cells in the CyTOF data. The remainder of T cells were either double negative for CD4 and CD8 (1.1%) or double positive (0.4%). Of the CD4+ T cells, 3.06% were T-regs.

Table 3.13 Cell subset classification of T cell clusters. The cell count (N.), cell subset (CS), associated lineage markers, and phenotype of each cluster (C).

C	Cells (N)	CS	Lineage marker	Phenotype
10	159,085	T CD8 mixed memory	$CD3^+$	$CD8^+$ $TIGIT^+$ $PD\text{-}1^+$
11	144,854	T CD4 EM	$CD3^+$	$CD4^+$ $c\text{-}Kit^+$ $PD\text{-}1^+$ $DNAM\text{-}1^+$
12	25,453	T CD4 CM	$CD3^+$	$CD4^+$ $PD\text{-}L2^+$
14	131,806	T CD8 EM / EMRA	$CD3^+$	$CD8^+$ $TIGIT^+$ $DNAM\text{-}1^+$ $NKG2C^+$
16	21,654	T-reg	$CD3^+$	$CD4^+$ $CD27^+$ $CD39^+$ $TIGIT^+$ $CD25^+$ $CD127^-$
17	8,394	T CD4 / T cell-monocyte complex	$CD3^+$ $CD14^+$	$CD4^+$
19	142,822	T CD8 Naïve	$CD3^+$	$CD8^+$ $CD31_{dim}$ $CXCR4^+$
24	242,513	T CD4 EM / CM	$CD3^+$	$CD4^+$ $CD27^+$ $DNAM\text{-}1^+$
29	265,224	T CD4 Naïve	$CD3^+$	$CD4^+$ $CD27^+$ $CXCR4^+$
33	4,345	T DP	$CD3^+$	$CD8^+$ $CD4^+$ $CD27^+$
35	63,608	γδ T	$CD3^+$	$TCR\gamma\delta^+$ $NKG2A^+$ $TIGIT^+$
36	13,047	T DN	$CD3^+$	$CD8^-$ $CD4^-$ $CD45^-$

As shown in **Section 3.7.3**, the overall proportion of T cells was lowest in PBSC graft Group A, elevated in Group B and highest in Group C. However, the difference was not uniform across CD4+ (**Figure 3.12A**, **Figure 3.12B**) and CD8+ (**Figure 3.12C**, **Figure 3.12D**) T cell clusters.

In Group C, Cluster 11 (T CD4+ EM) contributed 7.1% of cells but <3% of cells in groups B and A. Differences were also observed in some of the very small T cell clusters, in particular Cluster 36 (T DN) comprised 0.4% of Group C cells but 0.08% of Group B, and 0.17% of Group A.

In Group B, Cluster 29 (T CD4 Naïve) contributed 7.0% of cells but only 4.5% and 3.8% of cells in groups C and A, respectively. Cluster 33 (T DP) was also elevated in Group B alone with 0.1% of cells, compared with 0.06% of cells in the other two groups.

Whereas in pro-survival Group A, T cell cluster proportions were almost uniformly lower than the other two groups with the exception of Cluster 17 (T CD4 / T cell-monocyte complex), Cluster 33 (T DP) and Cluster 36 (T DN).

This indicates the increased proportion of T cells observed in Groups B and C relative to Group A were driven by distinct subsets within the T cell compartment. Group C had an elevated proportion of Cluster 11 (T CD4+ EM) cells, whereas Group B had a greater proportion of Cluster 29 (T CD4 Naïve) cells. Differences in rarer cell subsets were also observed such as Cluster 36 (T DN) being overrepresented in Group C, and Cluster 33 (T DP) in Group B.

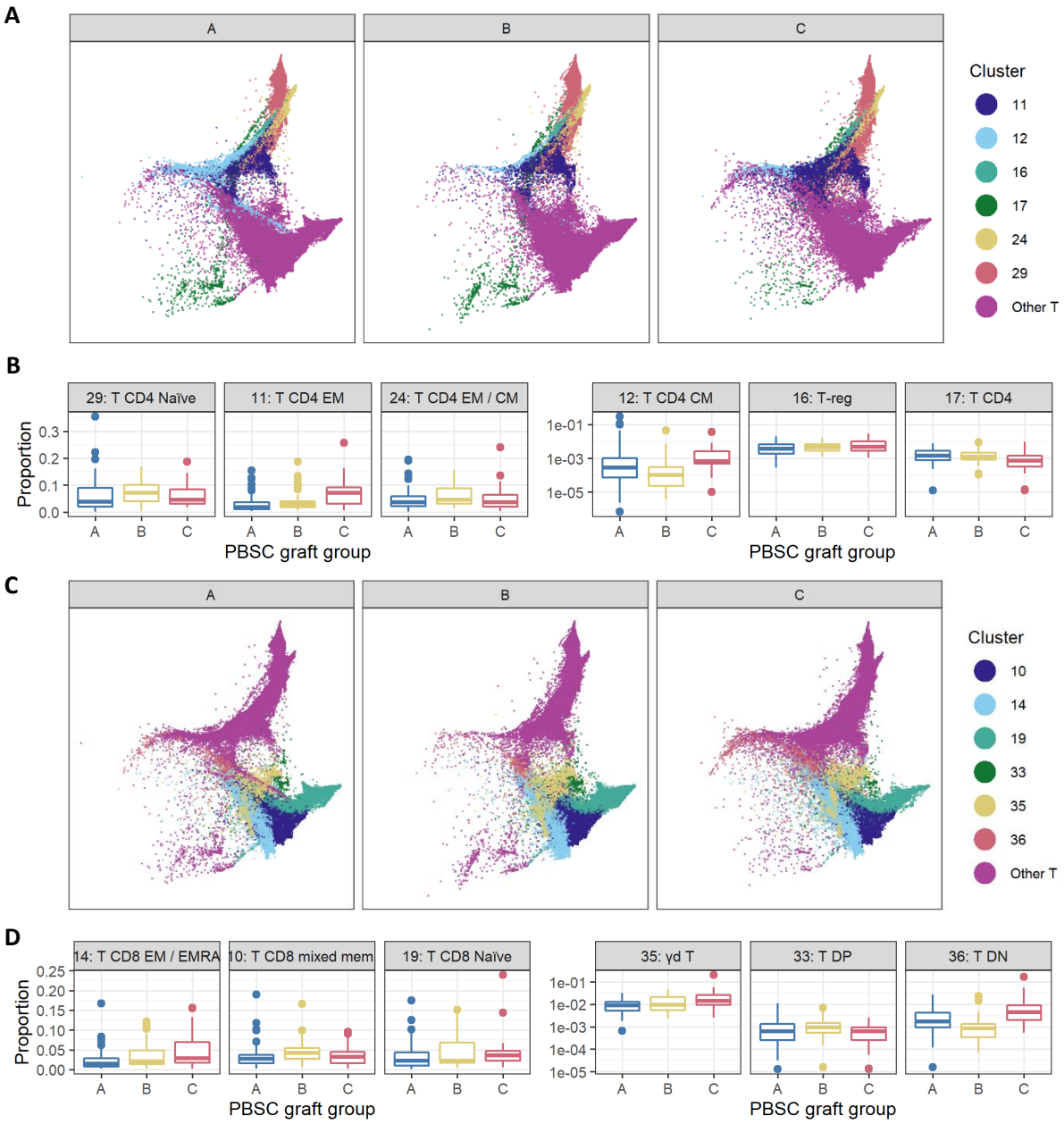


Figure 3.12 Differences in T cell subset composition by PBSC graft group. (A) PHATE visualisation of CD4+ T cell clusters by PBSC graft group. (B) The proportion of cells in each PBSC graft group by CD4+ T cell cluster. (C) PHATE visualisation of CD8+ T cell clusters by PBSC graft group. (D) The proportion of cells in each PBSC graft group by CD8+ T cell cluster. Boxes represent the median and IQR, whiskers extend to the lowest/highest datum within 1.5 IQR of the lower/upper quartile, • indicate outliers.

3.8.2 B cells

3.8.2.1 Classification of B cell subsets

The CD19$^+$ B cell clusters (cluster numbers 3,18,23,25,37) were divided into B cell specific cell subsets. The memory status of CD19$^+$ clusters was determined by the proportion of IgD and CD27 expressing cells. An intensity threshold of 1 was selected qualitatively for each marker. IgD$^+$ CD27$^-$ cells were labelled naïve, IgD$^+$ CD27$^+$ were labelled unswitched memory, and IgD$^-$ CD27$^+$ cells were considered switched memory(373). Cells that were IgD$^-$ CD27$^-$ were labelled double negative (DNEG) (**Figure 3.13A**).

Clusters 3, 25 and 37 contained mostly naïve B cells (**Figure 3.13B**). This was also the case for Cluster 23 but a small proportion of switched and unswitched cells were also observed. Cluster 18 contained very few naïve cells and was enriched in CD25 / IL-2R indicating a mature memory B cell phenotype (374).

The PhenoGraph clustering algorithm successfully divided naïve B cells (CD19$^+$ IgD$^+$ CD27-) from memory B cells. However, unswitched and switched memory B cells were grouped into the same cluster.

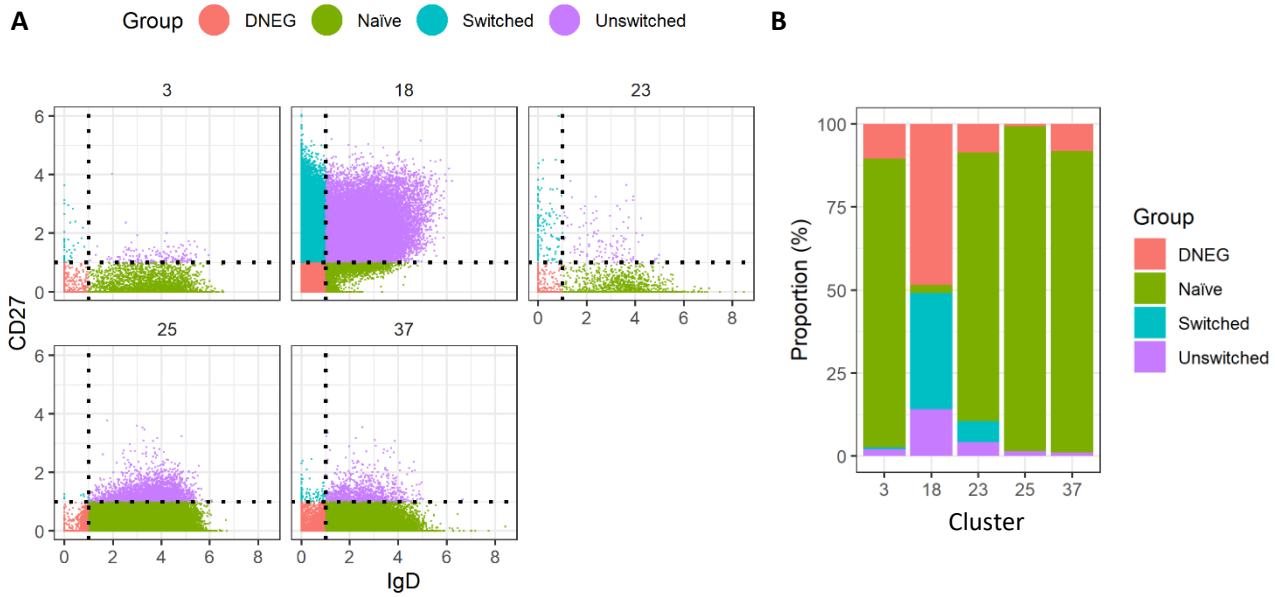


Figure 3.13 B cell clusters - CD27 & IgD expression. (A) Bi-axial plots showing marker intensity of cells in B cell clusters. Dashed lines indicate the threshold selected to determine memory status. (B) The proportion of B cells characterised as naïve, switched memory, unswitched memory, and CD27-IgD- (DNEG) by cluster.

3.8.2.2 Composition of B cell compartment

The cell subset classification and defining phenotypic characteristics of each B cell cluster are described in **Table 3.14**. As a proportion of cells assigned to the B cell lineage, 71.2% were naïve B cells (IgD$^+$ CD27$^-$), with the remaining 28.8% of B cells exhibiting a switched or unswitched memory phenotype (B memory).

Table 3.14 Cell subset classification of B cell clusters. The cell count (N.), cell subset (CS), associated lineage markers, phenotype and B cells memory markers association with each cluster (C).

C	Cells (N)	CS	Lineage marker	Phenotype	B memory markers
3	13498	B Naïve	$CD19^+$	$HLA\text{-}DR^+$ $CD10^+$ $CD45RA^+$ NKP46 int.	IgD^+ $CD27^-$
18	226262	B Memory	$CD19^+$	$CD25^+$ $CD45RA^+$	IgD^+ $CD27^+$ / IgD^- $CD27^+$ / IgD^- $CD27^-$
23	4038	B Naïve	$CD19^+$	$HLA\text{-}DR^+$ $CD11b^+$ $NKP46^+$ $CD39^+$	IgD^+ $CD27^-$
25	374077	B Naïve	$CD19^+$	$HLA\text{-}DR^+$ $CD10^+$ $CD45RA^+$ $NKP46^+$ $CCR7^+$ $CXCR4^+$	IgD^+ $CD27^-$
37	168645	B Naïve	$CD19^+$	$HLA\text{-}DR^+$ $CD10^-$	IgD^+ $CD27^-$

PBSC graft Group B, which had a high rate of TRM compared with Group A, was associated with a particularly high proportion of B cells. However, this difference was not observed uniformly across B cell clusters (**Figure 3.14A, Figure 3.14B**). In Group B, 12.5% of all cells were Cluster 25 (B Naïve $CD10^+$ $CD45RA^+$ $NKP46^+$ $CCR7^+$ $CXCR4^+$) cells, but this cluster contributed <6% of cells in the other two groups. The proportion of Cluster 18 (B memory $CD25^+$ $CD45RA^+$) was also far higher in Group B (7.1%) compared with Group A (4%) and Group C (3.8%). Cluster 3, which was relatively small (13,498 cells), was also elevated in Group B (0.5%). But there was little difference in the proportion of Cluster 37 (B Naïve $CD10^-$) and Cluster 23 (B Naïve $CD11b^+$ $NKP46^+$ $CD39^+$) between groups.

This indicates the increased proportion of B cells in PBSC graft Group B was mostly determined by an increase in the proportion of $CCR7^+$ $CXCR5^+$ Naïve B cells and memory B cells.

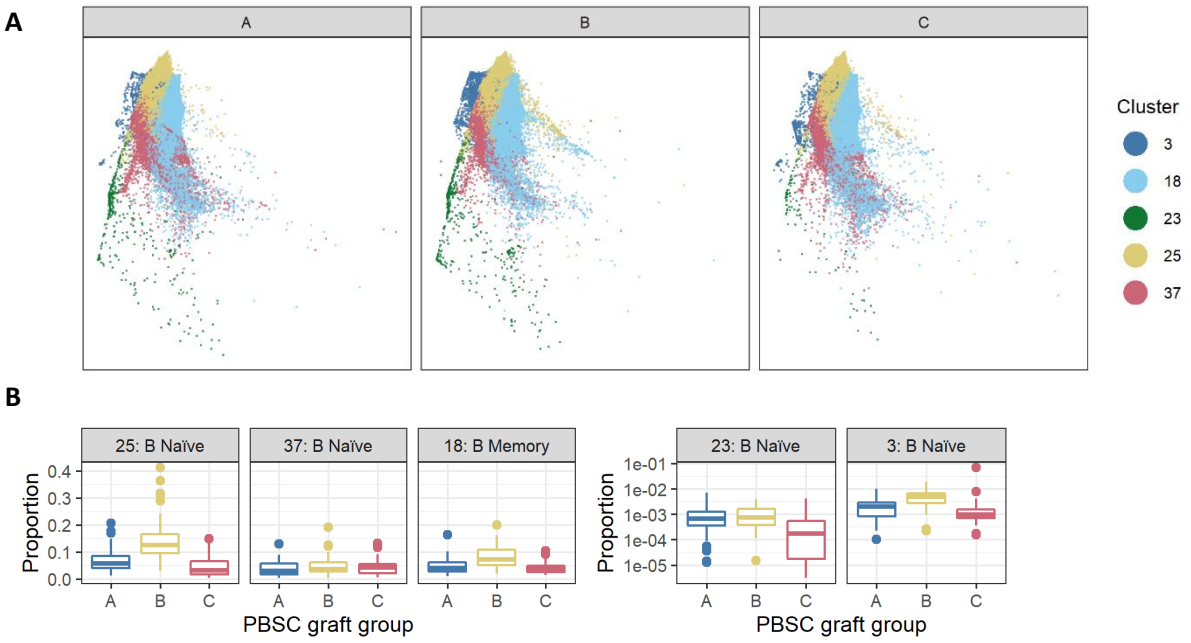


Figure 3.14 Differences in B cell subset composition by PBSC graft group. (A) PHATE visualisation of B cell clusters by PBSC graft group. (B) The proportion of cells in each PBSC graft group by B cell cluster. Boxes represent the median and IQR, whiskers extend to the lowest/highest datum within 1.5 IQR of the lower/upper quartile, • indicate outliers.

3.8.3 Monocytes

3.8.3.1 Classification of monocyte subsets

CD14^{+} monocytes are typically classified into classical (CD14^{+} CD16^{-}), intermediate (CD14^{+} CD16^{+}) and non-classical (CD14$_{dim}$ CD16^{+}) subsets (364). Classical monocytes make up the majority of circulating monocytes (~85-90%) and differentiate through the intermediate (~5%) to the non-classical subsets (~10%) (375-377). These phenotypes therefore exist as a continuum rather than discrete subsets, and as a result there is little consensus in gating monocytes in single cell analyses (377).

Clusters 5, 6, 7, 8 and 30 were consistently CD14^{++} or CD14$_{int}$ (**Figure 3.15**). With the exception of Cluster 8, these clusters were also CD16^{-} but also contained CD16$_{dim}$ cells. Collectively they represented over of half cells assigned to the monocyte lineage (54.2%) so are likely of a classical phenotype, however this can't be stated conclusively due to the variable CD16 expression. Clusters 7 and 30 were particularly bright for CD14 and CD11b (**Figure 3.16**). Cluster 7 was also CD11c^{++} and relatively enriched for the activation marker CD45RA (378) and the co-stimulatory molecule CD86 (375). Clusters 5 and 8 were characterised by intermediate (int.) CD14 and CD11b expression, with Cluster 8 being CD11c^{-}. Cluster 6 was the only classical-like cluster of monocytes that was HLA-DR^{+}. Collectively, these were assigned a label of classical-like monocytes.

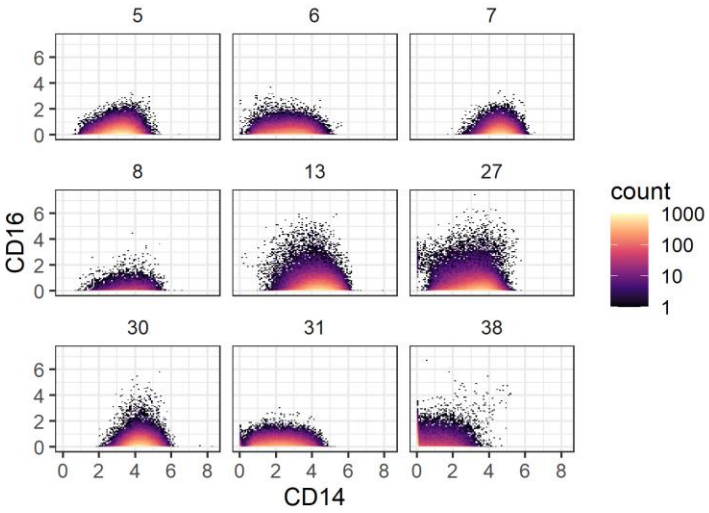


Figure 3.15 Monocyte clusters - CD14 & CD16 expression. Bi-axial plots showing marker intensity by cell count across monocyte clusters.

Cluster 13 contained a number of monocytes with intermediate (CD14^{+} CD16^{+}) and classical (CD14^{+} CD16^{-}) phenotypes. Cells of this cluster were consistently HLA-DR^{+} and enriched for CD11c and CD86 (**Figure 3.16, Figure 3.2**). Elevated HLA-DR, CD11c and CD86 expression is associated with intermediate monocytes further supporting this classification (364). Cluster 27 had cells corresponding to all three monocyte subsets including non-classical (CD14$_{dim}$ CD16^{+}). This cluster was also CD11b^{+} HLA-DR^{-} and relatively enriched for CD45RA (**Figure 3.16, Figure 3.2**).

Clusters 31 & 38 were outliers as they were CD14$_{dim}$ CD16^{-} and so did not conform to the expected monocyte subsets. Both were characterised by relatively low expression of CD14, CD11b, and CD45. However, Cluster 38 was also highly enriched in CCR7, associated with monocyte migration (379), and PD-L2 (**Figure 3.2**).

This suggests that PhenoGraph did not distinguish monocytes based on the prototypical monocyte subsets (classical, intermediate and non-classical). CD14^{+} clusters appeared to be divided based on continuous differences in marker expression. As would be expected, the majority of monocyte clusters

were associated with a classical-like phenotype. One cluster, Cluster 13, contained both intermediate and classical monocytes and exhibited intermediate like traits such as high HLA-DR, CD11c and CD86 expression. Whereas non-classical monocytes could not be distinguished by cluster.

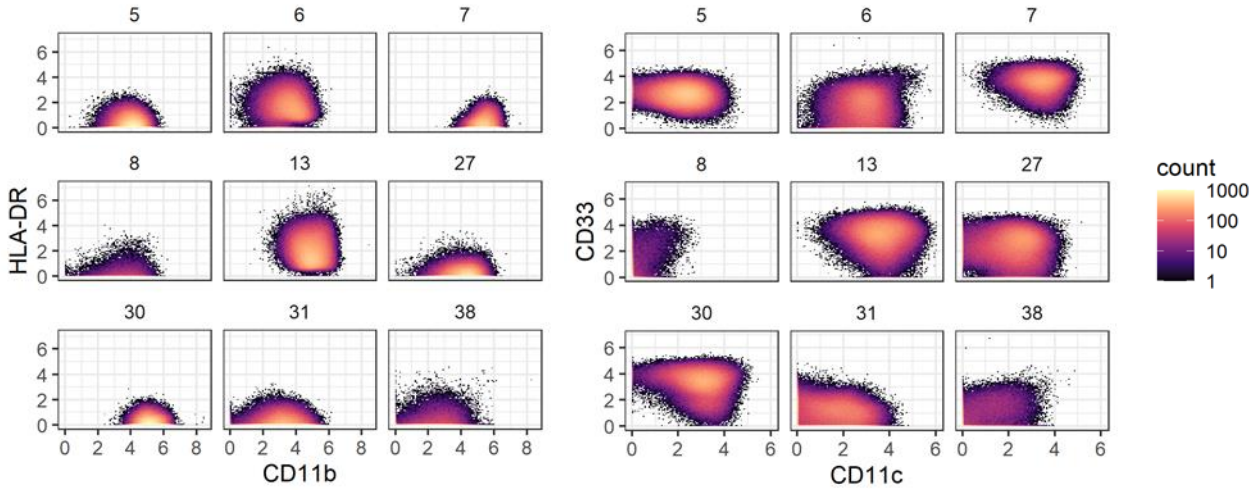


Figure 3.16 Monocyte clusters – HLA-DR, CD11b, CD33 & CD11c expression. Bi-axial plots showing marker intensity by cell count across monocyte clusters.

3.8.3.2 Composition of monocyte compartment

The cell subset classification of monocyte clusters and defining phenotypic characteristics are described in **Table 3.15**. Of cells assigned to the monocyte lineage, 54.2% exhibited a classical-like phenotype, 15.5% exhibited a mixture of intermediate and classical monocyte traits and a further 14.6% were of a mixture of all three subsets including non-classical. Additionally, 15.6% of monocytes exhibited a $CD14_{dim}$ $CD16^-$ phenotype.

Table 3.15 Cell subset classification of monocyte clusters. The cell count (N.), cell subset (CS), associated lineage markers, and phenotype of each cluster (C).

C	Cells (N)	CS	Lineage marker	Phenotype
5	327877	Classical-like M	$CD14^+$	CD11b int. $CD16^-$*
6	177214	Classical-like M	$CD14^+$	HLA-DR^+ $CD16^-$*
7	165873	Classical-like M	$CD14^+$	CD11b/$CD11c^{++}$ $CD45RA^+$ $CD86^+$ $CD16^-$*
8	27397	Classical M	$CD14^+$	$CD11c^-$ PD^-L2^+ $CD16^-$
13	267396	Classical & Intermed. M	$CD14^+$	CD11c/HLA-DR^{++} $CD86^+$ $CD16^{\pm}$
27	251728	Classical, Intermed. & Non-classical M	$CD14^+$	$CD11b^+$ HLA-DR^- $CD45RA^+$ $CD16^{\pm}$
30	233504	Classical-like M	$CD14^+$	$CD11b^{++}$ $CD16^-$*
31	213644	$CD14_{dim}$ M	$CD14_{dim}$	CD11b/CD45/$CD33_{dim}$ $CD16^-$
38	54433	$CD14_{dim}$ M	$CD14_{dim}$	CD11b/CD45/$CD33_{dim}$ $CD16^-$ $CCR7^{++}$ PD^-L2^+

* Clusters that were mostly $CD16^-$ but contained a few $CD16^+$ or $CD16_{dim}$ cells

The overall proportion of monocytes was highest in PBSC graft Group A, substantially lower in Group B and lowest in Group C (**Section 3.7.3**). The relative increase in Group A was observed uniformly across monocyte clusters with two exceptions (**Figure 3.17A, Figure 3.17B**). The proportion of Cluster 7 (Classical-like M CD11b/$CD11c^{++}$ $CD86^+$) was highest in Group B with 3.7% of cells compared with 1.9% in Group A and 0.26% of Group C. Whereas Cluster 38 ($CD14_{dim}$ M $CCR7^{++}$ PD-$L2^+$) contributed 2.6% of cells in Group C, but <0.7% of cells in the other groups.

Certain classical-like monocyte clusters contributed a considerable proportion of all cells assigned to the PBSC graft Group A, whilst being relatively rare in the other two graft groups. Cluster 5 (Classical-like M CD11b $_{int.}$) contributed 9.5% of cells in Group A but <3% of cells in groups B and C. Similarly, the proportion of Cluster 6 (Classical-like M HLA-DR^{+}) was more than double in Group A with 5.3% of cells, compared with <1.7% of cells in the other two groups. A comparable pattern was observed in the proportion of Cluster 30 (Classical-like M CD11b^{++}) cells.

Notably, Cluster 13 (Classical & Intermed. M CD11c/HLA-DR^{++} CD86^{+}) made up >4% of all cells in groups A and B, was comparatively rare in Group C (0.7%).

This indicates the relative abundance of monocytes in pro-survival Group A was almost uniform across monocyte clusters, and primarily driven by an increase in classical-like monocytes. There were some exceptions, Group C had the highest proportion of Cluster 38 (CD14$_{dim}$ M CCR7^{++} PD^{-}L2^{+}) with almost no Cluster 7 cells (Classical-like M CD11b/CD11c^{++} CD86^{+}), whereas Group B had the highest proportion of Cluster 7 cells.

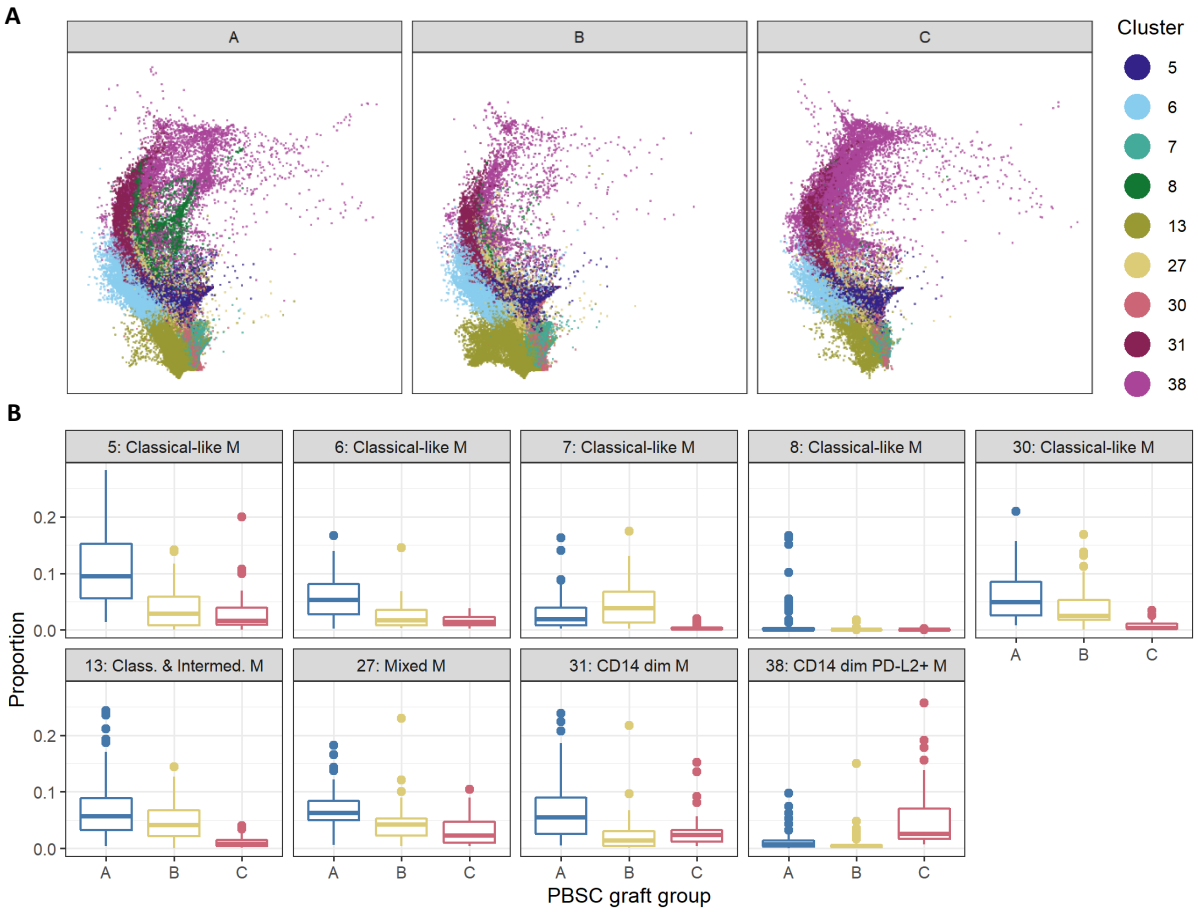


Figure 3.17 Differences in monocyte cell subset composition by PBSC graft group. (A) PHATE visualisation of monocyte clusters by PBSC graft group. (B) The proportion of cells in each PBSC graft group by monocyte cluster. Boxes represent the median and IQR, whiskers extend to the lowest/highest datum within 1.5 IQR of the lower/upper quartile, • indicate outliers.

3.8.4 Dendritic cells

3.8.4.1 Classification of DC subsets

The $CD14^-$ $HLA\text{-}DR^+$ DC clusters (cluster numbers 1, 4, 22, 2) were assigned to DC specific cell subsets. Cluster 2, which contained a mix of antigen presenting cells (DCs and monocytes), was also examined. Myeloid DCs (i.e., conventional DCs) were distinguished from plasmocytoid DCs (pDCs) by lacking CD11c, CD33 and CD11b expression. Cluster 22 was $CD11c^-$ $CD33^-$ $CD11b^-$ and enriched for CD45RA corresponding to a pDCs phenotype (380) (**Appendix Figure C.12**).

Within the myeloid DC compartment, classical DCs (cDCs) were distinguished from non-classical DCs by the absence of CD16 expression. Clusters 2 and 4 were both $CD16^-$ indicating cDCs. Cluster 1 contained both $CD16^+$ and $CD16^-$ cells and was enriched for the activation markers CD31 and CD86, suggesting a mix of non-classical and cDCs but with an activated phenotype.

Classical DCs, such as those within Clusters 2 and 4, can be further subdivided into cDC1 ($CD11c_{dim}$ $CD11b^-$ $CD141^+$) (381) and cDC2 ($CD11c^+$ $CD11b^+$ $CD1c^+$) subsets, however CD1c was not available in the CyTOF panel (**Table 3.2**). Cluster 4 was $CD11c^+$ $CD11b_{dim}$ $CD141^+$ which suggests a cDC1 phenotype. Cluster 2 was $CD11b^+$ so may contain cDC2 DCs, but in the absence of CD1c a conclusive classification could not be determined.

The PhenoGraph algorithm successfully divided myeloid DCs from pDCs. Within the myeloid DC compartment, only cDC1 DCs were readily distinguishable by cluster. A cluster of myeloid DCs (classical and non-classical) enriched for activation markers CD31 and CD86 was also identified.

3.8.4.2 Composition of DC compartment

The classification of DC clusters and defining phenotypic characteristics are described in **Table 3.16**. Of cells assigned to the DC lineage 33.7% were pDC, in line with reports that pDCs are the most numerous singular DC cell subset in human blood (380). 21.9% were cDC1 and 44.4% had an activated phenotype with a mixture of myeloid DC traits.

Table 3.16 Cell subset classification of DC clusters. The cell count (N.), cell subset (CS), associated lineage markers, and phenotype of each cluster (C).

C	Events / Cells (N)	CS	Lineage marker	Phenotype
1	76673	Activated DCs	$HLA\text{-}DR^+$	$CD11c^{++}$ $CD11b^+$ $CD16^{+/-}$ $CD31^+$ $CD86^+$ (Non-classical DC & cDC)
4	37797	cDC1	$HLA\text{-}DR^+$	$CD11c^+$ $CD11b^-$ $CD141^+$ $CD16^-$ $CD45RA^-$ (cDC1)
22	58167	pDC	$HLA\text{-}DR^+$	$CD11c^-$ $CD33^-$ $CD11b^-$ $CD16^-$ $CD45RA^+$
2	251417	Classical M & cDC	$CD14^{dim}$ & $CD14^-$ $HLA\text{-}DR^+$	$CD11c^{+/}{}_{dim}$ $CD11b^+$ $CD16^-$

No substantial difference in the proportion of DC cells was observed when compared between PBSC graft groups. However, differences were apparent when compared by cluster (**Figure 3.18**).

Cluster 1 (Activated DCs $CD31^+$ $CD86^+$) cells were relatively rare in Group A (0.7%), compared with Group B (1.4%) and Group C (1.8%). Whereas Cluster 22 (pDC) cells were more common in Group A with 1.2% of cells but only 0.8% in the other two groups.

This suggests that whilst there was little difference in the overall proportion of DCs between the PBSC graft groups, there was a relative abundance of pDCs and paucity of activated DCs in pro-survival Group A when composition was compared by cluster.

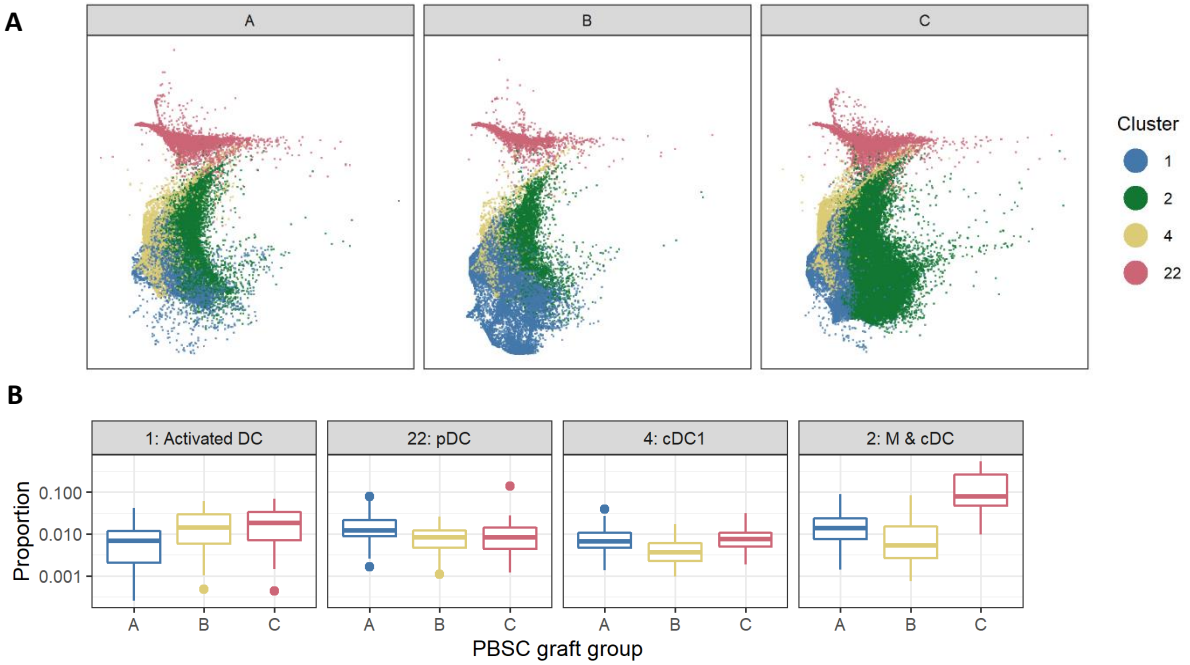


Figure 3.18 Differences in DC subset composition by PBSC graft group. (A) PHATE visualisation of DC clusters by PBSC graft group. (B) The proportion of cells in each PBSC graft group by DC cluster. Boxes represent the median and IQR, whiskers extend to the lowest/highest datum within 1.5 IQR of the lower/upper quartile, • indicate outliers.

3.8.5 NK cells

The NK & NKT lineages were represented by a single cluster of cells identified by the PhenoGraph algorithm. This section provides additional detail on the phenotypic characteristics of these clusters.

3.8.5.1 Classification of NK subsets

$CD56^{+}$ $CD3^{-}$ NK cells are typically gated into $CD56^{++}$ $CD16^{+/-}$ ($CD56^{++}$ NK) and $CD56_{dim}$ $CD16^{++}$ ($CD56_{dim}$ NK) subsets(382). Only one population of NK cells was identified by PhenoGraph, Cluster 15. This cluster contained a mix of both $CD56^{++}$ NK cells and $CD56_{dim}$ NK cells (**Figure 3.19**). NKG2A expression is expected to be higher in $CD56^{++}$ cells compared with $CD56_{dim}$ (383), a pattern which was visible across Cluster 15. Cluster 34 showed distinct $CD56^{+}$ $CD3^{+}$ expression supporting the NKT cell classification (**Figure 3.19**). These cells were also enriched for NKG2C, TIGIT & DNAM-1 (**Figure 3.2**).

This indicates that PhenoGraph successfully identified NK cells as a distinct cell lineage. However, it failed to distinguish between $CD56^{++}$ and $CD56_{dim}$ NK cells. NKT cells were also clearly delineated from NK and T cells.

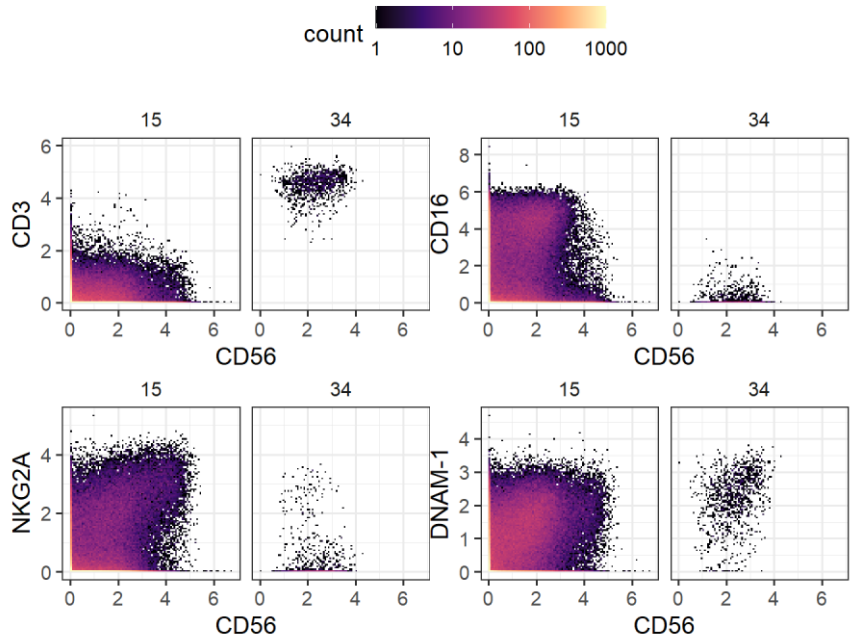


Figure 3.19 NK & NKT clusters - CD56, CD3, CD16, NKG2A & DNAM-1 expression. Bi-axial plots showing marker intensity by cell count across NK and NKT clusters.

3.8.5.2 Composition of NK compartment

The cell subset classification and defining phenotypic characteristics of the NK & NKT clusters are described in **Table 3.17**. As described earlier in this chapter (**Section 3.7.3**), PBSC graft Group A was characterised by the lowest proportion of NK cells (1.6%) (**Figure 3.9B**). The NKT cluster contributed ~0.01% of cells each graft group.

Table 3.17 Cell subset classification of NK & NKT clusters. The cell count (N.), cell subset (CS), associated lineage markers, and phenotype of each cluster (C).

Cluster	Cells (N)	CS	Lineage marker	Phenotype
15	134709	NK	$CD56^{+}$	$CD56^{+}$ $NKG2A^{+}$ $CD16^{\pm}$
34	1115	NKT	$CD56^{+}$ $CD3^{+}$	$CD56^{+}$ $CD3^{+}$ $NKG2C^{+}$ $DNAM\text{-}1^{+}$

3.8.6 Lineage negative cells

The HSC, lymphoid progenitor, and CLP lineages were each represented by a single cluster of cells identified by the PhenoGraph algorithm. This section provides additional details on the phenotypic characteristics of these clusters.

3.8.6.1 Classification of lineage negative subsets

Lineage negative (Lin^-) and $CD34^+$ clusters were characterised by the intensity of CD34, c-Kit and CD45RA expression and by cluster specific marker enrichment.

Cluster 39 was Lin^- $CD34^+$ c-Kit^+ $CD45RA^-$ (**Figure 3.20**) in line with the expected HSC phenotype. The expression of c-Kit was highly variable ranging from bright to dim. Cluster 32 was Lin^- $CD34^{dim}$ DNAM-1^+ $CXCR4^+$ $CD45RA^-$ $PD\text{-}L2^+$ so retained its lymphoid progenitor label (366, 367).

A small population of (<3000 cells) c-Kit positive CLP cells were identified. Cluster 21, with 2,691 CLP cells, was Lin^- $CD34^-$ c-Kit^+ $CD45RA^+$ IL-$7Ra^+$ IL-$2R^+$ (**Figure 3.20**). The c-Kit and IL-7RA expression aligns with the phenotype of common lymphoid progenitors reported in mice (370, 384).

PhenoGraph successfully identified HSC cells as a distinct population of $CD34^+$ cells. Other types of progenitor cells were also identified, including a small population of CLP cells, and a substantial population of DNAM-1^+ lymphoid progenitor cells.

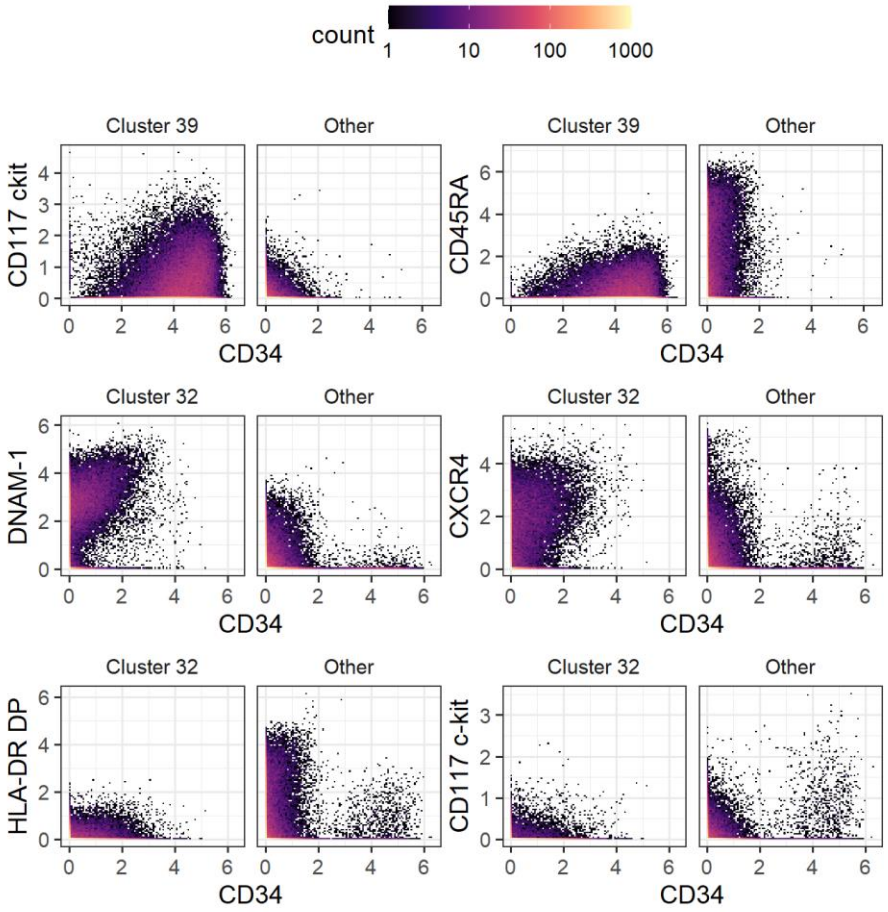


Figure 3.20 Lineage negative cluster – marker expression. Bi-axial plots showing marker intensity by cell count across HSC and lymphoid progenitor clusters.

3.8.6.2 Composition of lineage negative compartment

The cell subset classification and defining phenotypic characteristics of the Lin$^-$ clusters are described in **Table 3.18**. As described earlier in this chapter (**Section 3.6**), Group A was characterised by the highest proportion of HSCs (2.9%) (**Figure 3.21A, Figure 3.21B**). Cluster 32 (Lymph. progenitor) cells were relatively abundant in grafts from Group C (2.5%), whereas the CLP cluster contributed ~0.01% of cells each graft group. To emphasize the location of the lymph. progenitors between the HSC and NK populations, the NK cell populations are also shown on the PHATE embedding (**Figure 3.21A**).

Table 3.18 Cell subset classification of lineage negative clusters. The cell count (N.), cell subset (CS), associated lineage markers, and phenotype of each cluster (C).

C	Cells (N)	CS	Lineage marker	Phenotype
21	2,691	CLP	Lin-	CD34- c-Kit+ CD45RA+ IL-7Ra+ IL-2R+
32	42,216	Lymph. progenitor	Lin-	Lin- CD34dim DNAM-1+ CXCR4+ CD45RA- PD-L2+.
39	105,845	HSC	Lin-	CD34+ c-Kit+ CD45RA-

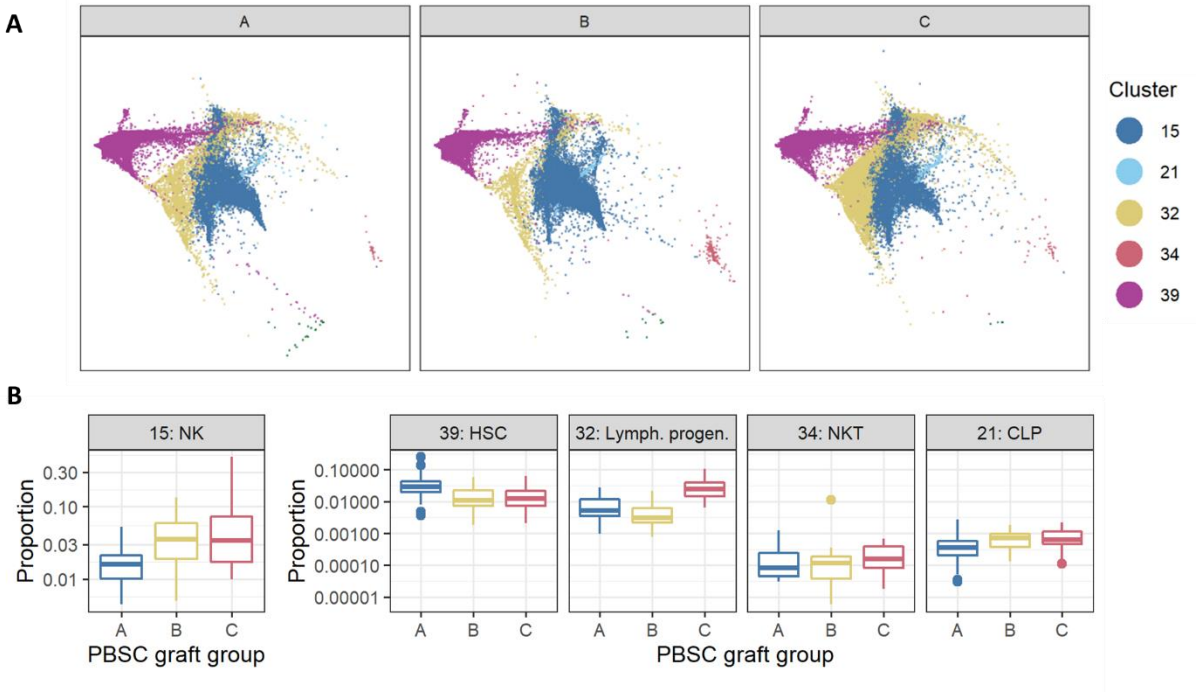


Figure 3.21 Differences lineage negative subset composition by PBSC graft group. (A) PHATE visualisation of Lin- clusters by PBSC graft group, NK clusters shown for comparison. (B) The proportion of cells in each PBSC graft group by cluster. Boxes represent the median and IQR, whiskers extend to the lowest/highest datum within 1.5 IQR of the lower/upper quartile, • indicate outliers.

3.9 Regression of survival and competing events

Three distinct types of PBSC graft were identified by cluster analysis (**Section 3.7**), the recipients of which exhibited significant differences in the probability of experiencing a deleterious event following transplantation. This supports the hypothesis that graft composition and outcome are correlated.

However, the donors were also found to differ in age, sex, CMV status, and in their relationship with the recipient patients. All of which are known to impact the prognosis of allo-HSCT patients (59, 157, 160, 161, 167). Multivariate analyses were therefore conducted to test the effects after adjusting for covariates. Cox PH regression models were developed for survival events, and FG regression models were utilised for competing events.

The three PBSC graft groups were identified based only on differences in cellular composition. Differences between the PBSC graft groups may therefore reflect the most variable cellular elements between donations, rather than the lineages most pertinent to a patient's response to the therapy. Thus, the relationships between the cellular composition of grafts and patient outcomes were also investigated.

To test the effect of graft composition on survival outcomes, a formulation of the Cox PH model compatible with compositional data was utilised (334-336). This approach was extended to the FG model for competing events. Compositions were transformed into unconstrained variables using the additive log-ratio (ALR) transformation prior to regression. I refer the reader to **Section 2.1.6.1** for an introduction to the ALR transformation.

This part of the chapter is structured into sections based on the type of event being investigated. The survival events studied were GvHD-free, relapse-free survival (GRFS) and overall survival (OS) (**Sections 3.9.1-3.9.2**). The competing events analysed were the onset of relapse, transplant-related mortality (TRM), AGvHD, and CGvHD (**Sections 3.9.3-3.9.5**).

Correlation between the exact log-ratio geometry and the ALR transformed data are shown in **Appendix Table C.1.** The events-per-variable ratios for multivariate models are shown in **Appendix Table C.2**.

3.9.1 GvHD-free, relapse-free survival

3.9.1.1 Effects of PBSC graft group on GRFS

The hazard $\lambda(t)$ of GRFS was regressed on clinical variables and PBSC graft groups using the Cox PH model (**Table 3.19**)**.** This tests the association between each variable and the rate of GRFS.

Based on univariate estimates, recipients of PBSC graft group C had a significantly lower rate of GRFS (HR 2.16; P <0.01) compared with group A. HCT-CI scores of 3+ (HR 1.5; P 0.1) were associated with a marginal decrease in the GRFS rate compared with patients with the lowest score. A test for time dependence in Schoenfeld residuals (PH test) was used to check the proportional hazards (PH) assumption for Cox PH models. There was evidence that CD34+ violated the PH assumption (P 0.04), so this variable was excluded from multivariate analysis.

In a multivariate comparison, recipients of PBSC graft Group C were significantly more likely to experience an event following transplant (**Table 3.20**). The probability of GRFS was 120% lower in Group C recipients (adj. HR 2.20; P <0.01) compared with Group A, with no difference in Group B (P 0.12). Higher HCT-CI scores were associated with an insignificant increased in the rate of GRFS (adj. 1.56; P 0.08).

Table 3.19 Univariate effects on GRFS - PBSC graft group. Univariate hazard ratios (HR) from the Cox PH model, with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Term	Ref	Univariate				
		HR	Lower	Upper	P. val	PH test
Age at transplant		1.00	0.99	1.02	0.76	0.14
CD34+ dose		1.10	0.92	1.31	0.28	0.04
DCMV status Pos	Neg	1.05	0.72	1.55	0.79	0.32
Donor age		1.00	0.99	1.02	0.84	0.89
Donor relationship UD	RD	1.09	0.74	1.62	0.66	0.21
Sex mismatch Yes	No	0.94	0.64	1.38	0.75	0.94
HCT-CI 1	0	0.86	0.44	1.68	0.66	0.12
HCT-CI 2	0	1.02	0.56	1.86	0.95	0.12
HCT-CI 3+	0	1.50	0.93	2.42	0.10	0.12
MAC or RIC RIC	MAC	0.98	0.65	1.49	0.94	0.30
Patient sex Female	Male	0.85	0.57	1.27	0.43	0.38
Patient CMV status Pos	Neg	1.26	0.80	1.99	0.32	0.69
PBSC graft group B	A	1.37	0.86	2.17	0.19	0.78
PBSC graft group C	A	2.16	1.35	3.45	<0.01	0.78

Table 3.20 Multivariate effects on GRFS - PBSC graft group. Multivariate hazard ratios (Adj. HR) from the Cox PH model, with 95% confidence interval (lower, upper), and significance test (P. val). P val < 0.1 highlighted in yellow.

Term	Ref	Multivariate			
		Adj. HR	Lower	Upper	P.val
Age at transplant		1.00	0.98	1.02	0.84
DCMV status Pos	Neg	0.95	0.61	1.46	0.80
Donor age		1.00	0.98	1.02	0.90
Donor relationship UD	RD	1.09	0.62	1.92	0.76
Sex mismatch Yes	No	0.89	0.56	1.42	0.63

Term	Ref	Multivariate			
		Adj. HR	Lower	Upper	P.val
HCT-CI 1	0	0.92	0.46	1.84	0.82
HCT-CI 2	0	1.22	0.64	2.31	0.55
HCT-CI 3+	0	1.56	0.95	2.57	0.08
MAC or RIC RIC	MAC	0.92	0.58	1.46	0.72
Patient sex Female	Male	0.92	0.58	1.44	0.71
Patient CMV status Pos	Neg	1.21	0.73	2.01	0.46
PBSC graft group B	A	1.48	0.91	2.40	0.12
PBSC graft group C	A	2.20	1.32	3.65	<0.01

3.9.1.2 Effects of graft composition on GRFS

The association between graft composition and the rate of GRFS was tested with the compositional Cox PH model (**Table 3.21**). Compositional variables were expressed as additive log-ratios (ALR) with the proportion of HSCs as the reference part. Clinical variables were included as additional covariates.

To test the significance of the association between the whole composition and GRFS, a likelihood-ratio (LR) test was applied. This resulted in a p-value of 0.007, therefore I found evidence of significant association between the rate of GRFS and the graft composition, beyond that attributable to other variables.

In respect of the individual parts of the composition, the rate of GRFS decreased significantly as the log-ratio of NK cells to HSCs increased (HR 1.32; P 0.05). The opposite was the case for monocytes, where an increased proportion was associated with improved rates of GRFS (HR 0.68; P 0.01).

Table 3.21 Multivariate effects on GRFS - graft composition. Multivariate hazard ratios (Adj. HR) from the Cox PH model, with 95% confidence interval (lower, upper), and significance test (P. val). Compositional variables expressed as additive log-ratios (alr). P. val < 0.1 highlighted in yellow.

Term	Ref	Multivariate			
		HR	Lower	Upper	P.val
Age at transplant		1.00	0.98	1.02	0.93
DCMV status Pos	Neg	0.89	0.57	1.39	0.60
Donor age		1.00	0.98	1.02	0.98
Donor relationship UD	RD	1.19	0.67	2.11	0.56
Sex mismatch Yes	No	0.85	0.52	1.37	0.49
HCT-CI 1	0	1.03	0.51	2.07	0.94
HCT-CI 2	0	1.50	0.76	2.95	0.25
HCT-CI 3+	0	1.96	1.16	3.33	0.01
MAC or RIC RIC	MAC	0.89	0.55	1.43	0.63
Patient sex Female	Male	0.86	0.54	1.38	0.54
Patient CMV status Pos	Neg	1.22	0.73	2.05	0.45
alr NK vs HSC		1.32	1.00	1.73	0.05
alr T vs HSC		1.14	0.79	1.63	0.49
alr DC vs HSC		1.35	0.93	1.94	0.11
alr Lymph. progen. vs HSC		1.04	0.80	1.35	0.78
alr B vs HSC		0.88	0.66	1.18	0.40
alr Mixed M/DC vs HSC		0.90	0.74	1.10	0.30
alr M vs HSC		0.68	0.53	0.89	0.01

3.9.2 Overall survival

3.9.2.1 Effects of PBSC graft group on OS

The hazard $\lambda(t)$ of OS was regressed on clinical variables and PBSC graft groups using the Cox PH model (**Table 3.22**).

Based on univariate estimates, recipients of PBSC graft groups B (HR 2.03; P 0.03) and C (HR 2.03; P 0.03) had a significantly lower rate of OS compared with group A. HCT-CI scores of 3+ (HR 2.3; P 0.01) were associated with a significant decrease in the OS rate compared to patients with the lowest score.

In a multivariate comparison, PBSC graft retained a significant effect on the survival rate (**Table 3.23**). The rate of OS was 143% lower in Group B recipients (adj. HR 2.43; P 0.01) compared with Group A, and 110% lower in Group C (adj. HR 2.1; P 0.04). The effect of HCT-CI was of a higher magnitude in the multivariate comparison (adj. HR 3; P <0.01).

Information on in vivo T cell depletion was available at a later stage of project development. When OS and GRFS were compared by type of depletion (Alemtuzumab/ATG), no significant differences were observed (**Appendix Figure C.14**).

Table 3.22 Univariate effects on OS - PBSC graft group. Univariate hazard ratios (HR) from the Cox PH model, with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Variable	Ref	Univariate				
		HR	Lower	Upper	P. val.	PH test
Age at transplant		1.01	0.99	1.03	0.33	0.20
CD34+ dose		1.03	0.83	1.28	0.77	0.78
DCMV status Pos	Neg	1.2	0.72	2.01	0.48	0.33
Donor age		1.01	0.99	1.03	0.57	0.57
Donor relationship UD	RD	0.87	0.52	1.45	0.59	0.24
Sex mismatch Yes	No	1.2	0.71	2.02	0.49	0.64
HCT-CI 1	0	0.85	0.33	2.16	0.73	0.27
HCT-CI 2	0	1.18	0.52	2.7	0.70	0.27
HCT-CI 3+	0	2.3	1.27	4.14	0.01	0.27
MAC or RIC RIC	MAC	1.43	0.8	2.57	0.23	0.83
Patient sex Female	Male	1.06	0.63	1.77	0.83	0.93
Patient CMV status Pos	Neg	1.91	0.97	3.78	0.07	0.53
PBSC graft group B	A	2.03	1.11	3.72	0.03	0.12
PBSC graft group C	A	2.03	1.08	3.83	0.03	0.12

Table 3.23 Multivariate effects on OS - PBSC graft group. Multivariate hazard ratios (Adj. HR) from the Cox PH model, with 95% confidence interval (lower, upper), and significance test (P. val). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate			
		HR	Lower	Upper	P. val.
Age at transplant		1	0.98	1.03	0.74
CD34+ dose		0.95	0.75	1.2	0.66
DCMV status Pos	Neg	0.93	0.51	1.7	0.82
Donor age		0.99	0.96	1.02	0.51

Variable	Ref	Multivariate			
		HR	Lower	Upper	P. val.
Donor relationship UD	RD	0.83	0.37	1.86	0.65
Sex mismatch Yes	No	1.04	0.57	1.91	0.90
HCT-CI 1	0	0.91	0.35	2.35	0.84
HCT-CI 2	0	1.66	0.68	4.04	0.27
HCT-CI 3+	0	3	1.57	5.72	<0.01
MAC or RIC RIC	MAC	1.4	0.74	2.66	0.31
Patient sex Female	Male	1.1	0.61	1.99	0.75
Patient CMV status Pos	Neg	2	0.93	4.32	0.09
PBSC graft group B	A	2.43	1.28	4.63	0.01
PBSC graft group C	A	2.1	1.05	4.18	0.04

3.9.2.2 Effects of graft composition on OS

The association between graft composition and the rate of OS was tested with the compositional Cox PH model (**Table 3.24**). Compositional variables were expressed as additive log-ratios with the proportion of HSCs as the reference part. Clinical variables were included as additional covariates.

An LR test indicated significant evidence (P 0.003) of an association between the rate of OS and the whole composition, beyond that attributable to other variables.

In respect of the individual parts of the composition, the rate of OS decreased significantly as the ratio of NK cells to HSCs increased (adj. HR 1.86; P <0.01). The opposite was observed for the proportion of lymphoid progenitor cells (adj. HR 0.73; P 0.06), and monocytes (adj. HR 0.73; P 0.06) where an increase in the proportion was associated with improved survival, however this did not achieve significance.

Table 3.24 Multivariate effects on OS - graft composition. Multivariate hazard ratios (Adj. HR) from the Cox PH model, with 95% confidence interval (lower, upper), and significance test (P. val). Compositional variables expressed as additive log-ratios (alr). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate			
		Adj. HR	Lower	Upper	P. val.
Age at transplant		1	0.98	1.03	0.71
CD34+ dose		0.93	0.73	1.18	0.53
DCMV status Pos	Neg	0.98	0.52	1.83	0.95
Donor age		0.99	0.96	1.02	0.38
Donor relationship UD	RD	0.84	0.38	1.86	0.67
Sex mismatch Yes	No	1.04	0.53	2.02	0.91
HCT-CI 1	0	0.93	0.34	2.57	0.90
HCT-CI 2	0	2.28	0.89	5.84	0.09
HCT-CI 3+	0	4	1.98	8.06	<0.01
MAC or RIC RIC	MAC	1.5	0.77	2.92	0.24
Patient sex Female	Male	1.03	0.55	1.95	0.92
Patient CMV status Pos	Neg	1.97	0.89	4.39	0.10
alr NK vs HSC		1.86	1.29	2.69	<0.01
alr DC vs HSC		1.36	0.83	2.26	0.23
alr T vs HSC		0.88	0.55	1.41	0.60
alr B vs HSC		0.87	0.58	1.31	0.52
alr Mixed M/DC vs HSC		0.98	0.74	1.3	0.91

Variable	Ref	Multivariate			
		Adj. HR	Lower	Upper	P. val.
alr Lymph. progen. vs HSC		0.7	0.49	1.01	0.06
alr M vs HSC		0.73	0.52	1.01	0.06

3.9.3 Relapse & transplant-related mortality

Overall and GvHD-free, relapse-free survival describe successful and unsuccessful responses to therapy; however, they don't provide insight into the causes of mortality or type of events experienced. Next, the cumulative incidence of relapse, transplant related mortality (TRM), AGvHD and CGvHD, were studied to see if they were associated with PBSC graft group and graft composition.

3.9.3.1 Effects of PBSC graft group on relapse & TRM

The subdistribution hazard $\lambda_k^{SD}(t)$ of relapse was regressed on clinical variables and PBSC graft groups with the FG model (**Table 3.25**). This was repeated for TRM, the competing event (**Table 3.26**). A significant subdistribution hazard ratio (HR_{SD}) indicates a significant change in the cumulative incidence (probability) of the event, however, the specific magnitude cannot be interpreted.

In univariate comparisons, there was no difference in relapse incidence by PBSC graft group (P > 0.05). The incidence of TRM, however, was higher in Group B relative to Group A (HR_{SD} 3.06; P 0.01). The HR_{SD} for Group C on TRM could not be reliably estimated due violation of the PH assumption (P 0.03). Having a sex mismatched donor was associated with a decrease in relapse incidence (HR_{SD} 0.46; P 0.05) but conversely an increase in TRM (HR_{SD} 2.11; P 0.04). This indicates the protective effect of sex mismatch on relapse may be determined by a higher rate of deaths from transplant related causes.

Table 3.25 Univariate effects on relapse – PBSC graft group. Univariate subdistribution hazard ratios (HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Variable	Ref	Univariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		0.98	0.96	1	0.09	0.37
CD34+ dose		1.2	0.91	1.57	0.20	0.20
DCMV status Pos	Neg	0.75	0.37	1.54	0.44	0.59
Donor age		1	0.97	1.03	0.92	0.22
Donor relationship UD	RD	0.55	0.27	1.1	0.09	0.50
Sex mismatch Yes	No	0.46	0.21	0.99	0.05	0.39
HCT-CI 1	0	0.88	0.25	3.1	0.85	0.66
HCT-CI 2	0	0.93	0.26	3.33	0.91	0.08
HCT-CI 3+	0	1.98	0.9	4.35	0.09	0.80
MAC or RIC RIC	MAC	0.8	0.38	1.66	0.55	0.58
Patient sex Female	Male	0.75	0.36	1.56	0.44	0.59
Patient CMV status Pos	Neg	1.24	0.54	2.85	0.61	0.43
PBSC graft group B	A	0.89	0.36	2.16	0.79	0.83
PBSC graft group C	A	1.47	0.65	3.31	0.36	0.18

Table 3.26 Univariate effects on transplant-related mortality – PBSC graft group. Univariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Variable	Ref	Univariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		1.03	0.99	1.06	0.13	0.44
CD34+ dose		0.93	0.7	1.22	0.58	0.85
DCMV status Pos	Neg	1.61	0.81	3.23	0.18	0.30

Variable	Ref	Univariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Donor age		1	0.98	1.03	0.78	0.42
Donor relationship UD	RD	1.53	0.73	3.21	0.26	0.51
Sex mismatch Yes	No	2.11	1.04	4.3	0.04	0.77
HCT-CI 1	0	0.91	0.27	3.08	0.88	0.38
HCT-CI 2	0	1.13	0.38	3.37	0.83	0.93
HCT-CI 3+	0	1.77	0.78	4.02	0.17	0.10
MAC or RIC RIC	MAC	2.18	0.9	5.28	0.08	0.63
Patient sex Female	Male	1.37	0.69	2.69	0.37	0.56
Patient CMV status Pos	Neg	1.65	0.68	3.99	0.27	0.81
PBSC graft group B	A	3.06	1.35	6.92	0.01	0.79
PBSC graft group C	A	2.24	0.92	5.46	0.08	0.03

In a multivariate comparison, no variables retained a significant association with the incidence of relapse (**Table 3.27**). There was an insignificant increase in relapse incidence in recipients of PBSC graft Group C (adj. HR_{SD} 2.17; P 0.09). Whereas patients with a sex mismatched (adj. HR_{SD} 0.47; P 0.09) or unrelated donor (adj. HR_{SD} 0.37; P 0.07) were associated with an insignificant decrease in relapse incidence.

In contrast, the incidence of TRM was significantly different between clinical and PBSC graft groups (**Table 3.28**). Receiving a graft from PBSC graft Group B was associated with an increase in TRM incidence (adj. HR_{SD} 3.45; P 0.01). Whereas the impact of Group C on TRM could not be reliably estimated due to violation of the PH assumption (P 0.04). Having an HCT-CI score of 3+ was also associated with increased TRM incidence (adj. HR_{SD} 3.53; P 0.02).

Table 3.27 Multivariate effects on relapse - PBSC graft group. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		0.98	0.95	1.02	0.30	0.40
CD34+ dose		1.12	0.81	1.54	0.49	0.26
DCMV status Pos	Neg	0.63	0.27	1.43	0.27	0.62
Donor age		0.99	0.94	1.04	0.61	0.16
Donor relationship UD	RD	0.37	0.13	1.09	0.07	0.53
Sex mismatch Yes	No	0.47	0.2	1.14	0.09	0.47
HCT-CI 1	0	1.07	0.26	4.46	0.92	0.60
HCT-CI 2	0	1.29	0.32	5.21	0.72	0.09
HCT-CI 3+	0	1.91	0.81	4.52	0.14	0.80
MAC or RIC RIC	MAC	1.07	0.47	2.47	0.87	0.51
Patient sex Female	Male	0.85	0.36	2.03	0.72	0.54
Patient CMV status Pos	Neg	1.75	0.6	5.08	0.30	0.52
PBSC graft group B	A	1.1	0.44	2.76	0.84	0.87
PBSC graft group C	A	2.17	0.9	5.27	0.09	0.20

Table 3.28 Multivariate effects on transplant-related mortality - PBSC graft group. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		1.02	0.98	1.05	0.40	0.47
CD34+ dose		0.93	0.68	1.29	0.68	0.89
DCMV status Pos	Neg	1.6	0.68	3.77	0.28	0.37
Donor age		1	0.95	1.04	0.94	0.37
Donor relationship UD	RD	2.16	0.57	8.19	0.26	0.59
Sex mismatch Yes	No	1.76	0.76	4.07	0.19	0.66
HCT-CI 1	0	0.96	0.24	3.84	0.96	0.33
HCT-CI 2	0	1.33	0.46	3.79	0.60	0.95
HCT-CI 3+	0	3.53	1.21	10.28	0.02	0.08
MAC or RIC RIC	MAC	1.83	0.67	4.96	0.24	0.58
Patient sex Female	Male	1.19	0.59	2.4	0.62	0.50
Patient CMV status Pos	Neg	1.26	0.45	3.53	0.66	0.72
PBSC graft group B	A	3.45	1.39	8.58	0.01	0.77
PBSC graft group C	A	1.67	0.61	4.57	0.32	0.04

3.9.3.2 Effects of graft composition on relapse & TRM

The association between graft composition and the incidence of relapse was tested with the compositional FG model (**Table 3.29**). This was repeated for TRM, the competing event (**Table 3.30**). Compositional variables were expressed as additive log-ratios with the proportion of HSCs as the reference part. Clinical variables were included as additional covariates.

An LR test indicated significant evidence (P 0.02) of an association between the incidence of TRM and the whole graft composition, beyond that attributable to other variables. However, no clear link with relapse was observed (P 0.45).

In respect of the individual parts of the composition, none of the terms had a significant association with relapse incidence (P > 0.05). There was an insignificant decrease in relapse incidence as the proportion of monocytes increased relative to HSCs (adj. HR_{SD} 0.6; P 0.09), with no apparent impact on the incidence of TRM.

In contrast, several lineages impacted the incidence of TRM. The incidence of TRM was significantly higher as the log-ratio of NK cells to HSCs increased (adj. HR_{SD} 2.22; P <0.01). The opposite was observed for the proportion of lymphoid progenitor cells (adj. HR_{SD} 0.54; P <0.01) where an increased proportion was associated with a reduction in TRM incidence.

Table 3.29 Multivariate effects on relapse - graft composition. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val), and proportional hazards test (PH test). Compositional variables expressed as additive log-ratios (alr). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		0.99	0.95	1.02	0.50	0.43
CD34+ dose		1.08	0.75	1.54	0.68	0.23
DCMV status Pos	Neg	0.65	0.26	1.62	0.36	0.65

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Donor age		0.98	0.93	1.04	0.52	0.17
Donor relationship UD	RD	0.36	0.11	1.13	0.08	0.46
Sex mismatch Yes	No	0.4	0.15	1.11	0.08	0.43
HCT-CI 1	0	1.16	0.27	5.05	0.84	0.57
HCT-CI 2	0	1.8	0.35	9.25	0.48	0.06
HCT-CI 3+	0	2.79	1.02	7.64	0.05	0.73
MAC or RIC RIC	MAC	1.05	0.46	2.43	0.90	0.50
Patient sex Female	Male	0.73	0.29	1.86	0.51	0.54
Patient CMV status Pos	Neg	1.78	0.6	5.26	0.30	0.53
alr DC vs HSC		1.67	0.81	3.45	0.16	0.53
alr Lymph. progen. vs HSC		1.19	0.68	2.08	0.54	0.61
alr NK vs HSC		1.11	0.68	1.81	0.67	0.43
alr T vs HSC		0.99	0.57	1.72	0.96	0.75
alr B vs HSC		0.91	0.46	1.8	0.79	0.82
alr Mixed M/DC vs HSC		0.81	0.54	1.21	0.31	0.13
alr M vs HSC		0.6	0.33	1.08	0.09	0.95

Table 3.30 Multivariate effects on TRM - graft composition. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val), and proportional hazards test (PH test). Compositional variables expressed as additive log-ratios (alr). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		1.01	0.97	1.05	0.64	0.54
CD34+ dose		0.96	0.69	1.34	0.82	0.91
DCMV status Pos	Neg	1.41	0.56	3.54	0.47	0.43
Donor age		0.99	0.94	1.05	0.81	0.30
Donor relationship UD	RD	2.24	0.57	8.82	0.25	0.59
Sex mismatch Yes	No	2.2	0.66	7.26	0.20	0.44
HCT-CI 1	0	0.77	0.17	3.44	0.73	0.32
HCT-CI 2	0	1.31	0.43	3.99	0.63	0.88
HCT-CI 3+	0	4.12	1.39	12.16	0.01	0.06
MAC or RIC RIC	MAC	2.4	0.95	6.11	0.07	0.61
Patient sex Female	Male	0.86	0.34	2.15	0.74	0.35
Patient CMV status Pos	Neg	1.09	0.35	3.41	0.88	0.74
alr NK vs HSC		2.22	1.4	3.54	<0.01	0.21
alr DC vs HSC		1.11	0.57	2.15	0.76	0.31
alr Mixed M/DC vs HSC		1.1	0.74	1.63	0.63	0.42
alr T vs HSC		0.97	0.5	1.9	0.94	0.46
alr M vs HSC		0.97	0.6	1.58	0.91	0.26
alr B vs HSC		0.76	0.46	1.24	0.27	0.50
alr Lymph. progen. vs HSC		0.54	0.35	0.84	<0.01	0.58

3.9.4 AGvHD

3.9.4.1 Effects of PBSC graft group on AGvHD

The subdistribution hazard $\lambda_k^{SD}(t)$ of AGvHD was regressed on clinical variables and PBSC graft groups with the FG model (**Table 3.31**). This was repeated for death (without AGvHD), the competing event for comparative purposes (**Table 3.32**).

None of the variables tested were significantly associated with incidence of AGvHD in univariate comparisons (P > 0.05). There was an insignificant increase AGvHD incidence for recipients of PBSC graft Group C (HR_{SD} 2.02; P 0.06) and for patients receiving a graft from an unrelated donor (HR_{SD} 1.84; P 0.08). There was evidence that the MAC or RIC variable violated the PH assumption when tested on AGvHD incidence (P 0.01), so was excluded from multivariate comparisons.

Table 3.31 Univariate effects on AGvHD - PBSC graft group. Univariate subdistribution hazard ratios (HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Variable	Ref	Univariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		1.02	0.99	1.05	0.16	0.95
CD34+ dose		1	0.73	1.38	0.98	0.38
DCMV status Pos	Neg	1.22	0.66	2.26	0.53	0.95
Donor age		1.01	0.99	1.03	0.43	0.39
Donor relationship UD	RD	1.84	0.93	3.65	0.08	0.61
Sex mismatch Yes	No	1.24	0.67	2.29	0.50	0.87
HCT-CI 1	0	1.24	0.46	3.34	0.66	0.35
HCT-CI 2	0	1.16	0.43	3.09	0.77	0.64
HCT-CI 3+	0	1.31	0.61	2.85	0.49	0.67
MAC or RIC RIC	MAC	1.12	0.58	2.15	0.74	0.01
Patient sex Female	Male	1.09	0.58	2.03	0.80	0.40
Patient CMV status Pos	Neg	1.25	0.6	2.6	0.55	0.50
PBSC graft group B	A	0.97	0.44	2.12	0.93	0.31
PBSC graft group C	A	2.02	0.98	4.17	0.06	0.22

Table 3.32 Univariate effects on Death (without AGvHD) - PBSC graft group. Univariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Variable	Ref	Univariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		0.99	0.97	1.01	0.36	0.34
CD34+ dose		1.13	0.93	1.39	0.22	0.87
DCMV status Pos	Neg	0.95	0.51	1.75	0.87	0.34
Donor age		1	0.98	1.02	0.96	0.74
Donor relationship UD	RD	0.63	0.34	1.15	0.13	0.73
Sex mismatch Yes	No	0.72	0.38	1.37	0.32	0.48
HCT-CI 1	0	0.86	0.28	2.64	0.80	0.09
HCT-CI 2	0	0.83	0.29	2.36	0.72	0.66
HCT-CI 3+	0	1.74	0.86	3.53	0.13	0.77

Variable	Ref	Univariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
MAC or RIC RIC	MAC	1.3	0.66	2.57	0.45	0.26
Patient sex Female	Male	0.92	0.49	1.72	0.79	0.86
Patient CMV status Pos	Neg	1.77	0.77	4.05	0.18	0.18
PBSC graft group B	A	1.94	0.97	3.89	0.06	0.68
PBSC graft group C	A	1.29	0.59	2.84	0.53	0.29

The association between PBSC graft Group C and AGvHD incidence was less certain in a multivariate comparison (adj. HR_{SD} 1.7; P 0.22). Whereas having an unrelated donor was associated with an increase in the incidence of AGvHD (adj. HR_{SD} 3.35; P 0.02) (**Table 3.33**). There was also a marginal increase in AGvHD with older donors (adj. HR_{SD} 1.03; P 0.09). Group B was not associated with AGvHD incidence (adj. HR_{SD} 0.93; P 0.87) but was linked with an increase in the competing risk, death (without AGvHD) (adj. HR_{SD} 2.5; P 0.01) (**Table 3.34**).

Table 3.33 Multivariate effects on AGvHD - PBSC graft group. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		1.01	0.98	1.04	0.54	0.96
CD34+ dose		1.01	0.75	1.35	0.97	0.42
DCMV status Pos	Neg	1.08	0.53	2.19	0.84	0.96
Donor age		1.03	1	1.07	0.09	0.35
Donor relationship UD	RD	3.35	1.17	9.56	0.02	0.58
Sex mismatch Yes	No	1.09	0.44	2.67	0.86	0.86
HCT-CI 1	0	1.34	0.46	3.87	0.59	0.40
HCT-CI 2	0	1.29	0.43	3.87	0.65	0.68
HCT-CI 3+	0	1.38	0.61	3.1	0.44	0.62
Patient sex Female	Male	1.16	0.48	2.77	0.75	0.34
Patient CMV status Pos	Neg	1.04	0.49	2.19	0.93	0.45
PBSC graft group B	A	0.93	0.41	2.11	0.87	0.35
PBSC graft group C	A	1.7	0.73	3.91	0.22	0.17

Table 3.34 Multivariate effects on Death (without AGvHD) - PBSC graft group. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		0.99	0.97	1.02	0.48	0.35
CD34+ dose		1.09	0.87	1.38	0.45	0.84
DCMV status Pos	Neg	0.67	0.32	1.44	0.31	0.28
Donor age		0.98	0.94	1.02	0.29	0.80
Donor relationship UD	RD	0.49	0.19	1.22	0.12	0.64
Sex mismatch Yes	No	0.75	0.36	1.55	0.43	0.42

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
HCT-CI 1	0	1.06	0.33	3.43	0.92	0.09
HCT-CI 2	0	1.14	0.36	3.57	0.83	0.70
HCT-CI 3+	0	1.91	0.88	4.11	0.10	0.77
Patient sex Female	Male	0.86	0.42	1.77	0.68	0.90
Patient CMV status Pos	Neg	2.26	0.85	6.02	0.10	0.19
PBSC graft group B	A	2.5	1.21	5.17	0.01	0.71
PBSC graft group C	A	1.64	0.72	3.7	0.24	0.32

3.9.4.2 Effects of graft composition on AGvHD

The association between graft composition and the incidence of AGvHD was tested with the compositional FG model (**Table 3.35**). This was repeated for death (without AGvHD), the competing event, for comparative purposes (**Table 3.36**). Compositional variables were expressed as additive log-ratios with the proportion of HSCs as the reference part. Clinical variables were included as additional covariates.

An LR test indicated no evidence (P 0.77) of an association between the incidence of AGvHD and the whole composition, beyond that attributable to other variables. Concerning the individual parts of the composition, none of the terms had a significant association with AGvHD incidence (P > 0.05). The incidence of death (without AGvHD) rose significantly as the log-ratio of NK cells to HSCs increased (adj. HR_{SD} 1.71; P 0.01), but this did not influence the AGvHD incidence (adj. HR_{SD} 0.9; P 0.68).

Table 3.35 Multivariate effects on AGvHD - graft composition. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). Compositional variables expressed as additive log-ratios (alr). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		1.01	0.98	1.04	0.50	0.95
CD34+ dose		1.01	0.71	1.43	0.95	0.40
DCMV status Pos	Neg	0.93	0.43	2.02	0.85	0.95
Donor age		1.03	0.99	1.08	0.10	0.40
Donor relationship UD	RD	3.77	1.22	11.64	0.02	0.61
Sex mismatch Yes	No	1.13	0.46	2.77	0.79	0.87
HCT-CI 1	0	1.44	0.49	4.28	0.51	0.40
HCT-CI 2	0	1.32	0.42	4.14	0.63	0.63
HCT-CI 3+	0	1.48	0.62	3.55	0.37	0.63
Patient sex Female	Male	1.09	0.47	2.52	0.84	0.38
Patient CMV status Pos	Neg	0.99	0.46	2.15	0.99	0.48
alr T vs HSC		1.46	0.75	2.84	0.26	0.13
alr DC vs HSC		1.13	0.61	2.09	0.69	0.27
alr Lymph. progen. vs HSC		1.06	0.65	1.75	0.81	0.19
alr Mixed M/DC vs HSC		1.01	0.73	1.41	0.95	0.11
alr M vs HSC		0.94	0.6	1.48	0.79	0.41
alr NK vs HSC		0.9	0.55	1.47	0.68	0.37
alr B vs HSC		0.76	0.47	1.24	0.28	0.55

Table 3.36 Multivariate effects on Death (without AGvHD) - graft composition. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). Compositional variables expressed as additive log-ratios (alr). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		0.99	0.96	1.02	0.64	0.35
CD34+ dose		1.04	0.8	1.35	0.79	0.85
DCMV status Pos	Neg	0.78	0.37	1.66	0.51	0.23
Donor age		0.98	0.94	1.01	0.20	0.78
Donor relationship UD	RD	0.46	0.19	1.14	0.09	0.71
Sex mismatch Yes	No	0.69	0.29	1.6	0.38	0.37
HCT-CI 1	0	1.07	0.31	3.62	0.92	0.07
HCT-CI 2	0	1.59	0.47	5.43	0.46	0.64
HCT-CI 3+	0	2.41	1.05	5.55	0.04	0.70
Patient sex Female	Male	0.78	0.36	1.69	0.53	0.91
Patient CMV status Pos	Neg	2.63	0.87	7.97	0.09	0.20
alr NK vs HSC		1.71	1.13	2.6	0.01	0.94
alr B vs HSC		1.34	0.84	2.14	0.22	0.93
alr DC vs HSC		1.15	0.71	1.85	0.58	0.78
alr Mixed M/DC vs HSC		0.95	0.68	1.31	0.74	0.50
alr T vs HSC		0.73	0.42	1.25	0.25	0.97
alr Lymph. progen. vs HSC		0.73	0.48	1.1	0.13	0.97
alr M vs HSC		0.71	0.47	1.06	0.09	0.48

3.9.5 CGvHD

3.9.5.1 Effects of PBSC graft group on CGvHD

The subdistribution hazard $\lambda_k^{SD}(t)$ of CGvHD was regressed on clinical variables and PBSC graft groups with the FG model (**Table 3.37**). This was repeated for death (without CGvHD), the competing event for comparative purposes (**Table 3.38**).

None of the variables tested were significantly associated with incidence of CGvHD in univariate comparisons (P > 0.05). There was a marginal decrease in CGvHD incidence for recipients with an HCT-CI score of 3+ (HR_{SD} 0.36; P 0.1). However, this may be related to an increase in the competing risk (HR_{SD} 2.24; P 0.01). The incidence of death (without CGvHD) was also marginally higher for CMV positive patients (HR_{SD} 2.11; P 0.06). There was evidence patient gender violated the PH assumption when tested on CGvHD incidence (P 0.02); as did PBSC graft Group C when tested on the competing risk (P 0.03).

Due to the lack of univariate effects on the outcome of interest (CGvHD) and evidence of time-dependence, the effect of PBSC graft group on CGvHD was not tested in a multivariate comparison.

Table 3.37 Univariate effects on CGvHD – PBSC graft group. Univariate subdistribution hazard ratios (HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Variable	Ref	Univariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		0.99	0.96	1.01	0.19	0.67
CD34+ dose		1.28	0.94	1.73	0.11	0.63
DCMV status Pos	Neg	0.66	0.35	1.23	0.19	0.31
Donor age		0.99	0.97	1.02	0.63	0.80
Donor relationship UD	RD	1.09	0.58	2.04	0.80	0.35
Sex mismatch Yes	No	1.07	0.57	1.99	0.84	0.74
HCT-CI 1	0	1.33	0.52	3.38	0.55	0.14
HCT-CI 2	0	1.31	0.6	2.86	0.50	0.10
HCT-CI 3+	0	0.36	0.11	1.2	0.10	0.69
MAC or RIC RIC	MAC	0.69	0.36	1.3	0.25	0.05
Patient sex Female	Male	1.14	0.61	2.13	0.69	0.02
Patient CMV status Pos	Neg	0.81	0.42	1.57	0.53	0.39
PBSC graft group B	A	0.94	0.45	1.97	0.86	0.28
PBSC graft group C	A	1.24	0.58	2.65	0.57	0.23

Table 3.38 Univariate effects on Death (without CGvHD) – PBSC graft group. Univariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). P. val < 0.1 highlighted in yellow.

Variable	Ref	Univariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		1.02	0.99	1.04	0.20	0.65
CD34+ dose		1	0.81	1.23	0.97	0.90
DCMV status Pos	Neg	1.27	0.74	2.18	0.39	0.24
Donor age		1	0.98	1.03	0.74	0.51
Donor relationship UD	RD	1.05	0.6	1.82	0.87	0.29

Variable	Ref	Univariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Sex mismatch Yes	No	1.03	0.59	1.8	0.91	0.42
HCT-CI 1	0	0.68	0.23	2.04	0.49	0.13
HCT-CI 2	0	0.89	0.33	2.38	0.82	0.30
HCT-CI 3+	0	2.24	1.22	4.11	0.01	0.74
MAC or RIC RIC	MAC	1.47	0.79	2.74	0.22	0.31
Patient sex Female	Male	0.98	0.57	1.7	0.95	0.84
Patient CMV status Pos	Neg	2.11	0.98	4.53	0.06	0.32
PBSC graft group B	A	1.75	0.89	3.42	0.10	0.59
PBSC graft group C	A	2.07	1.1	3.89	0.02	0.03

3.9.5.2 Effects of graft composition on CGvHD

Using a log-contrast FG model, the effects of graft composition on the incidence of CGvHD were tested (**Table 3.39**). The effects on death (without CGvHD), the competing risk, are also shown for comparative purposes (**Table 3.40**). Compositional variables were expressed as additive log-ratios with the proportion of HSCs as the reference part. Clinical variables were included as additional covariates.

An LR test indicated no evidence (P 0.24) of an association between the incidence of CGvHD and the whole composition, beyond that attributable to other variables.

In respect of the individual parts of the composition, an increase in the log-ratio of T cells to HSCs was associated with an increase in CGvHD incidence (adj. HR_{SD} 2.01; P 0.05). HCT-CI scores had opposite effects on each competing risk, with a score of 3+ being associated with lower CGvHD incidence (adj. HR_{SD} 0.28; P 0.05) but conversely, a greater incidence of death (without CGvHD) (adj. HR_{SD} 4.08; P <0.01). There was an insignificant decrease in CGvHD incidence in patients with CMV positive donors (adj. HR_{SD} 0.48; P 0.08), with no corresponding change in the incidence of the competing risk (adj. HR_{SD} 1.14; P 0.7).

PH tests indicated evidence of time-dependent effects between the mixed M/DC lineage and the incidence of death (without CGvHD) (P 0.05), so the HR_{SD} could not be reliably estimated.

Table 3.39 Multivariate effects on CGvHD - graft composition. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). Compositional variables expressed as additive log-ratios (alr). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		0.99	0.95	1.02	0.45	0.87
CD34+ dose		1.42	0.99	2.05	0.06	0.50
DCMV status Pos	Neg	0.48	0.22	1.08	0.08	0.47
Donor age		1.01	0.98	1.05	0.47	0.88
Donor relationship UD	RD	1.42	0.56	3.58	0.46	0.33
Sex mismatch Yes	No	1.16	0.57	2.38	0.69	0.74
HCT-CI 1	0	1.65	0.62	4.4	0.32	0.19
HCT-CI 2	0	0.87	0.31	2.49	0.80	0.09
HCT-CI 3+	0	0.28	0.08	1.02	0.05	0.57
MAC or RIC RIC	MAC	0.58	0.26	1.29	0.19	0.11

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Patient CMV status Pos	Neg	1.01	0.42	2.39	0.99	0.43
alr T vs HSC		2.01	1.01	3.98	0.05	0.69
alr DC vs HSC		1.25	0.59	2.62	0.56	0.85
alr Lymph. progen. vs HSC		1.04	0.64	1.67	0.88	0.46
alr Mixed M/DC vs HSC		1.02	0.74	1.42	0.88	0.64
alr M vs HSC		0.88	0.59	1.33	0.55	0.17
alr B vs HSC		0.83	0.51	1.33	0.43	0.16
alr NK vs HSC		0.72	0.44	1.18	0.20	0.18

Table 3.40 Multivariate effects on Death (without CGvHD) - graft composition. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with 95% confidence interval (lower, upper), significance test (P. val) and proportional hazards test (PH test). Compositional variables expressed as additive log-ratios (alr). P. val < 0.1 highlighted in yellow.

Variable	Ref	Multivariate				
		HR_{SD}	Lower	Upper	P. val.	PH test
Age at transplant		1.01	0.98	1.03	0.64	0.69
CD34+ dose		0.87	0.68	1.12	0.28	0.87
DCMV status Pos	Neg	1.14	0.59	2.19	0.70	0.23
Donor age		0.99	0.95	1.02	0.45	0.50
Donor relationship UD	RD	1.09	0.44	2.67	0.85	0.41
Sex mismatch Yes	No	0.85	0.42	1.73	0.65	0.66
HCT-CI 1	0	0.72	0.21	2.41	0.59	0.10
HCT-CI 2	0	1.98	0.71	5.48	0.19	0.48
HCT-CI 3+	0	4.08	1.88	8.87	<0.01	0.70
MAC or RIC RIC	MAC	1.52	0.74	3.16	0.26	0.28
Patient CMV status Pos	Neg	2.09	0.83	5.24	0.12	0.32
alr NK vs HSC		2.09	1.41	3.1	<0.01	0.77
alr DC vs HSC		1.29	0.83	2.02	0.26	0.12
alr Mixed M/DC vs HSC		0.98	0.76	1.27	0.88	0.05
alr B vs HSC		0.91	0.61	1.34	0.62	0.93
alr T vs HSC		0.76	0.46	1.25	0.28	0.87
alr Lymph. progen. vs HSC		0.73	0.53	1.02	0.06	0.36
alr M vs HSC		0.67	0.47	0.95	0.02	0.49

3.10 Summary

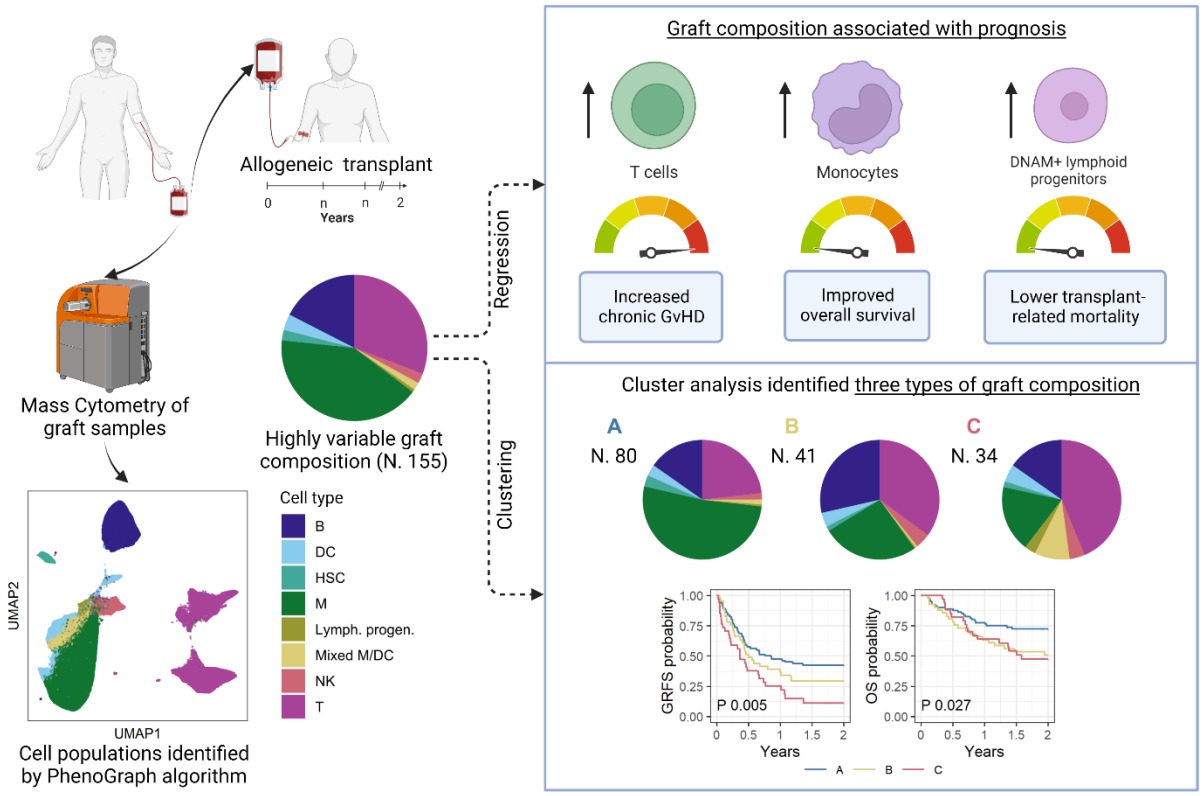


Figure 3.22 Graphical summary.

Cellular composition was highly variable between donors (185 PBSC grafts; 4,441,335 cells). The majority of cells were identified as monocytes (~36%; IQR 20 - 51%), T cells (~28%; IQR 17 - 39%) and B cells (~15%; IQR 10 - 23%). A smaller proportion of cells belonged to the DC (~3%; IQR 2 - 5%), NK cell (~2%; 1 - 4%) and HSC (~2%; 1-4%) compartments. Around 1% (IQR 0 – 2%) of cells were DNAM-1$^+$ lymphoid progenitor cells (Lin$^-$ CD34$_{dim}$ DNAM-1$^+$ CXCR4$^+$ CD45RA$^-$) which were not anticipated prior to this study. NKT and CLP cells each comprised < 0.1% of cells.

Cluster analysis for compositional data inferred the presence of three distinct groups of PBSC grafts based on cellular composition (**Figure 3.22**). One large group, A (N = 80), and two smaller groups, B (N = 41) and C (N = 34). Group A was characterised by a relative abundance of monocytes and HSCs but the lowest proportion of T and NK cells. Conversely, Group C had the highest proportion of T cells, DNAM-1$^+$ lymphoid progenitor cells, and the least monocytes. Whereas Group B was associated with an increased proportion of B and T cells with relatively few monocytes.

The probability of OS, GRFS, and TRM varied significantly between PBSC graft groups ($P < 0.05$). Patients receiving grafts from Group A had the most favourable rate of GRFS (42%; CI 33 – 55) and OS (72%; CI 63-83) 2 years post-transplant. As well as the lowest TRM incidence (12%; CI 5 - 20). The TRM incidence was more than double in Group B (35%; CI 20 – 49) and C (28%; CI 12 – 43). Group C also had a higher incidence of GvHD of any kind. However, no difference in relapse incidence was observed between the groups.

PBSC graft groups were defined by composition, however, donor factors, known to affect patient prognosis, also varied between groups. Group A was donated by the youngest donors (median 34; IQR 24 – 48), the least of which were CMV-positive (43%). Group B had the oldest donors (median 41; IQR 29 - 51), a greater percentage of which were female (39%) but relatively few that were unrelated

(51%). Whereas Group C had the most CMV positive (62%) unrelated donors (71%), and the lowest percentage of female donors (23%).

Survival and competing event analysis was used to test the effect graft group and composition had on outcomes following allo-HSCT. All associations noted below were significant in a multivariate comparison ($P < 0.05$) except where it is stated otherwise.

Regressing outcomes on the PBSC graft groups (with Group A as the reference group) indicated Group B was associated with a lower rate of OS and a higher probability of TRM. However, no difference was observed in the incidence of AGvHD, CGvHD, or relapse. This suggests the increase in TRM in Group B may be related to other causes of TRM.

Group C was associated with a lower rate of both OS and GRFS. Whilst an insignificant univariate increase in AGvHD was observed, this effect was less clear after adjusting for other factors such as donor age and relationship. There was also an insignificant increase in relapse incidence (P 0.09) in this group. Due to violation of the PH assumption, the relationship between Group C and TRM could not be tested.

I found evidence of a significant association between the whole composition of grafts and the rate of both GRFS and OS, beyond that attributable to other variables. Moreover, there was also an association between graft composition and the incidence of TRM.

Concerning the individual parts of the composition, several lineages were found to be predictive of patient prognosis (**Figure 3.22**). An increase in the proportion of T cells relative to HSCs was associated with greater CGvHD incidence. A greater proportion of NK cells relative to HSCs was associated with lower rates of OS and GRFS and increased TRM incidence; however, no impact was observed on the incidence of relapse or GvHD.

Whereas a greater proportion of monocytes relative to HSCs was associated with improvements in the GRFS rate, an insignificant increase in OS (P 0.06) and a decrease in relapse incidence (P 0.09). The proportion of DNAM-1^+ lymphoid progenitors relative to HSCs also had a protective effect on the incidence of TRM whilst being associated with a trend towards improved OS (P 0.06).

Several clinical variables had an association with prognosis. HCT-CI scores ≥ 3 were associated with lower rates of OS and GRFS and an increase in the incidence of relapse. Patients with a sex-mismatched donor had a lower relapse incidence, but conversely an increase in the probability of death from other causes. This suggests the protective effect on relapse may be due to an increase in the competing event. The relapse incidence was higher for patients when their donor was unrelated, and a marginal increase in AGvHD was observed with older donors and CMV-positive donors.

In summary, the cellular composition of grafts was highly variable between donations. The differences were sufficient to define three distinct groups of PBSC grafts, which were associated with significant differences in patient prognosis. Grafts from Group A, with a high proportion of monocytes and HSCs but relatively few T and NK cells, had the most favourable outcomes. In contrast, grafts from Group B and C which contained relatively few monocytes, but a higher proportion of lymphocytes had worse outcomes. The composition of grafts was shown to have a significant impact on prognosis. Concerning the individual lineages, a greater proportion of T cells (relative to HSCs) was predictive of increased CGvHD incidence. An increase in the proportion of NK cells was associated with lower rates of OS, GRFS, and higher TRM. Whereas monocytes and DNAM-1^+ lymphoid progenitors had a protective effect on the rate of GRFS and incidence of TRM, respectively.

4 Post-transplant serum protein study

4.1 Introduction

A large number of studies have identified the association of serum proteins with prognosis for allo-HSCT recipients (118). In collaboration with Kriti Verma, Wayne Croft, and Paul Moss, this chapter studies a set of candidate prognostic biomarkers identified in the sera of recipients on the 14th day following transplantation.

A prospective cohort of 154 allo-HSCT recipients was recruited between 2009 and 2017. On the 14th day post-transplant (Day-14), Kriti Verma collected serum samples and conducted OLINK proteomics analysis to determine the relative concentration of 92 inflammation-related proteins (338). Individuals within the cohort were followed up for a median of 3 years after transplantation (CI 1.98-3.9).

Wayne Croft conducted a differential protein expression analysis using this cohort. Differential expression analysis compared recipients who experienced an event (AGvHD (grade II-IV), CGvHD, relapse, and mortality) at any time during the follow-up period with those that were censored. After controlling for the multiple testing burden, this identified a subset of 19 proteins that were differentially expressed between event groups. As samples were taken at an early time point, these proteins could be useful prognostic biomarkers of patient prognosis.

My contribution to the study involved estimating the probability of clinical endpoints (time-to-event outcomes) and developing statistical models to test the impact these candidate biomarkers had on patient prognosis.

This chapter begins by providing an overview of the differential protein expression analysis, in **Section 4.2**. Next, **Section 4.3** details the cohort information and pre-processing. The next two sections focus on the use of survival analysis to analyse OS, **Section 4.4**, and GRFS, **Section 4.5**. This is followed by the use of competing risks analysis to study the dynamics of NRM and RM, in **Section 4.6**. Followed by competing risks analysis of secondary endpoints including relapse, AGvHD, and CGvHD in **Sections 4.8**, **4.9** and **4.10** respectively.

4.2 Differential expression analysis

This section summarises the findings of Kriti Verma and Wayne Croft which motivated the analysis in this chapter. As illustrated in **Figure 4.1**, a subset of 10 proteins measured on Day-14 were shown to be significantly upregulated in patients who did not survive during follow-up compared to those that were censored at the time of evaluation.

CUB-domain-containing protein 1 (CDCP1), Programmed death ligand-1 (PD-L1), and Interleukin-10 (IL-10) in particular showed the most significant increase. CDCP1 is a transmembrane molecule that is expressed in metastatic colon and breast tumours as well as on the surface of haematopoietic stem cells (385). PD-L1 is an immunosuppressive molecule that inhibits T cells by activating the PD-1 / PD-L1 signalling pathway(386, 387) . IL-10 is a cytokine that inhibits production of proinflammatory cytokines and stimulates the proliferation and differentiation of thymocytes, mast cells, and B cells (388).

In contrast to the other events tested where upregulation was predominantly observed, 5 proteins were significantly downregulated on Day-14 in patients who relapsed during follow-up compared to those that remained disease free. These included Interleukin-2 (IL2), Interleukin-6 (IL6), Interleukin-17 (IL17C), matrix metalloproteinase-10 (MMP-10), and C-X-C motif chemokine ligand 10 (CXCL10), with IL2 showing the most significant decrease in concentration. Both IL6 and CXCL10 in the plasma have previously been validated as prognostic biomarkers for allo-HSCT recipients (181-183).

IL-2 is a cytokine that promotes cytotoxic activity by CD8+ T cells and NK cells and is essential for the development and maintenance of T-regs (389, 390). IL6 is a pro-inflammatory cytokine that regulates the emergence of haematopoietic stem cells and has been linked with the parthenogenesis of AGvHD (391, 392). IL17C is a cytokine that regulates innate immune responses (393). MMP-10 is an enzyme required for the maintenance of a tumour-forming, stem-like population of cells in lung cancer patients (394). CXCL10 is a chemokine implicated in a variety of diseases including immune dysfunction, chronic inflammation, and cancer development (395).

Comparing patients diagnosed with AGvHD (grade II-IV) during follow-up to those that were censored identified 6 proteins that were significantly upregulated at Day-14. C-C motif ligand 23 (CCL23) and Leukaemia Inhibitory Factor Receptor (LIF-R) showed the most significant upregulation. Whereas Fms Related Receptor Tyrosine Kinase 3 Ligand was (FLT3L) was found to be downregulated at the same time point. CCL23 is a biomarker of inflammation involved in the regulation of angiogenesis, the formation of new capillaries from existing blood vessels (396, 397) . The ligand of LIF-R, LIF, is a multifunctional cytokine with a wide variety of roles in inflammatory responses and differentiation of leukaemia cells (398). FLT3L is a stem-specific growth factor essential for the survival of thymocytes (399, 400) .

Repeating the analysis for patients diagnosed with CGvHD identified a single upregulated protein at Day-14, Caspase-8 (CASP-8), which is a regulator of apoptosis and lymphocyte development and homeostasis amongst other functions (401).

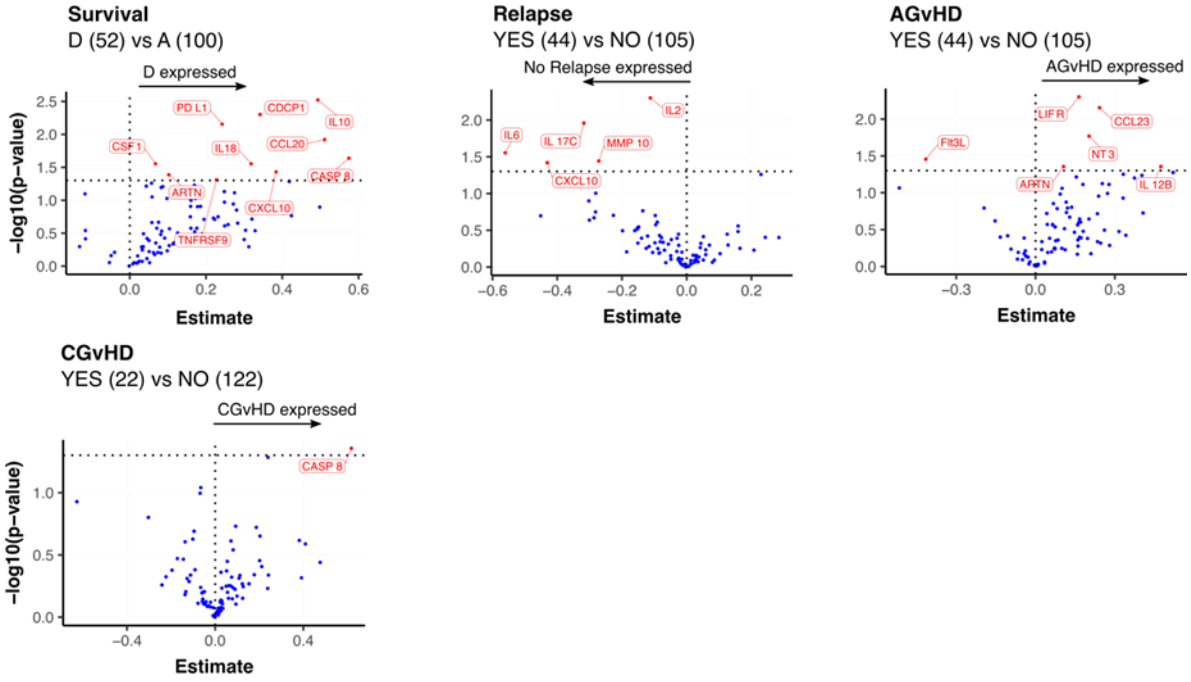


Figure 4.1 Differentially expressed serum protein markers by clinical event. Results of differential protein expression analysis of Day-14 serum protein levels stratified by clinical events Survival: Deceased (D) vs Alive (A); Acute GvHD: Yes vs No; Relapse: Yes vs No; Chronic GvHD: Yes vs NO. Proteins labelled in red showed a significant difference ($p<0.05$) by Man Whitney after Benjamini-Hochberg (BH) multiple testing correction.

4.3 Data & pre-processing

Missing values were present in both the event data and patient information (**Table 4.1** and **Table 4.2**). Due to the relatively low number of patients missing event data (N = 9, 6%), these patients were excluded. A complete-case strategy would have introduced a substantial bias if applied to the patient information, as 61 patients (40%) had at least one missing value. Therefore, multiple imputations of missing values were generated using the method described in **Section 2.2.4.1**.

The primary clinical endpoints defined were overall survival (OS), the time-to-death attributable to any cause; GvHD-free, relapse-free survival, the time-to-first-event-onset (AGVHD (grade II-IV), chronic GVHD, relapse, or death); non-relapse mortality (NRM), the time-to-death without prior relapse; and relapse-related mortality (RM) the time-to-death preceded by a relapse or progression of malignancy. Based on the time-to-event-onset, the secondary clinical endpoints were AGvHD (grade II-IV), CGvHD, and relapse.

The information available for subjects is summarised in **Table 4.2**. This set of variables will be referred to as the *clinical variables*. In pre-processing, ages were divided by 10, and the sex of the donor was encoded as the same sex as the recipient (sex match) or a mismatch (sex mismatch). GvHD prophylaxis was encoded as either ciclosporin or ciclosporin combined with mycophenolate or methotrexate (combination). Disease state at transplant was available but was excluded due to inconsistencies in reporting (41 different descriptions). T cell and mononuclear cell doses were missing for more than 40% of patients so were also excluded.

Table 4.1 Summary of cohort clinical endpoints prior to exclusion of subjects with missing data.

Variable	Group	Overall	Missing (%)
N		154	
AGVHD (%)	Censored	88 (59.1)	3.2
	Yes	61 (40.9)	
CGVHD (%)	Censored	122 (81.3)	2.6
	Yes	28 (18.7)	
Relapse (%)	Censored	105 (70.5)	3.2
	Yes	44 (29.5)	
Death (%)	Censored	100 (65.8)	1.3
	Yes	52 (34.2)	

Table 4.2 Summary of cohort characteristics and clinical endpoints.

Variable	Group	Overall	Missing (%)
N		145	
Age at transplant (median [IQR])		52.0 [43.0, 61.0]	0
CD34+ dose (cells x10^6; median [IQR])		5.0 [4.3, 5.9]	3.4
Donor age (median [IQR])		33.0 [25.0, 47.5]	34.5
Diagnosis (%)	Lymphoid	44 (30.3)	0
	Myeloid	101 (69.7)	
GvHD prophylaxis (%)	Ciclosporin	98 (67.6)	0
	Combination	47 (32.4)	
HLA mismatch status (%)	No	102 (88.7)	20.7
	Yes	13 (11.3)	

Patient sex (%)	Female	51 (35.2)	0
	Male	94 (64.8)	
RIC (%)	No	60 (41.4)	0
	Yes	85 (58.6)	
Sex matched donor (%)	No	63 (44.1)	1.4
	Yes	80 (55.9)	
HCT-CI (%)	0	46 (40.0)	20.7
	1	14 (12.2)	
	2	27 (23.5)	
	>3	28 (24.3)	
TBI	No	118 (81.4)	0
	Yes	27 (18.6)	
AGVHD (%)	Censored	88 (60.7)	0
	Yes	57 (39.3)	
CGVHD (%)	Censored	117 (80.7)	0
	Yes	28 (19.3)	
Relapse (%)	Censored	102 (70.3)	0
	Yes	43 (29.7)	
Death (%)	Censored	97 (66.9)	0
	Yes	48 (33.1)	
Competing fatal events (%)	Censored	97 (66.9)	0
	NRM	28 (19.3)	
	RM	20 (13.8)	

Regression analysis was restricted to proteins identified by Wayne Croft as being differentially expressed by event; these will be referred to as the *protein variables*. The protein variables are summarised in **Table 4.3** along with the outcome with which they were differentially expressed.

The protein variables are measured in Normalized Protein eXpression (NPX), an arbitrary unit on a log2 scale. NPX is a relative measure of protein concentration between subjects within this cohort, as such NPX values for different proteins are not directly comparable and will vary between studies. Olink Proteomics provide a hypothetical example where an NPX of 10 for one particular analyte could correspond to a concentration of 10^2pg/mL, however, for a second analyte an NPX of 10 corresponds to a concentration of 10^4pg/mL(402). This discrepancy in scale could lead to difficulties in the interpretation of effect size estimates.

NPX values were therefore standardised (std-NPX) by subtracting the mean and dividing by the standard deviation. As a result, effects in the regression analysis are estimated in standard deviations away from the average observed NPX value.

Table 4.3 Proteins identified as differentially expressed by outcome group. Protein names, NPX values before and after standardisation, and outcome for which differential expression was observed. Proteins were uniformly upregulated in the outcome group, except those marked ↓.

Protein	NPX Median [IQR]	Standardised NPX Median [IQR]	Outcome
ARTN	0.8 [0.7, 1.1]	-0.2 [-0.5, 0.5]	aGvHD, Survival
CASP-8	4.5 [3.6, 5.5]	0.0 [-0.6, 0.7]	cGvHD, Survival
CCL20	6.5 [5.9, 7.7]	-0.3 [-0.7, 0.4]	Survival

Protein	NPX Median [IQR]	Standardised NPX Median [IQR]	Outcome
CCL23	11.5 [11.1, 11.8]	0.1 [-0.8, 0.6]	aGvHD
CDCP1	4.5 [4.1, 5.0]	0.1 [-0.5, 0.7]	Survival
CSF-1	10.6 [10.5, 10.7]	0.0 [-0.5, 0.5]	Survival
CXCL10	10.0 [9.2, 10.6]	0.0 [-0.7, 0.5]	Relapse↓, Survival
Flt3L	11.0 [10.4, 12.2]	-0.2 [-0.7, 0.8]	aGvHD↓
IL-12B	3.7 [3.0, 4.8]	-0.1 [-0.7, 0.7]	aGvHD
IL-17C	2.4 [1.9, 2.9]	-0.1 [-0.6, 0.4]	Relapse↓
IL10	5.1 [4.7, 5.9]	-0.3 [-0.6, 0.3]	Survival
IL18	10.3 [9.8, 10.9]	-0.1 [-0.6, 0.7]	Survival
IL2	0.9 [0.7, 1.0]	-0.1 [-0.7, 0.6]	Relapse↓
IL6	6.1 [5.1, 7.1]	-0.2 [-0.7, 0.3]	Relapse↓
LIF-R	4.5 [4.3, 4.7]	0.0 [-0.6, 0.7]	aGvHD
MMP-10	8.5 [8.0, 9.0]	-0.1 [-0.7, 0.5]	Relapse↓
NT-3	1.9 [1.6, 2.2]	-0.1 [-0.6, 0.5]	aGvHD
PD-L1	6.9 [6.7, 7.3]	-0.2 [-0.7, 0.6]	Survival
TNFRSF9	7.3 [7.0, 7.7]	0.0 [-0.5, 0.5]	Survival

4.4 Overall survival

4.4.1 Overall survival function

The survival function was estimated to assess the probability of survival across the cohort. The overall survival function $S(t)$ gives the probability that a subject survives to a specified point in time, t, following transplantation.

Based on the KM estimator, the probability of OS at one year was 0.82 (CI 0.76 - 0.89) which decreased to 0.62 (CI 0.62 - 0.79) after three years. However, the median OS time ($S(t) = 0.5$) wasn't reached until approximately 9.4 years (CI 3.96 - Inf.). These values are illustrated as a survival curve in **Fig. 4.2** and describe the probability of survival across the Day-14 cohort, irrespective of individual patient characteristics.

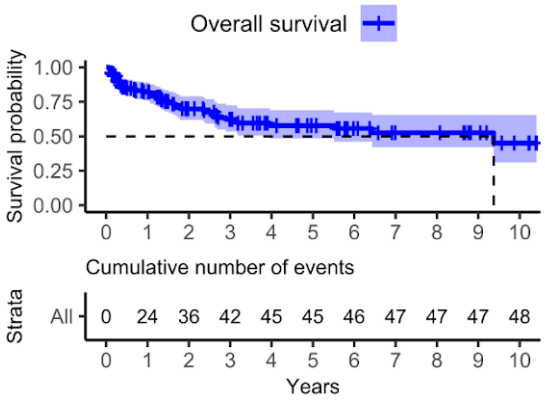


Figure 4.2 Overall survival. Kaplan-Meier estimate with a 95% confidence interval.

4.4.2 Univariate effects on OS

The hazard function $h(t)$ of OS was regressed on the clinical and protein variables (**Table 4.4**). The $h(t)$ of OS is the probability of experiencing mortality at time t for a subject who has survived until

that time. Hazard ratios (HR) represent the relative change in the hazard associated with a 1-unit increase in a covariate. When the covariate in question is categorical, the HR represents the relative change in the hazard compared to a reference group.

Of the clinical variables tested, RIC and having a sex-mismatched donor were associated ($P < 0.05$) with a decrease in the rate of all-cause mortality. Increased concentrations of IL18, PD-L1, CDCP1, CCL23, TNFRSF9, and CSF-1 were associated with higher rates of mortality ($P < 0.05$) and CXCL10 concentrations had an insignificant deleterious effect ($P < 0.1$). Proportional hazard tests (PH tests) based on Schoenfeld-type residuals indicated no violation of the PH assumption (**Table 4.4**).

Table 4.4 Univariate effects on overall survival – serum proteins. Univariate hazard ratios (HR) from the Cox PH model, with 95% confidence interval (CI), significance test (P. val) and proportional hazards test (PH test). $P < 0.05$ and $P < 0.1$ are highlighted in green and blue, respectively.

Variable	Reference	HR	CI	P val.	PH test
Age at transplant / 10		1.16	(0.91 - 1.48)	0.23	0.74
CD34+ dose		1.1	(0.98 - 1.24)	0.11	0.93
D. age / 10		1.01	(0.78 - 1.3)	0.96	0.60
Diagnosis Myeloid	Lymphoid	1.07	(0.58 - 1.97)	0.83	0.98
GvHD prophyl. Combition	Ciclosporin	1.27	(0.71 - 2.27)	0.42	0.90
P. sex Male	Female	1.19	(0.65 - 2.17)	0.57	0.81
RIC Yes	No	0.57	(0.32 - 1.01)	0.06	0.49
D. sex Mismatched	Matched	0.5	(0.27 - 0.92)	0.03	0.21
HCT-CI 1	HCT-CI 0	1.55	(0.58 - 4.12)	0.39	0.65
HCT-CI 2	HCT-CI 0	1.35	(0.57 - 3.16)	0.50	0.65
HCT-CI 3+	HCT-CI 0	1.88	(0.74 - 4.75)	0.21	0.65
TBI Yes	No	1.15	(0.57 - 2.32)	0.70	0.85
ARTN		1.02	(0.79 - 1.31)	0.89	0.58
CASP-8		1.15	(0.86 - 1.55)	0.34	0.63
CCL20		1.17	(0.91 - 1.5)	0.22	0.90
CCL23		1.41	(1.02 - 1.95)	0.04	0.49
CDCP1		1.41	(1.08 - 1.83)	0.01	0.74
CSF-1		1.38	(1.04 - 1.84)	0.03	0.34
CXCL10		1.26	(0.97 - 1.65)	0.09	0.69
Flt3L		0.94	(0.71 - 1.24)	0.65	0.54
IL10		1.16	(0.95 - 1.41)	0.14	0.90
IL-12B		1.22	(0.92 - 1.61)	0.17	0.08
IL-17C		0.92	(0.67 - 1.26)	0.59	0.79
IL18		1.45	(1.09 - 1.94)	0.01	0.64
IL2		1.16	(0.84 - 1.59)	0.37	0.80
IL6		1.1	(0.86 - 1.41)	0.46	0.99
LIF-R		1.14	(0.86 - 1.52)	0.38	0.11
MMP-10		1.04	(0.77 - 1.4)	0.80	0.76
NT-3		1	(0.76 - 1.33)	0.98	0.24
PD-L1		1.44	(1.11 - 1.88)	0.01	0.33
TNFRSF9		1.36	(1.02 - 1.82)	0.04	0.32

4.4.3 Variable selection for OS

Univariate hazard ratios are marginal effects without adjusting for other variables; therefore, a multivariate model is required. However, the number of variables was prohibitively large given the cohort size, therefore a variable selection strategy was employed to reduce the data dimension. I refer the reader to **Appendix B** for an introduction to variable selection.

Determining which variables to include in multivariate analyses is an important decision that can profoundly change the final interpretation of a prognostic study. Established risk factors need to be accounted for so the use of prior knowledge is essential. However, given a limited cohort size, a balance needs to be made between including relevant variables and ensuring a sensible events-per-variable (EPV) ratio, a common holistic intended to limit model overfitting. Simulation studies indicate an EPV ratio > 5 is desirable to limit overfitting to the observed data (361).

Univariate filtering involves selecting variables based on univariate or marginal effects of an arbitrary significance (e.g., $P < 0.1$). This is a common and intuitive approach but may overlook effects that are larger when observed in concert with over variables. Therefore, I chose to apply various variable selection algorithms and compared the resulting models.

Variable selection was conducted with two regularisation strategies, the least absolute shrinkage and selection operator (LASSO) and stepwise selection with the Akaike information criterion (AIC selection). AIC selection was applied in both directions, starting from the null model (forward direction (AIC fw)) and from all candidate variables (backward direction (AIC bw)). In addition, random survival forest variable importance analysis (RSF) was applied, which is a machine learning algorithm. Each of these methods has been utilised in biomarker discovery in allo-HSCT research (349-351) and applied to OLINK proteomics data(352, 353).

All protein and clinical variables were included as candidate predictors of OS. The results of variable selection analysis are summarised in **Table 4.5** with variables ranked from most to least important according to each algorithm. Univariate predictors ($P < 0.1$) are shown ordered by significance. AIC fw and bw selected the same set of predictors, however with a different rank order. Notably, CDCP1 & PD-L1 were selected as predictive of OS by all algorithms. Neither IL10 nor CD34+ had a univariate association with OS but were identified by AIC selection and RSF as being predictive.

Table 4.5 Variable selection results for overall survival. Variables are ranked by order of importance according to univariate filtering, LASSO, AIC selection, and RSF.

Univar. (P < 0.1)	LASSO	AIC fw	AIC bw	RSF
PD-L1	CDCP1	PD-L1	CDCP1	TNFRSF9
IL18	PD-L1	RIC	IL-17C	IL18
CDCP1		Sex matched	LIF-R	CCL23
CSF-1		ARTN	PD-L1	PD-L1
Sex matched		CDCP1	ARTN	Age at trans. / 10
CCL23		CD34+ dose	IL10	LIF-R
TNFRSF9		IL10	RIC	CDCP1
RIC		LIF-R	CD34+ dose	IL2
CXCL10		IL-17C	Sex matched	IL-12B
				CD34+ dose
				IL10

4.4.4 Model selection for overall survival

Variable selection resulted in a set of possible models for OS. An information theoretic approach can be used to compare multiple models simultaneously to determine which resulted in the best fit given the observed data.

The Bayesian Information Criterion (BIC) was developed to assess the fit of a regression model whilst penalising models for greater complexity. Lower values of BIC indicate a better apparent fit after penalisation. The difference in BIC between models (ΔBIC) indicates the level of evidence in favour of the model the lower BIC. Suggested ΔBIC thresholds correspond to weak (0-2), positive (2-6), strong (6-10) or very strong (10+) evidence in favour of the model with the lower BIC (360, 403).

The BIC and EPV ratio were calculated for each model and models were ranked from best to worst fit (**Table 4.6**). The AIC-selected models resulted in the lowest BIC (BIC 416.05) with 9 variables (EPV 5.33). The LASSO model had an incrementally higher BIC (ΔBIC 2.76) but was far less complex (EPV 24). Notably, both AIC and LASSO substantially improved model fit compared to selection based on univariate effects (ΔBIC 17.34) and RSF (ΔBIC 26.92). Only one model, the RSF-selected model, had an EPV < 5 (EPV 4.36).

Based on these comparisons, variables selected by LASSO were used for model development. The final model for OS was fitted adjusting for age, HCT-CI score, and clinical variables with a univariate association with OS (P < 0.1) (Sex matched and RIC). When these variables were included as additional covariates the EPV was 6 and BIC was 427.17.

Table 4.6 Model selection for overall survival. Bayesian information criterion (BIC) and events-per-variable (EPV) ratio by variable selection algorithm, with the number of variables (N var.), degrees of freedom (DF), and difference in BIC (ΔBIC).

Algorithm	N var.	DF	BIC	ΔBIC	EPV
AIC fw/bw	9	9	416.05	0.00	5.33
LASSO	2	2	418.81	2.76	24.00
Univar.	9	9	433.39	17.34	5.33
RSF	11	11	442.97	26.92	4.36

4.4.5 Multivariate effects on OS

The impact of selected variables on the rate of OS was tested using a multivariate CoxPH model (**Figure 4.3**). Adjusted hazard ratios (adj. HR) are multivariate estimates adjusting for other covariates.

Increased concentrations of PD-L1 (adj. HR 1.41; CI 1.01 - 1.97; P 0.05) retained a significant effect on the survival rate in the multivariate comparison. Thus, a 1 std-NPX increase in PD-L1 expression on the 14th day post-transplant was associated with a 41% increase in the mortality rate. The effect of CDCP1, however, was no longer significant (adj. HR 1.28; CI 0.94 - 1.75; P 0.13).

Higher HCT-CI scores, an established risk score for allo-HSCT patients, were associated with an insignificant increase in mortality rates. The adj. HR for an HCT-CI score of 3+ was 1.71 (CI 0.69 - 4.25; P 0.26) compared to those with the lowest score. In contrast, receiving a reduced intensity conditioning regimen (RIC) resulted in a significant protective effect (adj. HR 0.44; CI 0.23 - 0.83; P 0.02), as did having a sex mismatched donor (adj. HR 0.53; CI 0.28 - 0.99; P 0.05).

These findings suggest patients with higher concentrations of PD-L1 in the serum on Day-14 were associated with higher rates of all-cause mortality in this cohort. Whereas individuals had improved

chances of survival if they were subject to a RIC regimen or received transplanted cells from a donor of the opposite sex.

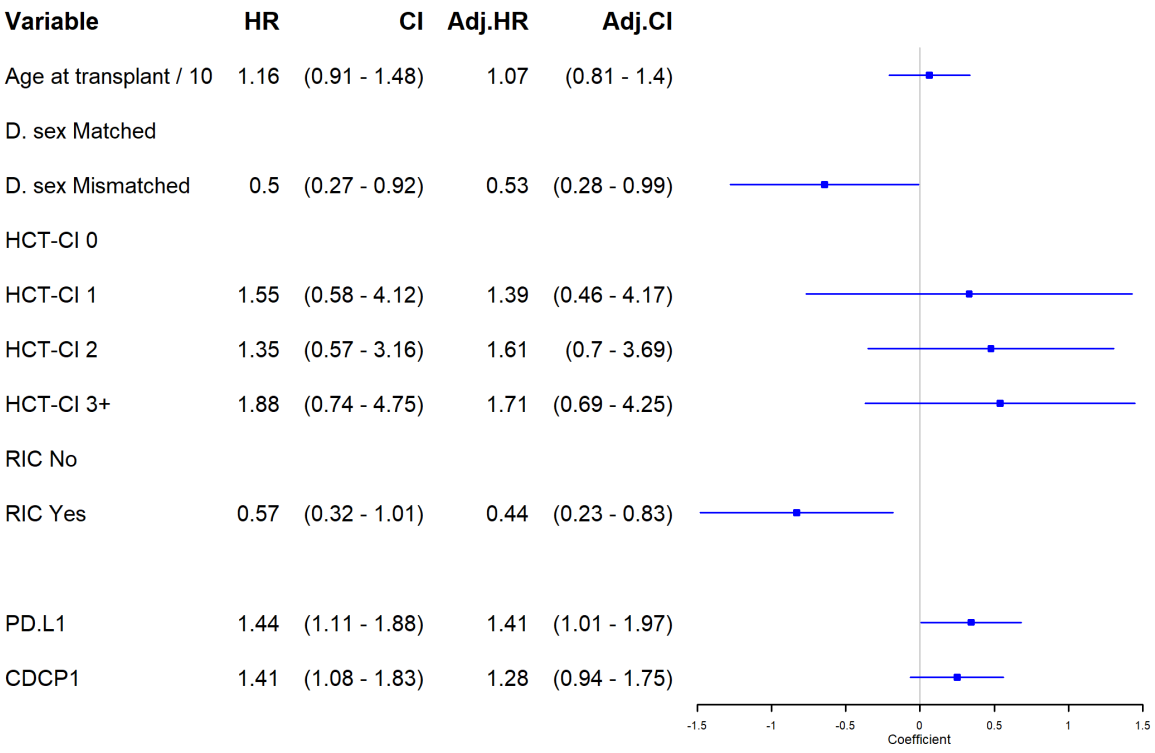


Figure 4.3 Overall survival associated serum proteins. Univariate (HR) and multivariate hazard ratios (adj. HR) with a 95% confidence interval (CI) based on the Cox PH model. Multivariate regression coefficients are visualised as a forest plot.

4.5 GvHD-free, relapse-free survival

OS is a direct measure of clinical benefit that is unambiguous and easy to measure as the final endpoint is the time of death. However, it requires a long length of follow-up and fails to capture ongoing morbidity which profoundly impacts a patient's quality of life.

The composite outcome of GRFS was developed to study the outcome of allogeneic-HSCT patients. It represents the time to first-event onset including AGvHD (grade II-IV), CGvHD, relapse, or death. By definition, GRFS events occur more quickly as it also includes non-fatal events that occur with a higher frequency early after transplantation. As a result, GRFS requires a shorter follow-up and captures the real recovery of patients without adverse events.

4.5.1 GvHD-free, relapse-free survival function

The $\mathrm{GRFS}(t)$ gives the probability of surviving event-free until a specified point in time, t, after transplantation. The median GRFS time was 0.42 years (CI 0.44 - 0.8, IQR 0.23 - 4.65) indicating that only half of the patients remained event-free after the first four months. At one-year post-transplant the probability of GRFS was 0.36 (CI 0.29 - 0.46) which dropped only marginally to 0.28 (CI 0.21 - 0.37) after three years as the majority of events occurred within the first 12 months of follow-up. The probability of surviving event-free is illustrated as a survival curve in **Fig. 4.4**.

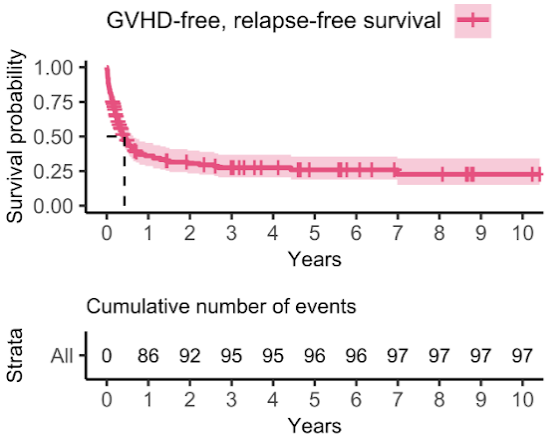


Figure 4.4 GvHD-free, relapse-free survival. Kaplan-Meier estimate with a 95% confidence interval.

4.5.2 Univariate effects on GRFS

The hazard function $h(t)$ of GRFS was regressed on the clinical and protein variables (**Table 4.7**). The hazard $h(t)$ of GRFS is the probability of experiencing an event (AGvHD (grade II-IV), CGvHD, relapse, or death) at time t for a subject who has survived event-free until that time.

Of the clinical variables, having a high HCT-CI score was associated (P < 0.05) with an increase in the hazard of GRFS. Increased concentrations of CCL23 were associated with a higher rate of events (P < 0.05) and PD-L1 concentrations had a marginally deleterious effect (P < 0.1). PH tests indicated no significant violation of the PH assumption.

Table 4.7 Univariate effects on GvHD-free, relapse-free survival – serum proteins. Hazard ratios (HR) with a 95% confidence interval (CI) based on the Cox PH model. P < 0.05 and P < 0.1 are highlighted in green and blue, respectively. Proportional hazards (PH) tests are shown for each variable.

Variable	Reference	HR	CI	P val.	PH test
Age at transplant / 10		1.12	(0.94 - 1.32)	0.21	0.60
CD34+ dose		1.06	(0.95 - 1.19)	0.30	0.09
D. age / 10		1.08	(0.93 - 1.26)	0.33	0.07
Diagnosis Myeloid	Lymphoid	1.22	(0.79 - 1.89)	0.37	0.51
GvHD prophyl. Combition	Ciclosporin	1.05	(0.69 - 1.6)	0.82	1.00
P. sex Male	Female	1.04	(0.69 - 1.57)	0.85	0.64
RIC Yes	No	0.82	(0.55 - 1.22)	0.33	0.42
D. sex Mismatched	Matched	0.87	(0.58 - 1.31)	0.52	0.53
HCT-CI 1	HCT-CI 0	1.27	(0.65 - 2.49)	0.49	0.93
HCT-CI 2	HCT-CI 0	1.1	(0.61 - 1.99)	0.75	0.93
HCT-CI 3+	HCT-CI 0	2	(1.14 - 3.52)	0.02	0.93
TBI Yes	No	1.21	(0.73 - 2)	0.45	1.00
ARTN		1	(0.85 - 1.19)	0.99	0.40
CASP-8		1.01	(0.83 - 1.24)	0.90	0.17
CCL20		1.09	(0.91 - 1.31)	0.36	0.15
CCL23		1.29	(1.06 - 1.58)	0.01	0.54
CDCP1		1.05	(0.87 - 1.27)	0.61	0.54
CSF-1		1.04	(0.86 - 1.28)	0.67	0.61
CXCL10		0.99	(0.8 - 1.21)	0.90	0.76
Flt3L		0.91	(0.74 - 1.11)	0.34	0.16

IL10		1.1	(0.91 - 1.34)	0.32	0.83
IL-12B		1.13	(0.92 - 1.39)	0.26	0.83
IL-17C		0.92	(0.75 - 1.13)	0.44	0.68
IL18		1.14	(0.93 - 1.4)	0.21	0.29
IL2		0.96	(0.77 - 1.21)	0.76	0.28
IL6		0.96	(0.79 - 1.17)	0.71	0.13
LIF-R		1.08	(0.89 - 1.31)	0.46	0.69
MMP-10		0.87	(0.71 - 1.08)	0.21	0.99
NT-3		1.03	(0.86 - 1.23)	0.75	0.29
PD-L1		1.2	(0.99 - 1.45)	0.07	0.61
TNFRSF9		1.05	(0.86 - 1.28)	0.62	0.49

4.5.3 Variable selection for GRFS

A variable selection strategy was employed to determine which subset of variables to include in the multivariate model of GRFS. All protein and clinical variables were included as candidate predictors of GRFS. The results of variable selection analysis for GRFS are summarised in **Table 4.8**.

CCL23 & HCT-CI were selected by all algorithms as being predictive of GRFS. The results of RSF and LASSO were the same except RSF selected two additional clinical variables, age at transplant and donor age. HCT-CI and CCL23 were selected by AIC in both directions; however, AIC fw selected MMP-10 whereas backward selection favoured IL-17C, neither of which had a univariate effect on GRFS.

Table 4.8 Variable selection results for GvHD-free, relapse-free survival. Variables are ranked by order of importance according to univariate filtering, LASSO, AIC selection, and RSF.

Univariate (P < 0.1)	LASSO	AIC fw	AIC bw	RSF
CCL23	HCT-CI	HCT-CI	IL-17C	HCT-CI
HCT-CI	CCL23	CCL23	CCL23	Age at transplant / 10
PD-L1		MMP-10	HCT-CI	CCL23
				D. age / 10

4.5.4 Model selection for GRFS

To determine which set of variable selection results to use for final model development, BIC and EPV ratios were calculated for each possible model (**Table 4.9**). The LASSO selected variables achieved the lowest BIC (BIC 854.70) and resulted in the least complex model (EPV 24.25). The AIC fw model had an incrementally higher BIC (ΔBIC 1.14) with one additional variable, MMP-10 (EPV 19.40) indicating weak evidence in favour of the LASSO model. There was positive evidence in favour of the LASSO model when compared to the AIC bw (ΔBIC 2.16) and univariate selected models (ΔBIC 4.43), and strong evidence against the results of RSF selection (ΔBIC 7.80) when compared to the LASSO model. Thus, BIC indicated the LASSO selected model had the best performance at predicting this particular set of outcome data.

GRFS events are relatively frequent compared with OS as it also describes non-fatal outcomes, as a result, the EPV ratios were more favourable compared to those reported in **Section 4.4.4**. All of the variable selection algorithms resulted in an EPV >10.

Given these comparisons, variables selected by LASSO were utilised for model development. When age was included as an additional covariate the EPV was 19.4 and BIC was 858.3455.

Table 4.9 Model selection for GvHD-free, relapse-free survival. Bayesian information criterion (BIC) and events-per-variable (EPV) ratio by variable selection algorithm, with the number of variables (N var.), degrees of freedom (DF), and difference in BIC (ΔBIC).

Variables	N var.	DF	BIC	ΔBIC	EPV
LASSO	2	4	854.70	0.00	24.25
AIC fw	3	5	855.84	1.14	19.40
AIC bw	3	5	856.85	2.16	19.40
Univar.	3	5	859.13	4.43	19.40
RSF	4	6	862.50	7.80	16.17

4.5.5 Multivariate effects on GRFS

The impact of selected variables on the rate of GRFS was tested using a multivariate CoxPH model (**Fig. 5**). Increased concentrations of CCL23 (adj. HR 1.3; CI 1.06-1.59; P 0.01) retained a significant effect on the GRFS rate in the multivariate comparison. Suggesting a 1 std-NPX increase in CCL23 expression at Day 14 was associated with a 30% increase in the rate of GvHD, relapse, or death. Higher HCT-CI scores were also associated with a significant increase in the hazard of GRFS. For an HCT-CI score of 3+ the adj. HR was 1.94 (CI 1.16-3.6; P 0.01) compared to those with a score of 0.

Thus, the model indicates high concentrations of CCL23 in the serum 14 days post-transplant were associated with increased rates of GvHD, relapse, or death. The HCT-CI score, originally developed for predicting all-cause mortality, may also be associated with the rate of GRFS.

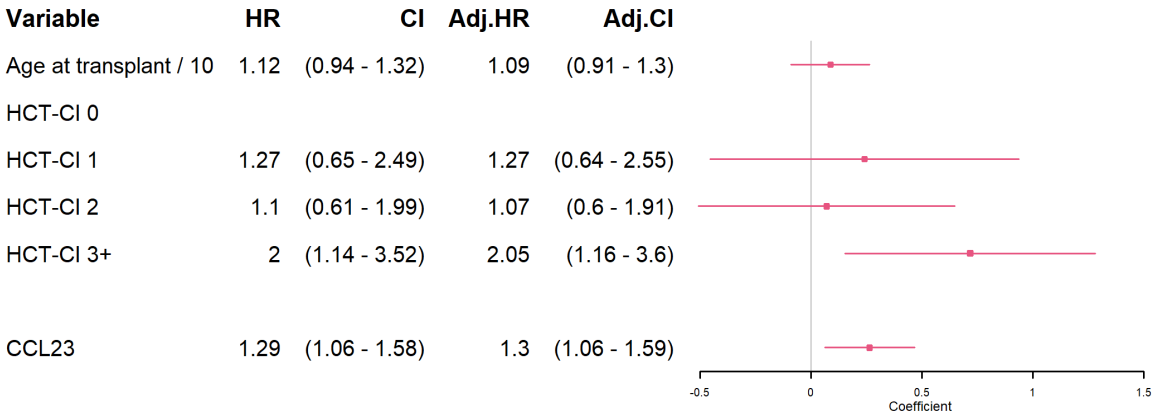


Figure 4.5 GvHD-free, relapse-free survival associated serum proteins. Univariate (HR) and multivariate hazard ratios (adj. HR) with a 95% confidence interval (CI) based on the Cox PH model. Multivariate regression coefficients are visualised as a forest plot.

4.6 Non-relapse and relapse-related mortality

Overall and GvHD-free, relapse-free survival describe successful and unsuccessful responses to therapy; however, they don't provide insight into the causes of mortality or type of events experienced. Allo-HSCT can cure through the cytotoxicity of the conditioning regimen or through the GvL effect. However, both effects are intrinsically linked to the toxicity of the therapy. It is therefore important to estimate the association of candidate biomarkers with the various competing causes of mortality.

The model development process was repeated in a competing risks framework with survival separated into relapse-related mortality (RM, death following disease relapse) and non-relapse mortality (NRM, death without disease relapse). Within the NRM group (N = 28), 14 subjects were diagnosed with AGvHD and 6 with CGvHD, including 2 diagnosed with both forms. Whereas, in the RM group (N = 20), all of whom relapsed, 4 also received a diagnosis of AGvHD, and 4 experienced CGvHD.

4.6.1 Cumulative incidence of RM and NRM

The cumulative incidence $CIF(t)$ of RM and NRM were estimated for the Day-14 cohort (**Figure 4.6**). This describes the probability of experiencing the event of interest, before time t, in subjects yet to experience a competing event. The one-year incidence of RM was 0.07 (CI 0.03 - 0.11) which rose substantially after three years to 0.18 (CI 0.10-0.25). For NRM the one-year incidence was 0.11 (CI 0.05 - 0.16) and increased to 0.2 (CI 0.12 - 0.28) three years after transplantation.

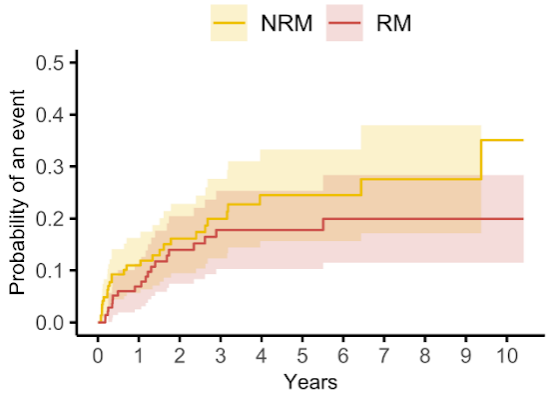


Figure 4.6 Cumulative incidence of non-relapse and relapse-related mortality. Cumulative incidence estimates with a 95% confidence interval.

4.6.2 Univariate effects on RM and NRM

The subdistribution hazard $\lambda_k^{SD}(t)$ of RM and NRM were regressed on clinical and protein variables with the FG model (**Table 4.10**). The $\lambda_k^{SD}(t)$ equates to the probability of experiencing the event at time t, for a subject who has survived until that time or who has experienced a competing event. Subdistribution hazard ratios (HR_{SD}) estimate the change in the $\lambda_k^{SD}(t)$ per unit change in the variable (or between groups).

As this resists an intuitive interpretation, estimates are often described in terms of the effect it has on the cumulative incidence. A significant HR_{SD}, indicates a significant change in the cumulative incidence (probability) of the event, however, the specific magnitude cannot be interpreted.

Of the clinical variables, having a myeloid diagnosis was associated ($P < 0.05$) with an increase in the incidence of RM. So too were increased concentrations of CDCP1, TNFRSF9, and LIF-R ($P < 0.05$). CCL23 concentrations also had an insignificant deleterious effect ($P < 0.1$). In contrast, IL-17C concentrations

had a protective effect on RM incidence (P < 0.05). PH tests based on the cumulative sum of residuals indicated IL-12B and IL2 violated the PH assumption (P < 0.05).

This was repeated for the NRM, the competing risk. Myeloid diagnosis (P 0.06) and having a sex-mismatched donor (P 0.05) were associated with a decrease in the incidence of NRM, whereas NRM incidence increased insignificantly with higher HCT-CI scores (P 0.07). Increased concentrations of CSF-1, IL12-B, CXCL10, and IL6 (P < 0.05) were also associated with increased NRM incidence and there was an insignificant increase with higher PD-L1 and IL2 concentrations (P <0.1). PH tests indicated no violation of the PH assumption.

Table 4.10 Univariate effects on relapse-related and non-relapse mortality – serum proteins. Subdistribution hazard ratios (HR_{SD}) with a 95% confidence interval (CI) based on the FG model. P < 0.05 and P < 0.1 are highlighted in green and blue, respectively. Proportional hazards (PH) tests are shown for each variable.

		Relapse-related mortality				Non-relapse mortality			
Variable	**Reference**	**HR_{SD}**	**CI**	**P val.**	**PH test**	**HR_{SD}**	**CI**	**P val.**	**PH test**
Age at transplant / 10		1.29	(0.86 - 1.94)	0.22	0.066	1.04	(0.76 - 1.43)	0.8	0.276
CD34+ dose		1.1	(0.95 - 1.27)	0.2	0.233	1.08	(0.93 - 1.26)	0.32	0.122
D. age / 10		1.02	(0.67 - 1.56)	0.91	0.472	1	(0.75 - 1.32)	0.98	0.403
Diagnosis Myeloid	Lymphoid	4.81	(1.16 - 19.99)	0.03	0.387	0.49	(0.24 - 1.03)	0.06	0.763
GvHD prophyl. Combination	Ciclosporin	1.31	(0.54 - 3.2)	0.55	0.305	1.21	(0.58 - 2.53)	0.61	0.252
P. sex Male	Female	1.76	(0.63 - 4.89)	0.28	0.178	0.87	(0.42 - 1.84)	0.72	0.379
RIC Yes	No	0.52	(0.22 - 1.23)	0.13	0.362	0.74	(0.36 - 1.54)	0.42	0.437
D. sex Mismatched	Matched	0.8	(0.32 - 1.97)	0.63	0.065	0.4	(0.17 - 0.94)	0.04	0.96
HCT-CI 1	HCT-CI 0	1.69	(0.37 - 7.72)	0.5	0.346	1.21	(0.3 - 4.87)	0.79	0.682
HCT-CI 2	HCT-CI 0	1.27	(0.36 - 4.47)	0.71	0.218	1.26	(0.42 - 3.75)	0.68	0.305
HCT-CI 3+	HCT-CI 0	0.51	(0.08 - 3.11)	0.46	0.542	2.88	(0.95 - 8.77)	0.07	0.452
TBI Yes	No	1.13	(0.37 - 3.47)	0.83	0.226	1.15	(0.49 - 2.7)	0.75	0.726
ARTN		1.03	(0.78 - 1.37)	0.81	0.533	1.02	(0.8 - 1.3)	0.9	0.717
CASP-8		1.27	(0.89 - 1.81)	0.19	0.108	1.08	(0.77 - 1.53)	0.65	0.295
CCL20		1.08	(0.73 - 1.58)	0.7	0.515	1.21	(0.9 - 1.62)	0.22	0.324
CCL23		1.51	(0.97 - 2.36)	0.07	0.228	1.2	(0.77 - 1.87)	0.43	0.822
CDCP1		1.84	(1.24 - 2.73)	<0.01	0.261	1.1	(0.82 - 1.46)	0.53	0.342
CSF-1		1.02	(0.66 - 1.58)	0.91	0.1	1.58	(1.1 - 2.27)	0.01	0.125
CXCL10		1.03	(0.73 - 1.44)	0.87	0.108	1.42	(1.01 - 1.98)	0.04	0.89
Flt3L		0.85	(0.6 - 1.2)	0.36	0.148	1.03	(0.71 - 1.48)	0.89	0.458
IL10		1.2	(0.8 - 1.78)	0.37	0.085	1.21	(0.87 - 1.69)	0.25	0.158
IL-12B		0.96	(0.74 - 1.26)	0.79	0.024	1.27	(1.05 - 1.55)	0.02	0.333
IL-17C		0.62	(0.38 - 0.99)	0.05	0.45	1.14	(0.8 - 1.63)	0.46	0.171
IL18		1.36	(0.91 - 2.02)	0.14	0.537	1.34	(0.89 - 2.02)	0.16	0.206
IL2		0.8	(0.5 - 1.26)	0.33	0.005	1.45	(0.93 - 2.26)	0.1	0.355
IL6		0.75	(0.5 - 1.11)	0.15	0.081	1.35	(1.01 - 1.79)	0.04	0.188
LIF-R		1.57	(1.02 - 2.43)	0.04	0.23	0.89	(0.62 - 1.28)	0.53	0.33
MMP-10		0.78	(0.49 - 1.25)	0.3	0.232	1.26	(0.86 - 1.85)	0.23	0.434
NT-3		1.02	(0.68 - 1.52)	0.93	0.127	1	(0.74 - 1.36)	0.99	0.196
PD-L1		1.29	(0.88 - 1.89)	0.19	0.227	1.4	(0.97 - 2.02)	0.07	0.068
TNFRSF9		1.5	(1.03 - 2.18)	0.03	0.093	1.08	(0.69 - 1.69)	0.75	0.085

4.6.3 Variable selection for RM and NRM

A variable selection strategy was employed to determine which subset of variables to include in the multivariate models. This was repeated for each competing event, RM and NRM.

Variable selection was conducted for RM and the results are summarised in **Table 4.11**. Both AIC and RSF selected a relatively large number of variables, particularly when compared to the number of variables found to have a univariate association with RM ($P < 0.1$). This was also true for LASSO, however, all of the protein variables selected by LASSO had a univariate effect ($P < 0.1$) on RM. LASSO selection indicated that concentrations of CDCP1, LIF-R, TNFRSF9, CCL23 were predictive of RM. IL-17C was the only univariate predictor of RM excluded by LASSO. LASSO selection also indicated diagnosis, HCT-CI, TBI, patient sex, age, and CD34+ dose should be retained in the RM model.

Table 4.11 Variable selection results for relapse-related mortality. Variables are ranked by order of importance according to univariate filtering, LASSO, AIC selection, and RSF.

Univar. (P < 0.1)	**LASSO**	**RSF**	**AIC fw**	**AIC bw**
Diagnosis	CDCP1	TNFRSF9	Diagnosis	Diagnosis
CDCP1	LIF-R	CDCP1	TBI	TBI
IL-17C	CCL23	LIF-R	TNFRSF9	TNFRSF9
LIF-R	TNFRSF9	IL18	LIF-R	LIF-R
CCL23	Diagnosis	D. age /10	PD-L1	PD-L1
TNFRSF9	HCT-CI	IL-17C	D. age /10	D. age /10
	TBI	IL6	CCL20	CCL20
	P. Sex	CCL23	P. Sex	P. Sex
	CD34+ dose	CD34+ dose	GvHD Prophylaxis	GvHD Prophylaxis
	Age at transplant /10	NT-3	CD34+ dose	CD34+ dose
		CSF-1	CXCL10	CXCL10
		MMP-10	ARTN	ARTN
			IL10	IL10
			Sex matched	Sex matched
			IL18	IL18
			HCT-CI	HCT-CI
			IL6	IL6
			RIC	RIC

When applied to select predictors of NRM (**Table 4.12**), a similar pattern emerged where AIC and RSF selected a disproportionately large number of variables relative to the number of univariate predictors. LASSO indicated that concentrations of IL6, IL2, IL10, CXCL10, and CSF-1 were predictive of NRM. These protein variables were a subset of the univariate predictors of NRM ($P < 0.1$). LASSO also indicated that HCT-CI score and CD34+ dose were predictive of NRM.

Table 4.12 Variable selection results for non-relapse mortality. Variables ranked by order of importance according to univariate filtering, LASSO, AIC selection, and RSF.

Univar. (P < 0.1)	LASSO	RSF	AIC fw	AIC bw
HCT-CI	IL6	IL18	HCT-CI	HCT-CI
D. sex matched	IL2	IL2	GvHD Prophylaxis	GvHD Prophylaxis
Diagnosis	PD-L1	PD-L1	IL6	IL6
CSF-1	IL10	HCT-CI	IL18	IL18
IL2	CXCL10	Age at transplant /10	IL10	IL10
CXCL10	CSF-1	IL10	PD-L1	PD-L1
PD-L1	HCT-CI	TNFRSF9	CXCL10	CXCL10
IL6	CD34+ dose	Diagnosis	D. age /10	D. age /10
IL10		IL-17C	CD34+ dose	CD34+ dose
		LIF-R	TNFRSF9	TNFRSF9
		IL6	ARTN	ARTN
		CCL23	D. sex matched	D. sex matched
		CXCL10	PSex	PSex
		CSF-1	CCL20	CCL20
		IL-12B	LIF-R	LIF-R
			RIC	RIC
			TBI	TBI
			Diagnosis	Diagnosis

The total number of deaths (N = 48) is divided between each competing event, in this case, relapse-related (N =20) and non-relapse mortality (N = 28). As a result, EPV ratios are substantially reduced. Reducing model complexity is therefore a priority. In this sense none of the algorithms applied were successful. Particularly in the case of AIC and RSF which selected a disproportionately large number of variables, this could indicate the algorithms overfitting to noise variables. As a result, the AIC and RSF results were discounted from onward analysis, and model selection was conducted focusing on the results of univariate and LASSO selection.

4.6.4 Model selection for RM and NRM

To determine which set of variable selection results to use for final model development, BIC and EPV ratios were calculated for each possible model of RM (**Table 4.13**) and NRM (**Table 4.14**). Multivariate PH tests indicated IL2 and IL10 violated the PH assumption (P < 0.01; P 0.02) and were therefore excluded.

For models of RM, univariate selection (P < 0.1) achieved the lowest BIC (BIC 178.88; EPV 3.33). LASSO selection resulted in a substantially higher BIC (ΔBIC 15.96), indicating strong evidence in favour of the simpler univariate selected model which included 3 fewer clinical variables but with the addition of IL-17C. When NRM was the outcome, the univariate predictors also resulted in the lowest BIC (BIC 254.65, EPV 2.89). However, the BIC was only incrementally higher for the LASSO selected model (ΔBIC 3.17), indicating positive but not strong evidence in favour of the univariate selected model.

All of the models tested had unfavourable EPV ratios (< 5) indicating that the number of events observed may be insufficient to estimate unbiased multivariate subdistribution hazard ratios in this cohort. This is further compounded by the need to test the influence of serum proteins on both competing risks.

The model with the best predictive power for one competing event is likely to be distinct from that of the competing cause of failure. As a result, the variable selection analysis selected a different set of variables for RM and NRM. However, when analysing competing events, it is recommended that one interprets the effect of variables on each competing event(223). This is because a change in the subdistribution hazard for one event may be determined by a substantial change in the rate of the competing risk.

A decision was made to use the variable selection method that identified the least protein variables, LASSO selection. The effects were then compared to a model fitted based on univariate filtering to ensure that the same conclusions would have been drawn. However, given the low EPV values that resulted after combining variable selection results (EPV 1.25 for RM, EPV 1.63 for NRM), there is a high risk of overfitting.

Table 4.13 Model selection for relapse-related mortality. Bayesian information criterion (BIC) and events-per-variable (EPV) ratio by variable selection algorithm, with the number of variables (N var.) and difference in BIC (ΔBIC).

Variables	N var.	BIC	EPV	ΔBIC
Univar.	6	178.88	3.33	
LASSO	10	194.84	1.67	15.96
Clin.	10	201.31	1.67	22.43

Table 4.14 Model selection for non-relapse mortality. Bayesian information criterion (BIC) and events-per-variable (EPV) ratio by variable selection algorithm, with the number of variables (N var.) and difference in BIC (ΔBIC).

Variables	N var.	BIC	EPV	ΔBIC
Univar.	7	254.65	2.89	
LASSO	6	257.82	3.25	3.17
Clin.	10	263.35	2.17	8.69

4.6.5 Multivariate effects on RM and NRM

The impact of selected variables on the incidence of RM was tested using a multivariate FG model (**Table 4.15**). The process was repeated for NRM (**Table 4.15**). The adjusted subdistribution hazard ratios (adj. HR_{SD}) are the multivariate estimates after adjusting for covariates. Given the low EPV ratios, the adj. HR_{SD} need to be interpreted with caution as there is a high probability of overfitting.

Increased CDCP-1 concentration was associated with an increase in the rate of RM (adj. HR_{SD} 2.41; CI 1.27 - 4.56; P 0.01) and an insignificant increase in NRM (adj. HR_{SD} 1.26; CI 0.85 – 1.94; P 0.29), as was the case in the univariate comparison. Whereas higher concentrations of IL6 were associated with lower rates of RM (adj. HR_{SD} 0.47; CI 0.22 - 1; P 0.05) but had the opposite effect on NRM (adj. HR_{SD} 1.5; CI 1 - 2.23; P 0.05); a similar insignificant trend was observed for CXCL10 expression when regressed on RM (adj. HR_{SD} 0.57; CI 0.32 - 1.02; P 0.06). However, it did not retain a significant effect on NRM in the multivariate comparison (adj. HR_{SD} 1.29; CI 0.76 - 2.2; P 0.341).

Whilst the direction of the effects of LIF-R remained the same, the significance changed substantially between univariate and multivariate estimates. Higher concentrations of LIF-R were associated with lower rates of NRM in the multivariate comparison (adj. HR_{SD} 0.47; CI 0.28 - 0.77; $P < 0.01$). However, this effect was not significant in the univariate test (HR_{SD} 0.89, CI 0.62 - 1.28, P 0.53). In contrast, LIF-

R had an uncertain effect on RM in the multivariate model (adj. HR_{SD} 1.48; CI 0.66 - 3.33; P 0.345) but had a univariate association with increased rates of RM (HR_{SD} 1.57; CI 1.02 - 2.43; P 0.042).

Patients with an HCT-CI score of 3+ experienced higher NRM incidence (adj. HR_{SD} 4.9; CI 1.17 - 20.59; P 0.04), as did those with increased CD34+ doses (adj. HR_{SD} 1.24; CI 1.02 - 1.49; P 0.03). Whereas patients with a Myeloid diagnosis had lower rates of NRM (adj. HR_{SD} 0.27; CI 0.1 - 0.73; P 0.01) compared to a Lymphoid diagnosis but higher rates of RM (adj. HR_{SD} 8.35; CI 1.87 – 37.32; P < 0.01).

These models suggest higher concentrations of CDCP1 in the serum at Day-14 post-SCT were associated with increased incidence of RM. Moreover, IL6 and CXCL10 concentrations were linked with protection from RM incidence but an increase in the risk of death from other causes. Subjects with higher HCT-CI scores and CD34+ doses were more likely to experience NRM. In contrast, patients with a myeloid diagnosis had a lower NRM incidence but a higher probability of death following relapse.

Table 4.15 Multivariate effects on relapse-related and non-relapse mortality – serum proteins. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with confidence interval (CI), significance test (P. val) and proportional hazards test (PH test). P < 0.05 and P < 0.1 are highlighted in green and blue, respectively.

	Relapse-related mortality				Non-relapse mortality			
Variable	**Adj. HR_{SD}**	**CI**	**P val.**	**PH test**	**Adj. HR_{SD}**	**CI**	**P val.**	**PH test**
Age at transplant / 10	1.19	(0.69 - 2.06)	0.54	0.10	1.11	(0.69 - 1.8)	0.66	0.40
CD34+ dose	1.02	(0.82 - 1.26)	0.88	0.26	1.24	(1.02 - 1.49)	0.03	0.30
Diagnosis Lymphoid								
Diagnosis Myeloid	8.35	(1.87 - 37.32)	0.01	0.32	0.27	(0.1 - 0.73)	0.01	0.86
HCT-CI 0								
HCT-CI 1	1.24	(0.15 - 10.16)	0.85	0.39	0.83	(0.11 - 6.24)	0.86	0.50
HCT-CI 2	2.22	(0.37 - 13.42)	0.39	0.33	1.22	(0.31 - 4.87)	0.77	0.34
HCT-CI 3+	0.32	(0.03 - 3.96)	0.39	0.46	4.9	(1.17 - 20.59)	0.04	0.49
P. sex Female								
P. sex Male	1.7	(0.4 - 7.14)	0.47	0.30	0.94	(0.35 - 2.49)	0.90	0.49
TBI No								
TBI Yes	3.98	(0.51 - 30.76)	0.19	0.28	1.35	(0.39 - 4.7)	0.63	0.70
CCL23	1.94	(0.78 - 4.87)	0.16	0.28	1.11	(0.62 - 1.97)	0.73	0.92
CDCP1	2.41	(1.27 - 4.56)	0.01	0.50	1.26	(0.82 - 1.94)	0.29	0.60
CSF-1	0.93	(0.34 - 2.54)	0.88	0.12	1.57	(0.95 - 2.6)	0.08	0.09
CXCL10	0.57	(0.32 - 1.02)	0.06	0.08	1.29	(0.76 - 2.2)	0.34	0.97
IL6	0.47	(0.22 - 1)	0.05	0.07	1.5	(1 - 2.23)	0.05	0.46
LIF-R	1.48	(0.66 - 3.3)	0.35	0.31	0.47	(0.28 - 0.77)	< 0.01	0.19
PD-L1	0.6	(0.24 - 1.5)	0.28	0.35	1.35	(0.71 - 2.55)	0.36	0.06
TNFRSF9	2.39	(0.81 - 7.03)	0.11	0.12	0.75	(0.4 - 1.41)	0.37	0.28

Model selection indicated that fitting the FG model with univariate filtering resulted in a slight improvement in the fit; therefore, a second model was fitted, and the effects were then compared to ensure that the same conclusions would have been drawn (**Table 4.16**). Note, this model included the same protein variables as the LASSO model with the addition of IL17C. IL17C, had no clear effect on either RM (adj. HR_{SD} 0.46; CI 0.14 - 1.54; P 0.213) or NRM (adj. HR_{SD} 0.8; CI 0.51 - 1.25; P 0.326) supporting its exclusion. One protein variable for which the effect size substantially changed was CCL23 which was associated with a greater increase in RM in this model (adj. HR_{SD} 2.55; CI 1 - 6.47)

whereas its effect on NRM remained uncertain (adj. HR_{SD} 1.04; 0.64 - 1.71). Thus, the LASSO and univariate selected models were in agreement as to the direction and significance of the main effects, however some of the estimates differed in magnitude.

Table 4.16 Alternative multivariate effects on relapse-related and non-relapse mortality after univariate filtering (P < 0.1) – serum proteins. Multivariate subdistribution hazard ratios (adj. HR_{SD}) with a 95% confidence interval (CI) based on the FG model. Proportional hazards (PH) tests are shown for each variable.

	Relapse-related mortality				Non-relapse mortality			
Variable	**Adj. HR_{SD}**	**CI**	**P val.**	**PH test**	**Adj. HR_{SD}**	**CI**	**P val.**	**PH test**
D. sex Matched								
D. sex Mismatched	0.57	(0.2 - 1.68)	0.31	0.043	0.39	(0.14 - 1.09)	0.074	0.901
Diagnosis Lymphoid								
Diagnosis Myeloid	5.99	(1.5 - 23.92)	0.011	0.306	0.36	(0.16 - 0.82)	0.015	0.717
HCT-CI 0								
HCT-CI 1	2.3	(0.19 - 28.09)	0.528	0.391	0.84	(0.1 - 6.82)	0.872	0.454
HCT-CI 2	3.64	(0.71 - 18.67)	0.123	0.277	1.34	(0.36 - 4.92)	0.662	0.405
HCT-CI 3+	0.66	(0.1 - 4.5)	0.671	0.41	3.39	(0.86 - 13.33)	0.092	0.43
CSF-1	0.68	(0.25 - 1.9)	0.467	0.113	1.53	(0.92 - 2.54)	0.099	0.056
CXCL10	0.64	(0.4 - 1.04)	0.07	0.181	1.46	(0.85 - 2.49)	0.166	0.959
PD-L1	0.64	(0.28 - 1.43)	0.276	0.468	1.4	(0.84 - 2.33)	0.192	0.085
IL6	0.58	(0.29 - 1.17)	0.13	0.054	1.32	(0.86 - 2.03)	0.199	0.527
CDCP1	2.29	(1.21 - 4.35)	0.012	0.544	1.2	(0.78 - 1.86)	0.412	0.47
CCL23	2.55	(1 - 6.47)	0.05	0.378	1.04	(0.64 - 1.71)	0.861	0.744
IL.17C	0.46	(0.14 - 1.54)	0.213	0.502	0.8	(0.51 - 1.25)	0.326	0.505
TNFRSF9	2.28	(0.85 - 6.06)	0.1	0.119	0.78	(0.42 - 1.45)	0.436	0.289
LIF-R	1.43	(0.7 - 2.9)	0.324	0.364	0.53	(0.31 - 0.9)	0.019	0.194

4.7 Candidate prognostic biomarkers associated with primary clinical endpoints

To recap the results reported for the primary clinical endpoints, two of the protein variables, PD-L1, and CCL23, were associated with an increase in the rate of OS and GRFS, respectively. The concentration of these proteins was associated with survival in both univariate and multivariate comparisons and were selected unanimously by the variable selection algorithms.

The competing risks analysis studied the cumulative incidence of NRM and RM. Univariate and multivariate FG models indicated CDCP1 was associated with an increase in the RM incidence, with no corresponding decrease in NRM. Whereas IL6 had different effects on the two competing risks, associated with increasing NRM incidence, but a decrease in RM.

To gain a complete understanding of the event dynamics, I next tested the effects of these proteins on the incidence of secondary endpoints including AGvHD, CGvHD, and relapse. To reduce the risk of interpreting spurious findings, multivariate analysis of secondary endpoints was restricted to protein variables selected for their association with OS, GRFS, RM, and NRM.

4.8 Acute graft-vs-host disease

Broadly, GvHD occurs in two main forms, acute (AGvHD) and chronic (CGvHD). These were originally distinguished by the time of onset, with AGvHD occurring in the first 100 days post-transplant. Since the introduction of RIC, AGvHD has been known to occur at later time points, and the definition has been updated to include clinical manifestations (75). AGvHD is the leading cause of short-term (day 100 and 1 year) mortality after allo-HSCT (75). Whereas CGvHD is a major cause of late non-relapse mortality that usually begins between 3 months and 2 years post-transplant (78). As these events happen at different time scales and have distinct parthenogeneses, the influence of serum proteins on the incidence of each form of GvHD was tested.

4.8.1 Cumulative incidence of AGvHD

The $CIF(t)$ of AGvHD (grade II-IV) was estimated as a competing risk with death (without AGvHD). The incidence of AGvHD at day 100 was 0.24 (CI 0.17 - 0.31) and reached a maximum of 0.29 (CI 0.21 - 0.37) within 5 months (**Figure 4.7**). In total, 41 subjects experienced AGvHD within the follow-up.

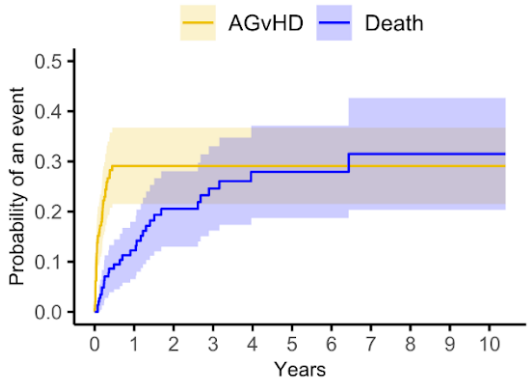


Figure 4.7 Cumulative incidence of acute graft-vs-host disease. Cumulative incidence estimates with a 95% confidence interval. Death (without AGvHD) is the competing risk.

4.8.2 Univariate effects on AGvHD

The subdistribution hazard $\lambda_k^{SD}(t)$ of AGvHD & death (without AGvHD) were regressed on clinical and protein variables with the FG model (**Table 4.17**). In this analysis, AGvHD is the outcome of interest and the estimates for death (without AGvHD) are shown for comparative purposes. This is to ensure differences in the AGvHD were not driven by a change in the rate of the competing risk.

Higher HCT-CI scores were associated with an insignificant increase in AGvHD incidence (HR_{SD} 2.04; CI 0.94 - 4.41; P 0.07). There was also a trend towards increased AGvHD with older donors (HR_{SD} 1.22; CI 0.94 - 1.59; P 0.14). Of the protein variables, there was a marked increase in AGvHD incidence with higher concentrations of CCL23 (HR_{SD} 1.51; CI 1.12 - 2.03; P 0.01) and LIF-R (HR_{SD} 1.32; CI 1.02 - 1.7; P 0.04). PH tests indicated no significant violation of the PH assumption.

CDCP1 concentration was associated with a significant increase in death (without AGvHD) (HR_{SD} 1.38; CI 1.03 - 1.86; P 0.03) and there was a marginal increase with elevated TNFRSF9 concentration (HR_{SD} 1.33; CI 0.96 - 1.84; P 0.09), however, neither protein impacted the incidence of AGvHD. There was evidence of LIF-R and CCL23 violated the PH assumption when regressed on death (without AGvHD) (P <0.01, P 0.01).

Table 4.17 Univariate effects on acute graft-vs-host disease – serum proteins. Subdistribution hazard ratios (HR_{SD}), with confidence interval (CI), significance test (P. val) and proportional hazards test (PH test). P < 0.05 and P < 0.1 are highlighted in green and blue, respectively.

	Acute graft-vs-host disease				Death (without AGvHD)			
Variable	**HR_{SD}**	**CI**	**P val.**	**PH test**	**HR_{SD}**	**CI**	**P val.**	**PH test**
Age at transplant / 10	1.28	(0.93 - 1.76)	0.12	0.72	1.02	(0.79 - 1.32)	0.86	0.77
CD34+ dose	1.06	(0.88 - 1.27)	0.53	0.47	0.99	(0.88 - 1.12)	0.92	0.73
D. age / 10	1.22	(0.94 - 1.59)	0.14	0.69	1	(0.68 - 1.47)	1.00	0.41
Diagnosis Lymphoid								
Diagnosis Myeloid	1.34	(0.66 - 2.71)	0.42	0.19	0.8	(0.39 - 1.66)	0.55	0.30
GvHD prophyl. Ciclosporin								
GvHD prophyl. Combination	1.06	(0.56 - 1.98)	0.87	0.40	1.54	(0.78 - 3.03)	0.21	0.06
P. sex Female								
P. sex Male	1.34	(0.7 - 2.56)	0.38	0.73	0.9	(0.44 - 1.83)	0.76	0.61
RIC No								
RIC Yes	0.99	(0.54 - 1.82)	0.98	0.80	0.73	(0.37 - 1.44)	0.36	0.07
D. sex Matched								
D. sex Mismatched	0.98	(0.54 - 1.79)	0.95	0.47	0.59	(0.29 - 1.23)	0.16	0.88
HCT-CI 0								
HCT-CI 1	1.23	(0.48 - 3.17)	0.66	0.22	0.96	(0.28 - 3.31)	0.95	0.35
HCT-CI 2	0.68	(0.23 - 1.99)	0.48	0.14	1.13	(0.47 - 2.71)	0.79	0.11
HCT-CI 3+	2.04	(0.94 - 4.41)	0.07	0.45	0.88	(0.31 - 2.49)	0.81	0.29
TBI No								
TBI Yes	1.23	(0.6 - 2.53)	0.57	0.51	1.12	(0.49 - 2.59)	0.78	0.72
ARTN	1.08	(0.87 - 1.35)	0.47	0.52	0.98	(0.77 - 1.25)	0.90	0.42
CASP-8	1.18	(0.85 - 1.65)	0.31	0.14	1.32	(1 - 1.75)	0.05	0.06
CCL20	1.22	(0.92 - 1.62)	0.17	0.10	1.08	(0.82 - 1.43)	0.57	0.28
CCL23	1.51	(1.12 - 2.03)	0.01	0.72	0.89	(0.63 - 1.25)	0.50	0.01
CDCP1	1.04	(0.79 - 1.38)	0.76	0.89	1.38	(1.03 - 1.86)	0.03	0.16
CSF-1	1.11	(0.8 - 1.55)	0.53	0.33	1.15	(0.81 - 1.63)	0.43	0.70
CXCL10	1.06	(0.75 - 1.5)	0.76	0.32	1.11	(0.84 - 1.46)	0.47	0.76
Flt3L	0.74	(0.54 - 1.02)	0.07	0.80	1.1	(0.81 - 1.49)	0.54	0.29
IL-12B	1.25	(0.9 - 1.73)	0.18	0.66	0.93	(0.67 - 1.27)	0.63	0.16
IL-17C	0.94	(0.67 - 1.3)	0.69	0.53	0.83	(0.53 - 1.31)	0.43	0.24
IL10	1.24	(1.05 - 1.46)	0.01	0.17	1.02	(0.79 - 1.33)	0.85	0.82
IL18	1.19	(0.92 - 1.54)	0.19	0.94	1.17	(0.78 - 1.76)	0.45	0.43
IL2	1.3	(0.85 - 1.97)	0.23	0.16	0.99	(0.71 - 1.36)	0.93	0.13
IL6	1.13	(0.8 - 1.59)	0.49	0.07	0.97	(0.7 - 1.36)	0.87	0.64
LIF-R	1.32	(1.02 - 1.7)	0.04	0.10	0.86	(0.58 - 1.27)	0.45	<0.01
MMP-10	0.8	(0.58 - 1.12)	0.20	0.54	1.02	(0.7 - 1.5)	0.91	0.39
NT-3	1.24	(1 - 1.54)	0.05	0.35	0.79	(0.55 - 1.14)	0.21	0.56
PD-L1	1.16	(0.88 - 1.53)	0.28	0.43	1.22	(0.86 - 1.73)	0.26	0.30
TNFRSF9	1.06	(0.83 - 1.36)	0.64	0.67	1.33	(0.96 - 1.84)	0.09	0.24

4.8.3 Multivariate effects on AGvHD

The impact of selected variables on the incidence of AGvHD was tested using a multivariate FG model (**Table 4.18**). The process was repeated for death (without AGvHD), the competing risk (**Table 4.18**). There was some evidence of time-dependence when CCL23 was regressed on death (without AGvHD) (P 0.01) but was included as the assumption was not violated in the AGvHD model (P 0.73), whereas LIF-R was excluded as the assumption appeared to be violated for both outcomes (P<0.1).

The concentration of CCL23 retained a significant impact on AGvHD incidence in the multivariate comparison (adj. HR_{SD} of 1.66, CI 1.12 - 2.46; P 0.01). Older subjects experienced a greater incidence of AGvHD with a 10-year increase in age having an adj. HR_{SD} of 1.46 (CI 1.03 – 2.08; P 0.04). There was an insignificant increase in AGvHD incidence with higher HCT-CI scores (adj. HR_{SD} 2.6; CI 0.97 - 6.95; P 0.06).

These findings suggest the incidence of AGvHD increased with the age of the patient at the time of transplant. CCL23 concentration may also have had a significant deleterious effect on the incidence of AGvHD.

Table 4.18 Multivariate effects on acute graft-vs-host disease – serum proteins. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with confidence interval (CI), significance test (P. val) and proportional hazards test (PH test). P < 0.05 and P < 0.1 are highlighted in green and blue, respectively.

	Acute graft-vs-host disease				Death (without AGvHD)			
Variable	**Adj. HR_{SD}**	**Adj.CI**	**P val.**	**PH test**	**Adj. HR_{SD}**	**Adj.CI**	**P val.**	**PH test**
Age at transplant / 10	1.46	(1.03 - 2.08)	0.04	0.64	0.93	(0.68 - 1.29)	0.68	0.65
CD34+ dose	1.06	(0.88 - 1.27)	0.55	0.60	1	(0.88 - 1.14)	0.94	0.73
Diagnosis Lymphoid								
Diagnosis Myeloid	0.9	(0.37 - 2.21)	0.81	0.19	0.92	(0.39 - 2.2)	0.86	0.39
HCT-CI 0								
HCT-CI 1	1.22	(0.37 - 3.99)	0.74	0.19	0.75	(0.16 - 3.55)	0.72	0.35
HCT-CI 2	0.83	(0.25 - 2.71)	0.75	0.20	1.08	(0.37 - 3.18)	0.88	0.22
HCT-CI 3+	2.6	(0.97 - 6.95)	0.06	0.38	0.91	(0.24 - 3.37)	0.89	0.10
P. sex Female								
P. sex Male	1.08	(0.53 - 2.17)	0.84	0.91	1.01	(0.41 - 2.5)	0.98	0.59
TBI No								
TBI Yes	1.48	(0.66 - 3.34)	0.34	0.65	1.89	(0.54 - 6.6)	0.32	0.62
CCL23	1.66	(1.12 - 2.46)	0.01	0.74	0.65	(0.35 - 1.2)	0.17	0.03
CDCP1	0.93	(0.64 - 1.35)	0.7	0.96	1.37	(0.98 - 1.92)	0.07	0.23
CSF-1	0.95	(0.6 - 1.51)	0.83	0.42	1.13	(0.64 - 1.97)	0.68	0.77
CXCL10	0.89	(0.59 - 1.35)	0.6	0.45	1.04	(0.65 - 1.64)	0.88	0.65
IL6	1.13	(0.77 - 1.65)	0.54	0.07	0.89	(0.62 - 1.29)	0.55	0.73
PD-L1	0.87	(0.5 - 1.52)	0.62	0.39	1.19	(0.72 - 1.96)	0.51	0.50
TNFRSF9	0.99	(0.69 - 1.42)	0.95	0.43	1.36	(0.91 - 2.04)	0.13	0.20

4.9 Chronic graft-vs-host disease

4.9.1 Cumulative incidence of CGvHD

The $CIF(t)$ of CGvHD and death (without CGvHD), were estimated as competing risks (**Figure 4.8**). At one-year post SCT, the CGvHD incidence was 0.17 (CI 0.10 - 0.24) which rose to 0.2 (CI 0.13 - 0.28) after three years. In total, 28 subjects experienced CGvHD during the follow-up.

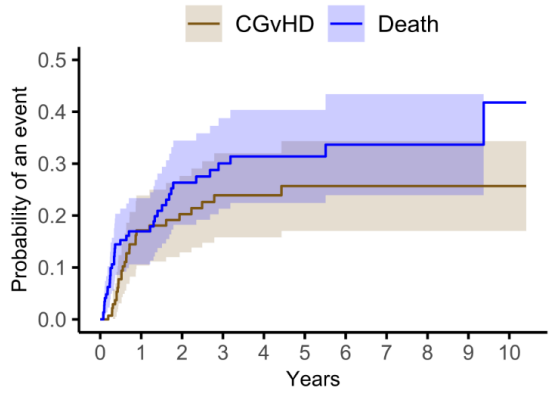


Figure 4.8 Cumulative incidence of chronic graft-vs-host disease. Cumulative incidence estimates with a 95% confidence interval. Death (without CGvHD) is the competing risk.

4.9.2 Univariate effects on CGvHD

The subdistribution hazard $\lambda_k^{SD}(t)$ of CGvHD & death (without CGvHD) were regressed on clinical and protein variables with the FG model (**Table 4.19**). In this analysis, CGvHD is the outcome of interest and the estimates for death (without CGvHD) are shown for comparative purposes.

None of the protein variables tested had a univariate association with CGvHD incidence (P > 0.05). There was an insignificant trend towards lower CGvHD incidence with increasing concentrations of Flt3L (HR_{SD} 1.33; CI 0.93 - 1.88; P 0.12) and a trend in the opposite direction for IL10 (HR_{SD} 0.77; CI 0.55 - 1.06; P 0.11). The PH test indicated IL-12B violated the PH assumption for CGvHD (P 0.05).

There was an apparent decrease in the incidence of CGvHD when male subjects were compared to female subjects (HR_{SD} 0.47; CI (0.23 - 0.97); P 0.04) with a corresponding increase in the incidence of death (without CGvHD) (HR_{SD} 2.01; CI 0.94 - 4.33; P 0.07). Moreover, total body irradiation was associated with a trend towards increased CGvHD incidence (HR_{SD} 1.91; CI 0.86 - 4.22; P 0.11). However, there was evidence both patient sex and donor sex mismatch violated the PH assumption for death (without CGvHD) (P 0.04, P 0.08).

Due to the absence of univariate associations with CGvHD incidence, a multivariate model was not interpreted.

Table 4.19 Univariate effects on chronic graft-vs-host disease – serum proteins. Subdistribution hazard ratios (HR_{SD}), with confidence interval (CI), significance test (P. val) and proportional hazards test (PH test). P < 0.05 and P < 0.1 are highlighted in green and blue, respectively.

	Chronic graft-vs-host disease				Death (without CGvHD)			
Variable	**HR_{SD}**	**CI**	**P val.**	**PH test**	**HR_{SD}**	**CI**	**P val.**	**PH test**
Age at transplant / 10	0.92	(0.73 - 1.17)	0.50	0.49	1.21	(0.89 - 1.64)	0.22	0.77
CD34+ dose	1.08	(0.89 - 1.32)	0.45	0.76	1.07	(0.95 - 1.2)	0.28	0.09
D. age / 10	0.97	(0.7 - 1.34)	0.84	0.41	1.05	(0.78 - 1.41)	0.76	0.41
Diagnosis Lymphoid								
Diagnosis Myeloid	0.83	(0.39 - 1.79)	0.64	0.25	1.14	(0.56 - 2.3)	0.72	0.24
GvHD prophyl. Ciclosporin								
GvHD prophyl. Combination	1.3	(0.62 - 2.73)	0.49	0.59	1.14	(0.6 - 2.18)	0.69	0.52
P. sex Female								
P. sex Male	0.47	(0.23 - 0.97)	0.04	0.88	2.01	(0.94 - 4.33)	0.07	0.04
RIC No								
RIC Yes	0.87	(0.42 - 1.82)	0.72	0.51	0.6	(0.32 - 1.13)	0.11	0.61
D. sex Matched								
D. sex Mismatched	1.49	(0.72 - 3.1)	0.28	0.62	0.48	(0.24 - 0.98)	0.05	0.08
HCT-CI 0								
HCT-CI 1	0.91	(0.23 - 3.68)	0.90	0.40	1.5	(0.49 - 4.62)	0.49	0.39
HCT-CI 2	0.76	(0.27 - 2.17)	0.61	0.50	1.03	(0.36 - 2.95)	0.96	0.24
HCT-CI 3+	0.71	(0.24 - 2.13)	0.55	0.51	1.98	(0.79 - 4.94)	0.15	0.44
TBI No								
TBI Yes	1.91	(0.86 - 4.22)	0.11	0.69	0.96	(0.43 - 2.14)	0.92	0.12
ARTN	0.96	(0.73 - 1.25)	0.75	0.15	1.01	(0.8 - 1.27)	0.96	0.59
CASP-8	1.15	(0.81 - 1.63)	0.45	0.42	1.12	(0.85 - 1.48)	0.40	0.58
CCL20	1.01	(0.75 - 1.37)	0.93	0.12	1.27	(0.98 - 1.64)	0.07	0.22
CCL23	1.11	(0.81 - 1.53)	0.50	0.62	1.36	(0.92 - 2.02)	0.13	0.71
CDCP1	0.82	(0.54 - 1.24)	0.35	0.15	1.39	(1.05 - 1.85)	0.02	0.90
CSF-1	0.81	(0.59 - 1.11)	0.20	0.17	1.53	(1.13 - 2.07)	0.01	0.41
CXCL10	0.91	(0.62 - 1.33)	0.62	0.13	1.24	(0.92 - 1.68)	0.17	0.65
Flt3L	1.33	(0.93 - 1.88)	0.12	0.68	0.84	(0.62 - 1.13)	0.25	0.13
IL-12B	0.9	(0.61 - 1.32)	0.58	0.05	1.21	(0.91 - 1.61)	0.20	0.07
IL-17C	0.98	(0.73 - 1.32)	0.91	0.51	0.82	(0.55 - 1.22)	0.32	0.15
IL10	0.77	(0.55 - 1.06)	0.11	0.46	1.23	(1.02 - 1.49)	0.03	0.42
IL18	0.83	(0.55 - 1.26)	0.38	0.36	1.43	(1.02 - 2.02)	0.04	0.50
IL2	0.89	(0.66 - 1.2)	0.45	0.20	1.28	(0.81 - 2.01)	0.29	0.91
IL6	0.96	(0.71 - 1.3)	0.81	0.40	1.13	(0.85 - 1.5)	0.41	0.88
LIF-R	1.09	(0.66 - 1.82)	0.73	0.86	1.11	(0.81 - 1.53)	0.51	0.04
MMP-10	1.1	(0.76 - 1.59)	0.63	0.32	0.86	(0.6 - 1.23)	0.41	0.22
NT-3	1.15	(0.8 - 1.65)	0.46	0.62	1.01	(0.78 - 1.3)	0.95	0.25
PD-L1	1.04	(0.69 - 1.55)	0.85	0.10	1.44	(1.07 - 1.96)	0.02	0.52
TNFRSF9	0.87	(0.61 - 1.24)	0.43	0.16	1.32	(0.92 - 1.89)	0.14	0.28

4.10 Relapse incidence

4.10.1 Cumulative incidence of relapse

The $CIF(t)$ of relapse was estimated with NRM as the competing risk (**Figure 4.9**). At one-year post-transplant, the relapse incidence was 0.19 (CI 0.12-0.26) which rose to 0.29 (CI 0.21-0.37) after three years. In total, 43 patients relapsed during the follow-up.

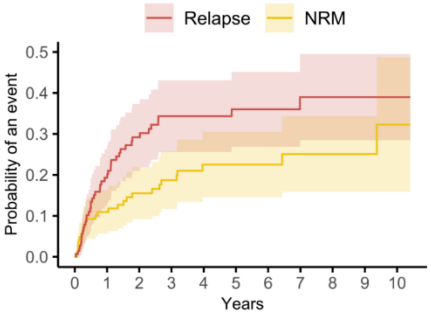


Figure 4.9 Cumulative incidence of relapse. Cumulative incidence estimates with a 95% confidence interval. Non-relapse mortality (NRM) is the competing risk.

4.10.2 Univariate effects on relapse

The subdistribution hazard $\lambda_k^{SD}(t)$ of relapse was regressed on clinical and protein variables with the FG model (**Table 4.20**). In this analysis, relapse is the outcome of interest. Effects on the competing risk, NRM, are shown for comparative purposes, but the estimates reported in **Table 4.10** will be interpreted.

Of the clinical variables, having a myeloid diagnosis was associated (P 0.04) with an increase in relapse incidence. Increased concentrations of IL2 were associated with a decrease in the incidence of relapse (P 0.01) and those of CXCL10 and IL6 with a marginal decrease (P 0.06; P 0.08). PH tests indicated GvHD prophylaxis and IL2 violated the PH assumption (P <0.05).

When relapse was modelled as the competing event, the HR_{SD} estimates for NRM (**Table 4.20**) were consistent with those reported in **Table 4.10**. However, in this comparison, the effect of IL12-B was less certain (HR_{SD} 1.2; CI 0.86 - 1.66) and IL10 was associated with an increase in NRM (HR_{SD} 1.3; CI 1.08 - 1.57).

Table 4.20 Univariate effects on relapse – serum proteins. Subdistribution hazard ratios (HR_{SD}), with confidence interval (CI), significance test (P. val) and proportional hazards test (PH test). P < 0.05 and P < 0.1 are highlighted in green and blue, respectively.

	Relapse				Non-relapse mortality			
Variable	**HR_{SD}**	**CI**	**P val.**	**PH test**	**HR_{SD}**	**CI**	**P val.**	**PH test**
Age at transplant / 10	1.12	(0.88 - 1.43)	0.36	0.532	1.05	(0.76 - 1.44)	0.78	0.238
CD34+ dose	1.07	(0.96 - 1.2)	0.22	0.404	1.08	(0.93 - 1.26)	0.30	0.146
D. age / 10	0.93	(0.73 - 1.19)	0.57	0.505	1.02	(0.77 - 1.35)	0.88	0.452
Diagnosis Lymphoid								
Diagnosis Myeloid	2.07	(1.03 - 4.17)	0.04	0.075	0.48	(0.23 - 1)	0.05	0.708
GvHD prophyl. Ciclosporin								
GvHD prophyl. Combination	0.74	(0.37 - 1.47)	0.40	0.007	1.25	(0.6 - 2.63)	0.55	0.307
P. sex Female								
P. sex Male	0.97	(0.53 - 1.8)	0.93	0.818	0.89	(0.43 - 1.88)	0.77	0.349

RIC No								
RIC Yes	0.82	(0.45 - 1.48)	0.50	0.164	0.77	(0.37 - 1.59)	0.47	0.55
D. sex Matched								
D. sex Mismatched	1.15	(0.64 - 2.09)	0.64	0.695	0.4	(0.17 - 0.94)	0.04	0.964
HCT-CI 0								
HCT-CI 1	1.21	(0.39 - 3.72)	0.75	0.387	1.2	(0.3 - 4.84)	0.80	0.687
HCT-CI 2	1.3	(0.59 - 2.85)	0.52	0.522	1.13	(0.38 - 3.39)	0.82	0.563
HCT-CI 3+	0.53	(0.19 - 1.53)	0.24	0.4	2.7	(0.9 - 8.11)	0.09	0.45
TBI No								
TBI Yes	0.56	(0.21 - 1.48)	0.24	0.064	1.2	(0.51 - 2.84)	0.67	0.676
ARTN	1.02	(0.77 - 1.33)	0.91	0.285	1.01	(0.8 - 1.28)	0.91	0.592
CASP.8	0.91	(0.68 - 1.21)	0.51	0.262	1.18	(0.84 - 1.65)	0.34	0.549
CCL20	0.86	(0.59 - 1.27)	0.46	0.738	1.25	(0.93 - 1.66)	0.13	0.515
CCL23	1.18	(0.86 - 1.61)	0.31	0.871	1.19	(0.78 - 1.81)	0.41	0.645
CDCP1	1.15	(0.86 - 1.55)	0.35	0.838	1.11	(0.82 - 1.49)	0.51	0.362
CSF.1	0.97	(0.73 - 1.28)	0.83	0.242	1.58	(1.1 - 2.28)	0.01	0.114
CXCL10	0.78	(0.6 - 1.01)	0.06	0.556	1.44	(1.03 - 2.01)	0.04	0.775
Flt3L	0.84	(0.63 - 1.11)	0.22	0.178	1.01	(0.7 - 1.45)	0.97	0.349
IL.12B	0.99	(0.73 - 1.35)	0.96	0.286	1.2	(0.86 - 1.66)	0.28	0.159
IL.17C	0.76	(0.53 - 1.1)	0.14	0.44	1.12	(0.78 - 1.61)	0.53	0.177
IL10	0.78	(0.58 - 1.05)	0.10	0.185	1.3	(1.08 - 1.57)	0.01	0.347
IL18	1.04	(0.77 - 1.42)	0.78	0.774	1.32	(0.89 - 1.95)	0.17	0.116
IL2	0.63	(0.43 - 0.91)	0.01	0.03	1.5	(1 - 2.25)	0.05	0.415
IL6	0.73	(0.52 - 1.03)	0.08	0.426	1.37	(1.03 - 1.82)	0.03	0.312
LIF.R	1.07	(0.8 - 1.44)	0.64	0.785	0.93	(0.65 - 1.35)	0.71	0.338
MMP.10	0.82	(0.6 - 1.11)	0.19	0.22	1.25	(0.85 - 1.83)	0.25	0.493
NT.3	0.83	(0.6 - 1.14)	0.25	0.507	1.03	(0.76 - 1.38)	0.86	0.154
PD.L1	1.02	(0.75 - 1.39)	0.90	0.546	1.41	(0.99 - 2.01)	0.06	0.082
TNFRSF9	1.1	(0.8 - 1.53)	0.55	0.081	1.1	(0.71 - 1.72)	0.66	0.086

4.10.3 Multivariate effects on relapse

The impact of selected variables on the incidence of relapse was tested using a multivariate FG model (**Table 4.21**). Effects on the competing risk, NRM, are shown for comparative purposes, but the multivariate estimates reported in **Table 4.15** will be interpreted. CSF-1 was excluded for violation of the PH assumption when regressed on NRM in this comparison (P 0.04).

The concentration of CXCL10 was associated with a significant decrease in relapse incidence in the multivariate comparison (adj. HR_{SD} 0.66; CI 0.48 - 0.9; P 0.01). There was also an insignificant trend towards lower relapse incidence with IL6 concentration (adj. HR_{SD} 0.72; CI 0.5 - 1.04; P 0.08). Of the clinical variables, patients with a Myeloid diagnosis had significantly higher rates of relapse (adj. HR_{SD} 2.3; CI 1 - 5.26; P 0.05) compared to those with a Lymphoid diagnosis.

This data indicates higher concentrations of CXCL10 were linked with lower relapse incidence but with a corresponding increase in NRM. The models reiterate that patients with a myeloid diagnosis experienced less NRM but also a higher probability of relapse. This comparison also demonstrates modelling relapse as the competing risk for NRM, as opposed to relapse-related mortality, did not substantively change the estimated coefficients compared to those in **Table 4.15**.

Table 4.21 Multivariate effects on relapse – serum proteins. Multivariate subdistribution hazard ratios (Adj. HR_{SD}), with confidence interval (CI), significance test (P. val) and proportional hazards test (PH test). P < 0.05 and P < 0.1 are highlighted in green and blue, respectively.

	Relapse				Non-relapse mortality			
Variable	**Adj. HR_{SD}**	**Adj.CI**	**P val.**	**PH test**	**Adj. HR_{SD}**	**Adj.CI**	**P val.**	**PH test**
Age at transplant / 10	1.04	(0.77 - 1.4)	0.80	0.55	1.13	(0.7 - 1.81)	0.62	0.41
CD34+ dose	1.07	(0.93 - 1.23)	0.35	0.34	1.21	(1 - 1.46)	0.05	0.31
Diagnosis Lymphoid								
Diagnosis Myeloid	2.3	(1 - 5.26)	0.05	0.05	0.24	(0.08 - 0.7)	0.01	0.78
HCT-CI 0								
HCT-CI 1	1.43	(0.36 - 5.57)	0.61	0.41	1	(0.12 - 8.31)	1.00	0.59
HCT-CI 2	1.63	(0.7 - 3.79)	0.26	0.53	1.24	(0.3 - 5.12)	0.77	0.34
HCT-CI 3+	0.55	(0.18 - 1.67)	0.29	0.35	4.62	(1.01 - 21.26)	0.06	0.18
P. sex Female								
P. sex Male	0.78	(0.38 - 1.61)	0.50	0.52	0.95	(0.35 - 2.6)	0.93	0.59
TBI No								
TBI Yes	0.67	(0.21 - 2.16)	0.50	0.11	1.38	(0.42 - 4.55)	0.60	0.48
CCL23	1.42	(0.87 - 2.31)	0.16	0.76	1.16	(0.65 - 2.08)	0.61	0.93
CDCP1	1.29	(0.91 - 1.84)	0.16	0.88	1.26	(0.82 - 1.95)	0.29	0.56
CXCL10	0.66	(0.48 - 0.9)	0.01	0.47	1.49	(0.89 - 2.5)	0.13	0.97
IL6	0.72	(0.5 - 1.04)	0.08	0.37	1.54	(1.04 - 2.28)	0.03	0.47
LIF-R	1.04	(0.73 - 1.48)	0.82	0.83	0.48	(0.28 - 0.82)	0.01	0.16
PD-L1	0.9	(0.49 - 1.65)	0.72	0.67	1.44	(0.79 - 2.63)	0.23	0.07
TNFRSF9	1.15	(0.6 - 2.23)	0.67	0.10	0.86	(0.49 - 1.51)	0.61	0.37

4.11 Absolute protein quantification

The Olink immunoassay provides a relative measure of serum concentration between individuals in this cohort, measured in arbitrary units, NPX. For a subset of the candidate biomarkers, Kriti Verma conducted an enzyme-linked immunosorbent assay (ELISA) assay to quantify absolute protein concentration in the original samples. The absolute concentration of CCL23, CDCP1, PD-L1, and IL-6 showed a significant positive correlation with the OLINK NPX data (**Figure 4.10**). However, LIF-R showed no clear relationship. This raises the possibility of an error in the relative quantification of this analyte by the OLINK immunoassay. The correlation of age at transplant and CD34+ dose with the OLINK NPX data was also tested (**Table 4.22**). This revealed CCL23 and PD-L1 both had a weak ($\rho < 0.5$) positive correlation with age, whereas LIF-R had a weak negative relationship. No relationship with CD34+ dose was detected, except for LIF-R, which had a weak negative correlation.

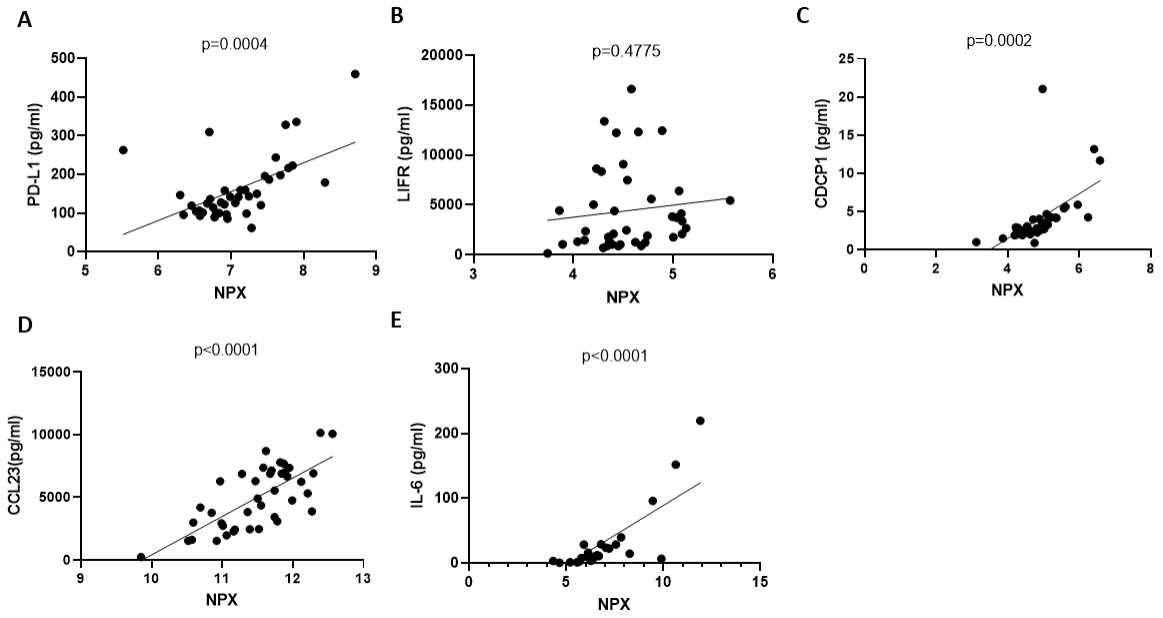


Figure 4.10 Correlation of relatively quantified Olink Normalised protein expression (NPX) data with absolute quantification of protein expression by ELISA in pg/ml for selected proteins. Figure by Kriti Verma. (A) PD-L1 (B) LIFR (C) CDCP1 (D) CCL23 (E) IL-6.

Table 4.22 Correlation of relatively quantified Olink Normalised protein expression (NPX) data with age and CD34+ dose for selected proteins. Spearman's correlation (ρ).

Variable	Protein	Spearman's ρ	P. value
Age	CCL23	0.33	< 0.01
Age	CDCP1	0.09	0.27
Age	PD-L1	0.27	< 0.01
Age	IL6	0.08	0.37
Age	LIF-R	-0.27	< 0.01
CD34+ dose	CCL23	-0.07	0.43
CD34+ dose	CDCP1	-0.12	0.16
CD34+ dose	PD-L1	0.08	0.34
CD34+ dose	IL6	0.03	0.71
CD34+ dose	LIF-R	-0.19	0.02

4.12 Summary

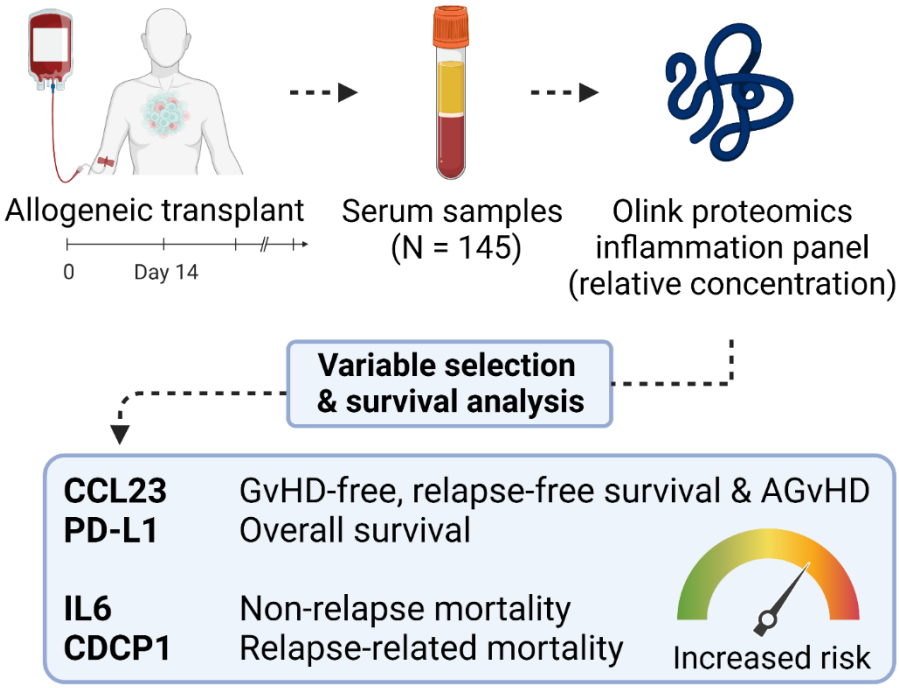


Figure 4.11 Graphical summary.

The estimated OS probability at three years was 62% (CI 0.62 - 0.79) for this cohort. However, the probability of surviving without GvHD or relapse, GRFS, was substantially lower at 28% (CI 0.21-0.37). Indicating relatively few individuals survived without ongoing morbidity.

Within five months of transplantation, 29% (CI 0.21 - 0.37) of individuals were diagnosed with AGvHD (grade II-IV). Whereas for CGvHD, the cumulative incidence was 20% (CI 0.21-0.38) after three years. The cumulative incidence of relapse-related mortality, RM, at 18% (CI 0.10 - 0.25) was similar to that of NRM (20%; CI 0.12 - 0.28), at the same time point. Although, the probability of relapse was higher at 29% (CI 0.21-0.37).

A subset of the candidate biomarkers was shown to be predictive of patient prognosis when measured on the 14th day following transplantation. Both univariate and multivariate analysis suggested high concentrations of PD-L1, CCL23, and CDCP1 expression were associated with lower rates of OS, GRFS, and increased RM incidence, respectively ($P < 0.05$).

Moreover, IL-6 concentration was associated with a lower incidence of RM but an increase in the competing risk NRM ($P < 0.05$). CXCL10 was also associated with lower relapse incidence in the multivariate comparison ($P < 0.05$), however, this did not achieve significance in the univariate test (P 0.06).

Relatively quantified Olink normalised protein expression (NPX) data correlated with absolute quantification of protein expression by ELISA in pg/ml for PD-L1, CDCP1, CCL23, and IL-6. However, there was no relationship between the Olink NPX data and absolute protein concentration for LIF-R.

5 Discussion & Conclusions

This thesis investigated factors influencing prognosis following allo-HSCT in an exploratory data analysis. The studies focused on the cellular composition of the graft pre-transplant, and the expression of serum proteins 14 days post-transplant.

Specifically, the four main aims addressed were to:

(1) Describe the cellular composition of PBSC grafts prior to transplantation using CyTOF data.
(2) Determine if distinct types of PBSC graft can be defined based on variation in cellular composition.
(3) Investigate associations between PBSC graft composition and recipient prognosis.
(4) Investigate association between serum proteins post-transplant and recipient prognosis.

In **Chapter 3**, unsupervised cluster analysis was used to identify cell populations in single-cell CyTOF data generated from grafts, and to define distinct types of graft composition. In **Chapter 4**, supervised machine learning indicated which subset of serum proteins were predictive of patient outcomes. In both projects, survival and competing risks regression were used to test the influence this data had on the on the prognosis of allo-HSCT recipients.

The clinical endpoints studied were overall survival (OS), GvHD-free, relapse-free survival (GRFS), relapse-related mortality (RM) and transplant-related mortality (TRM) (referred to as non-relapse mortality (NRM) when the cause of death was not adjudicated). The onset of acute graft-vs-host disease (AGvHD), chronic graft-vs-host disease (CGvHD) and relapse were also assessed.

This chapter begins with a summary of the key findings (**Section 5.1**), then an interpretation of the results of the allo-HSCT graft composition study (**Sections 5.2-5.4**) and the post-transplant serum protein study (**Section 5.5**). Next, the potential limitations of the cohorts, data and methodology are described (**Section 5.6**). This is followed by a discussion of future directions for research to expand on the findings reported in this thesis and to address the limitations (**Section 5.7**). Lastly, the main conclusions are described (**Section 5.8**).

5.1 Key findings

The data indicated HSCs represent around 2% (1 - 4%) of the transplanted cellular content. The remaining 98% of cells within grafts consisted mostly of monocytes, T cells and B cells. Whereas NK and DCs were less prevalent.

Three distinct groups of grafts were identified based on differences in cellular composition. These were independent predictors of patient prognosis. GRFS two years post-transplant ranged from 42% to 11% between graft groups.

I found evidence of significant association between the whole graft composition and the rates of GRFS (0.007) and OS (P 0.003), and incidence of TRM (P 0.02), beyond that attributable to other variables.

Concerning individual cell lineages, graft T-cell content was a risk factor for CGvHD incidence. In contrast, grafts with a greater proportion of monocytes and DNAM-1^{+} lymphocyte progenitor cells were associated with improvements in GRFS and TRM, respectively.

The data confirmed that a previously validated prognostic biomarker, IL6, is predictive of non-relapse mortality. The analysis also suggested that high-levels of PD-L1, CCL23 and CDCP1 expression in the serum at Day-14 could be prognostic biomarkers for OS, GRFS and RM, respectively.

5.2 Allo-HSCT graft composition

Relatively few studies have systematically described the composition of allo-HSCT grafts (404-409). Only 1% of cells in a peripheral blood stem cell graft are reported to be stem cells (409). Analysis in **Chapter 3** found the figure to be closer to 2% (1 - 4%) across a cohort of 185 grafts. This illustrates that the defining functional constituent, HSC, represent only a fraction of the transplanted cellular content.

5.2.1 Lymphocyte content of grafts aligns with estimates from previous studies

Cell populations were identified in the CyTOF data with an unsupervised clustering algorithm, PhenoGraph (245). Populations were then characterised into immune cell types based on the expression of cell surface markers. There was broad correspondence between the abundance of cell populations in **Chapter 3** and estimates reported in previous studies.

Both major and rarer T cell subsets were identified automatically. CD4+ and CD8+ T cell subsets were identified at a ratio of 1.6:1 across the grafts. Typically, these cells are found at a ratio between 1.7:1 and 2:1 in the peripheral blood of healthy individuals (409-411). Around 5.2% of T cells in grafts were γδ TCR+. Previous studies report γδ T cells to account for between 2.6 and 10% of T cells in PBSC grafts (412, 413). PhenoGraph was also able to identify subgroups within the main CD4/CD8 compartments. T-regs comprised 3.06% of the graft CD4+ T cell content. This corresponds closely with a previous study of 94 PBSC grafts where T-regs accounted for 2.96% of CD4+ T cells (362).

T cells are instrumental in the parthenogenesis of GvHD and are the primary mediators of the GvL effect, along with NK cells (40-42). As a proportion of lymphocytes, NK cells constituted 6% of cells, in line with reference values for the peripheral blood (366, 414). Typically, two NK cell populations are identified based on the expression levels of CD56 (i.e., $CD56^{bright}$ and $CD56^{dim}$) (382). However, these were not partitioned by the unsupervised clustering algorithm.

B cells were the other major lymphocyte component of grafts (15%, IQR 10 - 23%). B cells were more abundant than previous reports of PBSC grafts (404). The majority of B cells had a naïve phenotype (71.2%), as is the case in the peripheral blood, but this decreases with age (415).

5.2.2 Monocytes were relatively abundant

Monocytes are likely involved in immune reconstitution and regulation of immune responses following allo-HSCT (92). In **Chapter 3**, monocytes contributed around double the proportion of graft cellular content than has been previously reported (404). G-CSF mobilisation has been shown to markedly increase peripheral blood monocyte content (416).

PhenoGraph did not distinguish monocytes based on the prototypical monocyte subsets (classical, intermediate, and non-classical). Instead, $CD14^{+}$ clusters appeared to be divided based on continuous differences in marker expression. The majority (54.2%) had a classical-like phenotype, in line with estimates for peripheral blood (375).

pDC play a role in suppression of immune responses and have been linked with GvHD protection (417-419). Reports suggest pDCs are the most numerous DC subset in the peripheral blood (380). In line with this, pDCs were the most numerous singular DC subset in the grafts, comprising 33.7% of the DC compartment.

As well as the well-defined immune cell types, PhenoGraph identified a population of $DNAM\text{-}1^{+}$ lymphoid progenitor cells that contributed 1% (IQR 1 - 4%) of graft cellular content. Similar $DNAM\text{-}1^{+}$ lymphoid progenitors typically reside in the bone marrow but are increased in the circulation of individuals with chronic inflammation (367). Notably, these are also recruited to the peripheral blood

by G-SCF mobilisation (368). Studies indicate these cells rapidly differentiate into cytotoxic NK cells and functional T cells and may help to quickly generate effector lymphocytes in anti-viral responses (420-422).

Together these findings suggest that the immune cell abundance estimates based on the CyTOF data were realistic relative to previously reported estimates for PBSC grafts. It also indicates the PhenoGraph algorithm was capable of delineating most major immune cell types, in addition to identifying unexpected cell populations.

Considerable variation in composition was observed between different grafts. For example, whilst the median proportion of T cells was 28% the interquartile range was 17 - 39%. The total dose of infused mononuclear cells was highly heterogenous (MNC dose 0.650 - 11.600 x10^8 cells / kg). Moreover, there was a 90-fold difference in measured T cell concentrations (4.91 - 444.74 x 10^6 cells / kg).

5.3 Distinct types of allo-HSCT graft by composition

The second aim of **Chapter 3** was to test if distinct groups of grafts could be defined based on differences in graft composition between donations. Three distinct types of graft (PBSC graft groups) were identified within the graft cohort in using cluster analysis for compositional data (clustCoDa) (272). This focused on a subset of 155 grafts (153 PBSC; 2 BM) for which recipient outcome data was available.

The groups were identified based solely on differences in graft cellular composition but were found to stratify patients by risk. GRFS two years post-transplant ranged from 42% to 11% between graft groups (P 0.005).

5.3.1 Group A grafts had the best survival rates

Based on survival and cumulative incidence estimates, recipients of grafts from Group A had the most favourable rate of GRFS, OS, and the lowest TRM incidence. Meaning the recipients were not just more likely to survive, they were more likely to survive free of ongoing causes of morbidity and had a lower probability of dying from transplant-related complications.

Group A was characterised by a relative abundance of HSC, monocytes and pDCs, and had the lowest proportion of T cells, NK cells, and activated DCs (**Figure 3.9**, **Figure 5.1**). These grafts were donated by the youngest donors (median 34; IQR 24 – 48), the least of which were CMV positive (43%).

Younger donors are known to confer better rates of OS (156-158). In contrast, CMV-positive donors are associated with worse OS for recipients that are CMV-negative (168) (**Table 5.1**).

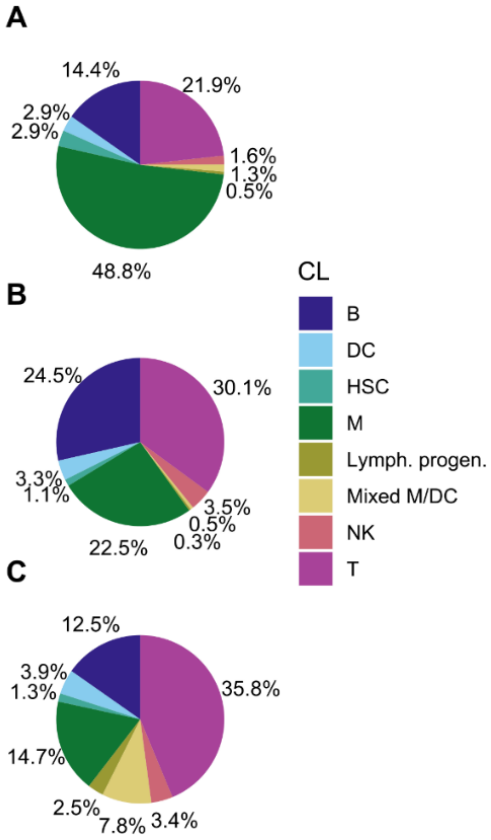


Figure 5.1 Median composition of PBSC graft groups. Proportion of cells by cell lineage (CL).

In both univariate and multivariate comparisons, Group C had an increased hazard of GRFS (adj. HR 2.20; <0.01) and OS (adj. HR 2.1; P 0.04). Moreover, there was a decrease in OS in Group B (adj. HR 2.43; P 0.01). This suggests that after accounting for donor & patient characteristics, PBSC graft group had a significant impact on patient prognosis.

Table 5.1 Defining characteristics of each PBSC graft group.

PBSC graft group	Donors	Prevalent cell types	Scarce cell types	GRFS 2-years
Group A	Youngest, mostly CMV neg.	HSC, classical-like monocytes, pDCs	T cells, NK cells, activated DCs	42%
Group B	Oldest, more female	B cells (memory; CCR7+ naïve), T cells (CD4 Naïve; DP)	HSC, classical-like monocytes,	29%
Group C	Variable age, mostly CMV pos., mostly unrelated	T cells (CD4 EM; γδ; DN), DNAM-1^{+} lymph. progenitors, PD-L2+ monocytes	HSC, classical-like monocytes	11%

5.3.2 Group C grafts had the highest GvHD incidence

In contrast to Group A, Group C consisted of the lowest proportion of monocytes, relatively few HSC, and had the highest proportion of T cells. Two populations of PD-L2+ monocytes and lymph. progenitors were also over-represented.

HSC were relatively scarce in Group C. Higher doses of CD34+ cells can improve engraftment following allo-HSCT (125, 126). Except for very high doses (125, 135, 409), higher doses are generally associated with improved outcomes (126, 130, 131) but the effect on GvHD is less clear (126-129).

Overall, T cells are the primary mediators of GvHD following allo-HSCT (136, 423). One might expect the group with the highest proportion of T cells to have had the highest GvHD incidence. Indeed, the cumulative incidence of GvHD in Group C was 59%, compared with around ~38% in the other two groups.

Notably, no significant differences in the incidence of relapse were observed between PBSC graft groups. Of the three, Group C had the highest incidence with 28% compared with <19% in the other groups.

Whilst there was an elevated probability of GvHD in Group C, there was no significant difference in AGvHD or CGvHD after adjusting for these donor characteristics. Whereas having an unrelated donor retained a significant effect on AGvHD (adj. HR_{SD} 3.35; P 0.02).

Group C grafts were mostly from CMV-positive (62%) unrelated donors (71%), and relatively few were female (23%). The elevated GvHD risk in Group C may therefore relate to the abundance of unrelated donors. Equally, these donor characteristics could have influenced graft composition making it difficult to speculate on causal association. However, the composition of Group C grafts aligns with the current understanding that a high T cell content promotes GvHD, and HSC content is broadly protective.

Whilst most of the major T cell subsets were up in this group, some rare T cell subsets, such as γδ T cells, and double negative T cells were also relatively abundant and may have protective effects. The concentration of γδ T cells in grafts has been linked with increased AGvHD (412, 413). However, high levels in the peripheral blood after allo-HSCT is protective against relapse and not associated with GvHD(424). Whereas double negative T cells are thought to suppress GvHD by inhibiting other lymphocytes (425-427). Repeating the comparison as a proportion of the T cell compartment could provide additional insight.

Group C grafts had the lowest monocyte content. Rather than influencing prognosis, the monocyte content may simply have been highly variable between groups. However, several previous studies indicate that this might not be the case.

A low concentration of monocytes has been linked to increased CGvHD (404). A lack of monocyte-dependant T cell inhibition could also contribute to the relatively high rate of GvHD (428, 429). Rapid monocyte recovery following allo-HSCT has also been linked with improved OS (430-432) and could influence graft function (433). For comparison, Group A, which had the highest proportion of monocytes, had the lowest rate of TRM at just 12% compared with >27% in the other groups.

These findings suggest Group C had an increased risk of GvHD & TRM relative to Group A, potentially related to an abundance of unrelated donors. These grafts were defined by a high concentration of T cells, and a relative scarcity of monocytes and HSC.

5.3.3 Group B grafts were associated with increased TRM

In both univariate and multivariate comparisons, Group B was associated with a higher incidence of TRM (adj. HR_{SD} 3.45; P 0.01) with little change in relapse incidence (adj. HR_{SD} 1.1; P 0.84). Indicating that the lower rate of survival was driven primarily by an increased risk of TRM, an effect that was still apparent after adjusting for donor and patient characteristics.

The composition of Group B grafts was characterised by an abundance of B cells and T cells, with a paucity of monocytes and HSC. In particular, memory B cells and CCR7+ Naïve B cells were over-represented. These grafts were donated by the oldest donors (median 41; IQR 29 - 51), a greater percentage of which were female (39%) and relatively few that were unrelated (51%).

As the GvHD incidence was not dissimilar from that of Group A at ~38%, the increase in TRM in Group B might have resulted from complications other than GvHD. However, the apparent absence of a GvHD effect could also be due to the relatively low number of events observed in each PBSC graft group.

The low levels of HSC in Group B grafts could augur worse immune reconstitution and influence susceptibility to opportunistic infections. Infections are also major causes of post-transplant non-relapse mortality that were not investigated and could have influenced the TRM incidence (52).

The B cell content of grafts has not been extensively studied in allo-HSCT (409). Small studies have indicated a potential link with graft B cell content and increased TRM and AGvHD (404, 434). Murine models suggest antigen presentation by B cells plays a role in GvHD parthenogenesis (435, 436).

Depletion of B cells from grafts has been found to influence the risk of transplant-related complications. The application of in vivo B cell depletion with rituximab early after transplant may be protective against AGvHD without impairing engraftment (437). Moreover, ex vivo depletion of T & B cells is known to effectively prevent PTLD (409).

Group B was also associated with a relative abundance of T cells and a scarcity of monocytes. Specifically, Naïve CD4+ T cells and DP T cells were more prevalent than in the other two graft groups. High naïve CD4+ T cell content (as a proportion of CD4+ T cells) has been linked with higher rates of AGvHD (438). Again, this suggests a targeted comparison within the T cell compartment could prove insightful.

Groups B and C had relatively few pDCs and a higher proportion of activated DCs (CD86+). A high proportion of pDCs has been linked with protection against GvHD in both PBSC (417) and BM grafts (405). pDCs are also capable of promoting T-reg expansion and suppression of antigen-specific

immune responses (418, 439). In contrast, DC activation status has been associated with increased AGvHD albeit with a different activation marker, CMRF-44 (440).

This indicates Group B had an increased risk of TRM relative to Group A, potentially related to an abundance of older and female donors. These grafts also had a high concentration of both B cells and T cells, and a relative scarcity of monocytes and HSC.

Compositional differences between these three groups may represent highly variable elements between grafts that unrelated to prognosis. However, a common pattern emerged where grafts with a high proportion of T cells, and low monocyte and HSC content were associated with lower rates of OS, GRFS, and worse TRM. The effects of which may be modulated by or as a result of differences in donor characteristics.

Notably, no significant differences in relapse incidence were observed between graft groups. A relatively low number of relapse events were observed in this cohort (31 relapse events), which could weaken any apparent association with the graft group. Alternatively, the clustering analysis may have failed to capture the compositional differences strongly associated with relapse. Or perhaps, compositional differences had little impact on the GvL effect in this cohort.

5.4 Allo-HSCT graft composition is associated with prognosis

The third goal of **Chapter 3** was to directly test the relationship between graft composition and recipient prognosis. Association between graft composition and prognosis was estimated by including parts of the composition as covariates in regression models (after log-ratio transformation).

The analysis was conducted on the same subset of 155 grafts (153 PBSC; 2 BM) in which the PBSC graft groups were identified. I found evidence of a significant association between the whole graft composition and the rate of GRFS (0.007), rate of OS (P 0.003), and incidence of TRM (P 0.02), beyond that attributable to other variables.

The nature of compositional data is that if one large part of the composition increases, others must decrease, and vice versa. If an increase in monocytes was associated with improved outcomes, this could also reflect a decrease in the proportion of other cell types with pathogenic effects. Results must therefore be interpreted from this relative perspective.

An increase in the proportion of T cells relative to HSCs was associated with significantly greater CGvHD incidence (adj. HR_{SD} 2.01; P 0.05). Given the established role T cells play in GvHD parthenogenesis, it's reasonable to suspect increased T cell concentrations in the graft were indicative of greater CGvHD risk. The link between donor T cells and GvHD has resulted in widespread exploration of T cell depletion as a form of GvHD-prophylaxis (441). This suggests other associations observed in **Chapter 3** could be informative for future clinical research.

5.4.1 The proportion of monocytes was associated with improved GRFS

In particular, a greater proportion of monocytes relative to HSCs was associated with significant improvements in the rate of GRFS (adj. HR 0.68; P 0.01). An insignificant increase in OS (adj. HR 0.73; P 0.06) and a decrease in relapse incidence (adj. HR_{SD} 0.6; P 0.09) were also observed with enhanced monocyte content.

A high concentration of monocytes has previously been linked with protection from CGvHD in PBSC grafts (404). Similarly, a decrease in the proportion of non-classical monocytes in BM grafts may covey lower AGvHD risk (442). Along with the role monocytes play in T cell inhibition (428) (429), it's possible

that graft monocyte content had a protective effect on the rate of GRFS. If confirmed, this would have important ramifications for research into graft manipulation as a cellular therapy.

The cytokine, G-CSF, administered prior to apheresis to mobilise donor HSC from the bone marrow, is known to influence circulating levels of myeloid cells at different stages of maturation (443). Moreover, G-CSF mobilises a regulatory set of monocytes to the peripheral blood (444). Together these indicate the potential for ex-vivo expansion of monocytic cells in the graft prior to infusion.

Levels of intermediate and non-classical monocytes >30 days post-transplant are associated with reduced CGvHD incidence (445). Another study associated the amount of classical monocytes early post-transplant with improved rates of survival, relapse incidence and TRM (430). Therefore, graft monocyte content should also be investigated for its potential to impact post-transplant monocyte reconstitution.

5.4.2 DNAM-1+ lymphocyte progenitors were associated with improved TRM

In **Chapter 3** there was evidence that an enhanced proportion of DNAM-1^{+} lymphoid progenitors could be linked with improved TRM. The proportion of DNAM-1^{+} lymphoid progenitors relative to HSCs had a protective effect on the incidence of TRM (adj. HR_{SD} 0.54; $P < 0.01$) and was associated with an insignificant improvement in OS (adj. HR 0.7; P 0.06).

The presence of similar DNAM-1^{+} lymphoid progenitors was only recently characterised in the serum of individuals with chronic infection and is understudied in the context of allo-HSCT (367). This population is thought to rapidly produce highly functional NK cells in response to inflammation which could be useful for immunotherapeutic purposes (368).

The impact of these cells on lymphocyte reconstitution and function post-transplant could therefore be informative for future research (446). Moreover, changes in DNAM-1+ lymphoid progenitor cell content had no effect on relapse incidence so this population should be studied for their potential to impede TRM without obstructing beneficial GvL effects.

A surprising result was that the proportion of NK cells relative to HSCs was associated with lower rates of OS (adj. HR 1.86; P <0.01) and GRFS (adj. HR 1.32; 0.05) as well as an increase in TRM incidence (adj. HR_{SD} 2.22; P <0.01). This was unexpected as particular NK cell subsets have been associated with protective effects in previous studies (121, 447, 448).

Specifically, the number of CD56^{++} NK cells has been associated with lower rates of CGvHD (447, 448). Whereas CD56^{dim} NK cell content has been linked with protection against relapse (121). The findings in **Chapter 3** do not directly contradict these reports, as NK cells were not associated with either GvHD or relapse incidence in this cohort. Other transplant-related complications such as infection may therefore have driven the observed reduction in survival rates.

Moreover, the unsupervised clustering analysis failed to divide NK cells into separate CD56^{++} and CD56^{dim} populations. Partitioning cells assigned to the NK cluster based on CD56 expression would be useful for comparison with other studies. An increase in the proportion of NK cells could also reflect an increase in the overall lymphocyte content as was observed between PBSC graft groups (**1**).

Overall, these findings suggest that an increased prevalence of T cells and a decrease in the monocyte content in grafts relative to HSCs was correlated with worse outcomes for allo-HSCT recipients.

5.5 Post-transplant serum proteins as prognostic biomarkers

Numerous studies have identified the association of serum proteins with prognosis for allo-HSCT recipients (118). The study in **Chapter 4** focused on a set of candidate prognostic biomarkers identified by colleagues in the sera of recipients on the 14th day following transplantation.

Variable selection is a technique in machine learning to select a subset of relevant variables for model construction (449). Three variable selection algorithms, LASSO, AIC and RSF were employed to select biomarkers and clinical risk factors associated with prognosis, whilst discarding irrelevant or redundant features with a minimum reduction in performance (341). These variable algorithms are discussed in detail in **Appendix B**.

The analysis identified IL6 expression as being associated with a higher risk of non-relapse mortality. This protein has previously been validated as a prognostic biomarker for allo-HSCT recipients (181-183). Moreover, the data suggests high levels of PD-L1, CCL23, and CDCP1 expression could be prognostic biomarkers for OS, GRFS, and relapse-related mortality, respectively. Demonstrating the methodology identified both established and potentially novel prognostic biomarkers.

5.5.1 PD-L1

PD-L1 is an immunosuppressive molecule that inhibits T cells by activating the PD-1 / PD-L1 signalling pathway. Soluble PD-L1 is elevated in the sera of individuals with multiple cancer types (450). Upregulation of PD-L1 is also observed during acute and chronic viral infections (386, 387, 451).

In **Chapter 4**, both univariate and multivariate analyses indicated the concentration of PD-L1 on the 14th day post-transplant could be used to predict worse OS (adj. HR 1.41; P 0.05). There are a variety of biological functions in which PD-L1 is implicated that could explain this link with prognosis.

During chronic infection and cancer, memory T cells can enter a dysfunctional immune state, termed exhaustion(452). This is characterised by a loss of effector functions, and upregulation of inhibitory receptors such as PD-1, amongst other differential changes(452). In this exhausted state, T cell mediated immune responses to cancer and infection are impaired(453, 454) .

Blocking the PD-1/PD-L1 pathway can restore the function of exhausted T cells(455) (452). Immunotherapy with anti–PD-L1 or anti–PD-1 antibodies has been approved for the treatment of several cancers including Hodgkin's lymphoma (456, 457) . Inhibiting this pathway has been shown to restore anti-cancer activity (458), and has potential as an adjunctive therapy for chronic infections (459).

PD-L1 is expressed on the surface of cells such as T cells, B cells, DCs, and macrophages (460, 461) and by cancerous cells in a large number of malignancies, including multiple myeloma (460-462). Many cancer types co-opt this pathway for immune evasion (463, 464). In B cell lymphoma, the level of soluble PD-L1 in the serum may reflect the amount expressed by cancer cells (465). In vitro, the soluble form of PD-L1 has been shown to be produced by DCs and cancer cell lines (466).

The concentration of soluble PD-L1 in recipients on the 14th day post-transplant could, therefore, be related to the level of PD-1 mediated immune suppression or even by the amount expressed by residual cancerous cells. The receptor PD-1 is upregulated in activated T cell populations in the early days following transplantation (467), so changes in the concentration of PD-L1 at this early time point could have a substantial effect on T cell activity.

Given the role of PD-L1 in immune evasion by cancer and infections, its credible that the level of PD-L1 could be associated with allo-HSCT prognosis. Several studies of non-haematological malignancies

have indicated the use of PD-L1 as a prognostic biomarker, including prostate and breast cancer (468, 469). Recent meta-analyses have indicated PD-L1 expression is a prognostic biomarker for worse survival in lung cancer, when measured in the serum and on cancer cells (470, 471).

However, the data in **Chapter 4** did not indicate which post-transplant complications contributed to the lower rate of survival. An increase in PD-L1 concentration did not confer a greater risk of AGvHD, CGvHD, or relapse risk in this cohort. Although there was a univariate increase NRM[1] (HR_{SD} 1.4; P 0.07) and a decrease in GRFS (HR 1.2; P 0.07).

In healthy individuals, the PD-1/PD-L1 pathway plays an important role in maintaining tolerance in autoreactive T cells (472) and terminating immune responses (462). It is therefore surprising, that murine models of allo-HSCT indicate PD-L1 promotes AGvHD by improving the survival of alloreactive T cells(473). Selective inhibition of PD-L1 on donor-derived T cells has been suggested as a GvHD prophylaxis strategy(473). Other, studies indicate PD-L1 is essential for AGvHD control, so the relationship is unclear.

Whilst no effect on relapse was observed in this cohort, a larger cohort may be required to identify this effect. Other outcomes also warrant investigation; in a large cohort study during which 44,930 deaths were observed after allo-HSCT, 23.8% were attributed to infections (52). The main factors determining late infectious mortality remain largely unknown (52). Given the upregulation of PD-L1 seen in acute and chronic infection, the association with late infectious mortality may be of interest.

The expression of soluble PD-L1 in allo-HSCT patients is relatively understudied, however, there are several biologically plausible reasons why it may have an association with prognosis. The data in **Chapter 4** suggests the potential utility of PD-L1 as prognostic biomarker for long-term outcomes when measured early post-transplant.

5.5.2 CDCP1

CDCP1 is thought to play a role in cancer survival and metastasis (474-477) and has been associated with poor prognosis in a variety of solid cancers (478, 479). It has also been suggested as a biomarker for leukaemia diagnosis (385).

In both univariate and multivariate analysis, CDCP1 concentration in the serum post-transplant was associated with an increased probability of relapse-related mortality (adj. HR_{SD} 2.41; P 0.01) (**Chapter 4**).

Overexpression of CDCP1 has been reported by a variety of studies in colorectal, breast, and lung cancers (385, 480). A recent study of AML patients identified CDCP1 expression on leukemic cells as a prognostic biomarker for worse OS in patients receiving intensive therapy (481).

Therapeutic antibody candidates to block CDCP1 have been investigated in mouse models(482). Stable over-expression of CDCP1 led to faster tumour growth, and downregulation inhibited tumour growth (482). Whilst there was an increase in RM with CDCP1 expression, the increase in relapse incidence was insignificant (adj. HR_{SD} 1.29; P 0.16).

The action of CDCP1 may contribute to resistance to targeted therapeutics in chronic myeloid leukaemia (483) and breast cancer cell lines (484). Moreover, it is thought to confer resistance to cytotoxic chemotherapy in pancreatic cancer(485).

[1] NRM (non-relapse mortality) was analysed instead of TRM (transplant-related mortality) in **Chapter 4**, because the specific cause of death was not formally adjudicated, and NRM includes all deaths without prior relapse.

If there is a true association between RM probability and CDCP1, it may be related to their particular malignancy and stage of the disease. It would be useful to compare the relative expression of CDCP1 between different disease stages and diagnoses to clarify this effect.

As well as its role in cancer, CDCP1 is expressed on HSCs under normal conditions and may play a functional role in the early stages of haematopoiesis (385). It is also expressed by non-haematopoietic cells including epithelial cells and mesenchymal stem cells (474).

A study of allo-HSCT grafts derived from cord blood suggested CDCP1 might act as a diagnostic biomarker of graft CD34+ content(478). However, no correlation between CDCP1 and CD34+ dose was observed in **Chapter 4** (Spearman's ρ -0.07; P 0.43).

The role of CDCP1 in cancer survival and expression in leukemic cells, suggests a prognostic link with relapse-related mortality is credible, and worth validation in future studies.

5.5.3 CCL23

CCL23 is a chemokine produced by eosinophils, myeloid DCs, and monocytes (486). Differential regulation of CCL23 has been identified before diagnosis of AGvHD in allo-HSCT recipients (487).

Both univariate and multivariate analyses indicated the concentration of CCL23 could be used to predict worse GRFS (adj. HR 1.3; P 0.01) and increased AGvHD (adj. HR_{SD} 1.66; P 0.01) (**Chapter 4**). There was also a link with worse OS in a univariate comparison (HR 1.41; P 0.04) but was not retained after variable selection.

CCL23 is the ligand for the CCR1 receptor(488). The interaction between CCR1 and other ligands on donor-derived cells contributes to the development of both GvHD and GvL effects following allo-HSCT(489). The action of the CCL23/CCR1 pathway on allogeneic immune responses is worth further investigation.

CCL23 had an insignificant impact on relapse (adj. HR_{SD} 1.42; P 0.16) and RM (adj. HR 1.94; P 0.16) in this cohort, however, there are a variety of studies indicating a link between CCL23 and cancer promotion. CCL23 acts as a chemoattractant to DCs, monocytes and resting T cells (486). Due to its capacity to recruit these cells, CCL23 may contribute to anti-cancer cytotoxic effects (486, 490, 491).

However, it also has many pro-cancer properties. CCL23 can induce angiogenesis, the formation of new capillaries from existing blood vessels (396, 397). Solid tumours require an adequate blood supply to proliferate, so angiogenesis is instrumental in this process (492). A similar relationship between angiogenesis and cancer promotion has been described in several haematological malignancies (396).

CCL23 has been shown to enhance the action of vascular endothelial growth factor (VEGF), which also has angiogenic properties (396, 486, 493). Expression of VEGF correlates with clinical characteristics of leukaemia and non-Hodgkin's-lymphoma (396). Moreover, activation of the CCL23 CCR1 pathway has also been linked to the suppression of T cell activity (494, 495).

This indicates there are several possible mechanisms through which CCL23 may be associated with poor prognosis following allo-HSCT. CCL23, therefore, represents a promising candidate biomarker for the rate of GRFS and AGvHD. The impact of CCL23 on short-term morbidity and mortality following transplant should be validated in independent data.

5.5.4 IL6

IL6 is a cytokine with pro and anti-inflammatory functions, associated with several inflammatory diseases (496, 497). Studies of allo-HSCT patients have suggested IL6 plays an important role in the

pathophysiology of AGvHD (392, 498-500). Notably, IL6 has already been identified and validated as prognostic biomarker for increased NRM and AGvHD (181, 182). IL6 measured at Day 7 post-transplant is associated with greater risk of AGvHD and NRM as well as lower survival rates in allo-HSCT patients (392).

In line with this understanding, elevated IL6 concentration was associated with an increase in the incidence NRM (adj. HR_{SD} 1.5; P 0.05) in **Chapter 4**. There was also a decrease in RM incidence (adj. HR_{SD} 0.47; P 0.05) and a trend towards lower relapse (adj. HR_{SD} 0.72; P 0.08).

This data could indicate several possible scenarios in this cohort. IL6 concentration could have been linked with a rate of NRM so high that fewer patients survived until relapse. Alternatively, IL6 could be linked with enhanced GvL effects that protected against relapse but increased NRM risk. As IL6 is a validated prognostic biomarker for NRM, the first scenario is more plausible.

IL6 is expressed by multiple cell types including DCs, macrophages, and T cells as well as non-haematopoietic cells (501). Blockade of IL6 signalling in mice models has been associated with lower GvHD severity and prolonged survival (502, 503). Early clinical trials with anti-IL6 antibodies, such as tocilizumab, have investigated the use of IL6 as a therapeutic target to prevent AGvHD (500, 504). The benefit of IL6 blockade is thought to be mediated by an increase in the number of T-reg cells, and inhibition of differentiation in other effector T cells(502, 505) .

IL6 expression is increased early post-transplant, particularly following TBI (500, 506) . The intensity of conditioning could therefore have influenced IL6 expression. In **Chapter 4**, a significant protective effect of RIC on OS was observed (adj. HR 0.44; P 0.02), but TBI had no univariate association (HR 1.15; P 0.70). The influence of the conditioning regimen on IL6 expression should be investigated in this cohort.

Whilst the association of IL6 with NRM is not a novel finding, this emphasises that the methodology identified validated prognostic biomarkers for recipient outcomes. This provides additional confidence that the other candidate biomarkers identified warrant validation.

5.5.5 CXCL10

CXCL10 is a cytokine that specifically activates CXCR3. This protein is implicated in a variety of diseases including immune dysfunction, chronic inflammation, and cancer development (395). Studies have suggested the interaction between CXCL10 and CXCR3 plays a role in graft rejection in solid organ transplants (507-509). In allo-HSCT, expression of CXCL10 is a validated prognostic biomarker for CGvHD a major cause of NRM (118, 183).

In line with this, CXCL10 concentration was associated with a univariate increase in NRM (HR_{SD} 1.42; P 0.04) in **Chapter 4**. Though this effect did not retain significance after adjusting for covariates (adj. HR_{SD} 1.29; P 0.34). The role of CXCL10 in attracting immune cells means it exacerbates inflammation and tissue damage in inflammatory disease (510). Along with its role in CXCR3 activation, this could explain a potential link between CXCL10 and increased NRM.

In contrast, CXCL10 concentration was associated with lower relapse incidence in both univariate and multivariate comparisons (adj. HR_{SD} 0.66; P 0.01). There was also a trend toward lower RM (adj. HR_{SD} 0.57; P 0.06). Whilst this could be related to the high rate of NRM, its possible CXCL10 had a protective effect.

CXCL10-mediated apoptosis has been shown to inhibit the growth of cervical cancer (511). Its potential anti-cancer properties are thought to be related to the stimulation of immune effector responses (512), by acting as a chemoattractant for cells including T cells, NK cells, DCs, and

macrophages and B cells (513, 514). Furthermore, CXCL10 plays a role in the attenuation of angiogenesis (395, 515).

However, the relationship with cancer is complicated as elevated expression of CXCL10 and its receptor CXCR3 has been associated with lymphoma and multiple myeloma amongst other cancers (395, 516, 517). CXCL10 colocalises with a proliferation marker in cancer cells (395) and its receptor CXCR3 plays a role in both suppression and promotion of cell proliferation (518).

CXCL10 may therefore have both pro- and anti-cancer properties, so any potential link with decreased relapse incidence should be interpreted with care. Moreover, the increase in NRM with CXCL10 expression did not appear to be driven by CGvHD (adj. HR_{SD} 0.96; P 0.83), for which CXCL10 is a validated biomarker.

Whilst the prognostic association of CGvHD with CXCL10 could not be confirmed in this cohort, the data indicated a potential link between CXCL10 and the incidence of NRM and relapse. That the analysis identified a second validated risk factor for allo-HSCT recipients along with IL6, provides additional confidence in the findings of the study.

5.5.6 LIF-R

Leukaemia inhibitory factor, LIF, is a multifunctional cytokine with a wide variety of roles in inflammatory responses and differentiation of leukaemia cells (398). Along with IL6, LIF is part of the IL6 superfamily and binds to its specific receptor LIF-R (398).

The absolute protein concentration of LIF-R was quantified by Kriti Verma, and no relationship between protein concentration and Olink NPX values was observed in **Chapter 4**. The Olink data for this analyte may therefore be erroneous, and HR_{SD} estimates may be artefacts of noise, rather than true effects.

The data in **Chapter 4** suggested a potential link between LIF-R expression and increased RM incidence, as well as a decrease in NRM. The direction of the effects was the same in both univariate and multivariate comparisons, however, the significance and magnitude changed substantially. This casts doubt on whether these were true effects, particularly given the low EPV ratio of the multivariate models (see **Section 5.6.1.7** on "overfitting" in the limitations).

The impact of LIF-R / LIF signalling is relatively understudied in the context of allo-HSCT. However, a few studies indicate functions that may be relevant. Increases in LIF expression have been linked with decreased thymic function (519). LIF also has a complex relationship with cancer progression, with both inhibitory (520) and cancer promoting effects (521-523). Moreover, the expression of LIF has been linked with worse prognosis in a variety of non-haematological cancers (521, 524, 525).

Overall, the study in **Chapter 4** was inconclusive and the association between LIF-R and prognosis was unclear. Given the potential for the Olink data to be erroneous, the analysis should be repeated with LIF-R excluded.

5.6 Limitations

5.6.1 Limitations of Allo-HSCT graft composition study

The following limitations refer to the study described in **Chapter 3.**

5.6.1.1 Donor and recipient cohorts

When conducting an exploratory data analysis, with a single cohort, there is a risk of over-interpreting the scarce available data. This is particularly true when the cohort is relatively small and clinically diverse - so these results should be interpreted cautiously. However, if nothing else, the study in

Chapter 3 has increased the amount of data that is available for speculation as to the impact graft composition has on prognosis.

There was some heterogeneity in the source of the 185 allo-HSCT grafts. Whilst the majority were extracted from PBSC, 4 were from BM, and 1 was from cord blood. Compositional data analysis was focused on a subset of 155 grafts (153 PBSC; 2 BM) for which outcome data was available for the recipients. The BM grafts could have been excluded from the analysis to improve interpretability.

Multiple studies have shown stem cell mobilisation regimes influence the T cell, NK, DC, and monocyte content of grafts (404, 409, 444). Standard G-SCF mobilisation was used for all the PBSC grafts, therefore the results may not generalise to the use of alternative mobilisation agents such as Plerixafor (409).

The grafts were unmanipulated at the time of CyTOF data collection. Information on T cell depletion was available at a later stage of project development. This showed that 93% of the recipients were subject to in vivo T cell depletion with ATG or Alemtuzumab post-transplant. Subsequent testing indicated no differences in the rate of OS and GRFS between ATG or Alemtuzumab recipients (**Figure C.14**). Moreover, there was no difference in the cumulative incidence of GvHD, relapse, or TRM. However, the interaction between the composition and subsequent pharmacological interventions could be a confounding factor.

Another underexplored effect was the effect of donor characteristics on composition. Qualitative comparisons were made between PBSC graft groups. However, the effects of donor age, CMV status, and gender on graft composition were not formally tested.

Some important risk factors were omitted from the analysis. Whilst information on disease stage was available for some patients, a large number of categories (10 categories) and disease-specific interpretation rendered this information unsuitable for inclusion in the statistical analysis given the cohort size. The results are only applicable to HLA-matched allo-HSCT. This variable was not included in the regression analysis due to the low prevalence of mismatched donors (7%).

Diagnosis was dichotomised into myeloid and lymphoid malignancies. This does not accurately reflect the diversity of risk between diseases, and the action of particular cellular components may vary depending on the specific haematological malignancy. Moreover, 10 patients had other malignancies or aplastic anaemia, further complicating the comparison. Therefore, diagnosis was omitted from the regression models.

The mismatch between the sex of the recipient and donor was also encoded as a binary variable, either matched or mismatched sex. However, female recipients of grafts from male donors may have an increased risk of GvHD (160).

5.6.1.2 Batch effects

Grafts were subject to CyTOF data collection across 8 batches. If strong batch effects are present, this can result in clusters that are specific to batch differences – masking differences that are of interest to the study. Batch correction is an active research topic for CyTOF data (526-528). Many effective solutions require the inclusion of technical replicates during data acquisition (528). Recently, robust integration methods to minimise technical variability without the use of replicates have been published but were not available during this project's development (529).

Visualisation of graft centroids – the average marker expression across all markers – by principal component analysis did not indicate significant batch effects (**Figure C.1**). PhenoGraph appeared to be fairly robust to batch differences. Of the 39 clusters, 4 were excluded for being observed in < 50%

of samples, and 32 were observed in > 90% of samples. However, batch effects could be influential so correction needs to be integrated into the analysis (529).

5.6.1.3 Missing data

Missing data was present in the recipient and donor information. A complete case strategy (excluding patients with any missing observations) was not suitable given the impact it would have on the sample size. Therefore, a multiple imputation strategy was applied (as described in **Section 2.1.10.1**) under the assumption that the data was missing at random or missing completely at random.

Variables with missing observations consisted of routinely collected information for allo-HSCT patients. The missingness likely resulted from data administration, rather than not being recorded in practice. It is therefore reasonable to assume that the observations were missing at random. Sensitivity analysis should be conducted to assess the influence imputation had on the resulting coefficient estimates.

In the context of survival analysis, when imputing missing covariate values, it is recommended that the outcome data is included in the imputation model (324) in the form of an event indicator and the baseline cumulative hazard (530). It is unclear whether this approach should be applied when modelling the sub-distribution hazard such as with the FG model (324, 531). Investigating multiple competing outcomes would also require different imputation models. Furthermore, this increases the risk of data leakage. Thus, only independent variables available at the time of model evaluation were included in the imputation model, and the same set of imputed data was utilised for each comparison.

5.6.1.4 Cell state

By focusing on the prognostic effect of cell abundance, the current study overlooks the wealth of biological information encoded in the state of cells. Certain markers may be expressed during signalling or other functional states such as inflammation (532). The impact differences in cell state had on prognosis needs to be explored and could be used to refine the PBSC graft groups.

5.6.1.5 Defining composition of PBSC graft groups

Cells were divided into cell populations with the PhenoGraph clustering algorithm. The unsupervised nature of this approach meant the cell populations are not easily reproducible in independent data. Samples were pooled into a single matrix prior to clustering; however, the number of cells was highly heterogenous between donations. As a result, grafts with larger cell counts could have had an outsized influence on the cluster analysis.

When comparing the composition of the three PBSC graft groups, cellular proportions were calculated from the total count of cells. This gave an unbiased comparison of every cell lineage identified within the CyTOF data. To improve interpretability of the results, additional summaries targeted at the T or B cell compartment would give insight to the relative contribution of naïve and effector cell subsets. Likewise, the relative proportion of pDC and cDC within the DC compartment could be revealing.

The defining characteristics of the PBSC graft groups were determined qualitatively. Statistical tests to identify proportionally abundant features could add clarity to graft group comparisons (533).

5.6.1.6 Amalgamated cell populations

Compositional data analysis was conducted with clusters of cells detected by PhenoGraph combined into lineages. These are referred to as summed or amalgamated log-ratios, in which parts of the composition are combined using domain specific knowledge (270).

Amalgamating parts of the composition was necessary due to the restricted sample size and the prohibitively large number of cell populations within the data (see next heading). The downside is that that this overlooks phenotypic and functional diversity within each lineage.

It would be useful to divide the composition into individual clusters or more finely grained cell subsets (e.g., T cells divided into CD4+, CD8+, etc). However, methodological changes would be required to analyse high-dimensional compositions which are discussed in **Section 5.7.4**.

5.6.1.7 Overfitting

The number of clinical variables adjusted for in the multivariate models was substantial given the number of events observed in the cohorts. This was compounded by the number of cell lineages included in the compositional models, even after amalgamation. The higher the events-per-variable (EPV) ratio, the lower the risk of bias in the model estimates. The multivariate models for GRFS had EPV ratios between 5.8 and 8. Simulation studies suggest models with an EPV > 5 can be interpreted without a substantial risk of bias (361).

The multivariate OS models had less favourable EPVs, ranging from 3.1 to 4.2. However, the remaining FG models all had an EPV < 3. The multivariate estimates should therefore be interpreted with caution as there is a substantial risk of bias and generalisation error. Dividing the data into a training and testing set would be useful to validate the hazard ratios but would be impractical given the sample size and use of unsupervised clustering.

A reduced set of clinical variables should have been selected for inclusion at the outset of the study, or a variable selection strategy should have been employed to construct more parsimonious models.

5.6.1.8 Additive log-ratio transformation

Survival analysis for compositional data was conducted using the additive log-ratio (ALR) transformation. Other types of log-ratio transformations have been proposed, such as the centred log-ratio (CLR) and the isometric log-ratio (ILR) as introduced in **Section 2.1.6.1**.

McGregor et al provide a persuasive study purporting the use of ILR transformation for survival analysis (335). Whilst theoretically justified, CLR and ILR introduce considerable complexity when an approximately isometric approach such as ALR can lead to the same inference (270). Moreover, of the three transformations, only the ALR has an intuitive interpretation, as it is constructed from the ratio of two parts of the composition.

Greenacre 2021(262) argues that alternative transformations complicate the practice of compositional data analysis whilst losing the benefit of easy interpretation. John Aitchison, the statistician instrumental in establishing the principles behind compositional data analysis, suggests that

> "The ALR transformation methodology has, in my view, withstood all attacks on its validity as a statistical modelling tool. Indeed, it is an approach to practical compositional data analysis which I recommend particularly for non-mathematicians. The advantage of its log-ratios involving only two parts, in contrast to CLR and ILR […], which use log-ratios involving more than two and often many parts, makes for simple interpretation and far outweighs any criticism, more imagined than real, that the transformation is not isometric." (534)

As a result, the importance of isometry is widely debated (270). Ultimately the choice of log-ratio transformation depends on the particular application, the data, and the audience with which the analysis it to be communicated.

The ALR transformation requires a choice of a suitable reference part. As HSC are the essential functional constituent of grafts, the ratio of other cell types to HSC is relatively intuitive in this context. With HSC as the reference part, the Procrustes correlation for the dataset was 0.9 (**Appendix Table C.1**), whereas the maximum correlation achieved was 0.94 with DCs as the reference. This indicates that the ALR transformed data was approximately isometric and that the use of alternative reference parts did not lead to substantial improvements.

5.6.1.9 Competing risks analysis for compositional data

Survival analysis using the Cox PH model has recently been extended for use with log-transformed compositional covariates (334) (335). This approach was used for studying the rate of OS and GRFS, which are not subject to competing risks.

As described in **Section 1.3.2**, the CSH and FG models are the two most common multivariate methods when studying competing risks in allo-HSCT data. To the best of my knowledge, neither has been applied to compositional covariates in published literature.

The CSH model is effectively the Cox PH model applied to each cause-specific hazard. The analysis could be conducted as demonstrated for survival events (334-336), only repeated for each competing risk and interpreted accordingly. Whereas to the best of my knowledge there is no published software designed to apply the FG model to compositional covariates. Therefore, code from Walmsley et al. 2021 (334, 336) was adapted for this purpose. Consequently, applying the FG model to compositional covariates is an experimental analysis that warrants formal evaluation.

A literature search identified several papers on mortality forecasting in which competing causes of death were modelled as a composition (535-537). However, competing risks regression with compositional covariates appears to be an understudied area of research.

5.6.1.10 Cluster analysis for compositional data

Grafts were partitioned into clusters, or PBSC graft groups, based on compositional differences. This relied on cluster analysis for compositional data (clustCoDa) a technique developed in the field of geochemistry (272, 279). Compositional data analysis is well-established in geochemical research; however, there has recently been exponential growth in the use of such methods in other fields of study (538).

The use of clustCoDa to partition the composition of CyTOF data into groups represents a new use case, but the methodology is the same: to apply an isometric log-ratio transformation (263, 272, 273), then perform cluster analysis on the transformed data. Clustering in this context is used as a form of feature extraction. The multi-dimensional composition is transformed into a single discrete variable, the cluster label, that is compatible with standard survival and competing risks analysis.

The GMM clustering algorithm was selected for clustCoDa due to reports of that it can out-perform other common clustering algorithms such as *k*-means (272, 275). A variety of other clustering algorithms such as Density-Based Spatial Clustering of Applications with Noise would be interesting to explore in this context (539).

A major benefit of GMM clustering is that the fitted model can be re-used to classify new observations into the same set of clusters (or PBSC graft groups as we have referred to them). This feature facilitates validation of PBSC graft groups and the associated risk in independent data sets. If validated, such a model could form the basis of a risk stratification tool to stratify grafts into risk groups based on multivariate compositions. This would, however, also require the cellular populations used to define the graft composition to be reproducible (see **Section 5.7.3**).

5.6.2 Limitations of post-transplant serum protein study

The following limitations refer to the study described in **Chapter 4.**

5.6.2.1 Recipient cohort

Results based on this single cohort represent an exploratory data analysis, for which independent data is required to confirm findings. Dividing the cohort into a training and testing set would have further lowered the EPV ratios. Moreover, the initial selection of the candidate biomarkers utilised the entire cohort, so this would have introduced bias.

Some important risk factors, such as disease stage and HLA-mismatch were omitted from the analysis. Whilst information on disease stage was available for a subset of patients, inconsistencies in the reporting rendered this information unsuitable for use in statistical analysis. Due to the low prevalence of mismatched donors (11%) this variable was not included in regression analysis.

Several variables including diagnosis and donor/recipient sex mismatch were simplified to enable inclusion with the regression models. Missing data was also present. The implications of these limitations are described in **Section 5.6.1.1** and **Section 5.6.1.3**.

5.6.2.2 Day 14

The Olink data was generated from samples taken on the 14th day post-transplant. There is an inherent assumption that the protein expression at this time point is associated with prognosis. However, it could be that the pattern of protein expression was fixed by the state of the subject's immune system prior to transplantation. Data from earlier time points would be required to assess this. Kriti Verma and Wayne Croft are conducting an extension of the study with samples taken on the day of transplantation.

5.6.2.3 OLINK relative protein concentration

The Olink NPX assay provides a relative measure of serum concentration between individuals in this particular cohort. However, the values are not comparable between studies. This is suitable for the current exploratory analysis but limits the reproducibility of findings and complicates comparison with other experiments.

5.6.2.4 Initial selection of candidate biomarkers

The initial set of candidate biomarkers were selected from the Olink data by colleagues based on univariate differential expression analysis, with BH correction to control for the false discovery rate. Whilst univariate filtering raises the possibility of overlooking effects that may be larger when observed in concert, it limits the risk of interpreting spurious findings.

A follow-up time was not specified in the differential expression analysis. In the comparison of mortality, for example, patients who died within weeks are treated as equivalent to those who died many years later. Proteins with a large impact on the rate of mortality but not the overall number of deaths may, therefore, have been overlooked. However, the major benefit of a two-group comparison is it is simple to convey to the audience that protein concentration was elevated/decreased in a particular group of patients (e.g., those that subsequently died).

5.6.2.5 Multiple testing burden

Statistical models were used to estimate hazard ratios for each candidate biomarker, factoring in time and censoring. Whilst this analysis was restricted to differentially expressed proteins, retained after BH correction, this introduces additional unaccounted multiple testing. This is further complicated by the use of multiple closely correlated clinical endpoints.

Applying additional multiple testing corrections would raise the possibility of false negatives – overlooking important associations. For example, if a single protein was tested for its effect on the incidence of RM and relapse, and the coefficients had significance values of P 0.06 and P 0.05, applying the BH correction would mean neither result was significant. This is counter intuitive as the two outcomes are closely related. Therefore, P values were left unadjusted for the statistical models – and the significance of the results is debatable.

5.6.2.6 Overfitting

The number of candidate biomarkers and clinical variables were substantial relative to the number of events observed. A variable selection strategy was employed to create more parsimonious models.

After variable selection, the EPV ratios for the multivariate models of OS and GRFS were 6 and 19.4, respectively. Indicating low risk of over-fitting. This was not the case for the FG models due to the number of variables retained after variable selection. The multivariate models of AGvHD and relapse had less favourable EPVs of 3.8 and 2.9 respectively. However, the remaining FG models all had an EPV < 2, so there is a high risk of overfitting. The risk of interpreting spurious findings was limited by focusing primarily on effects that were apparent in both the univariate and multivariate comparisons.

A restricted set of clinical variables should have been selected for inclusion, ideally at the outset of the study. Additional filtering of the clinical variables post-selection would also improve EPV ratios.

5.6.2.7 Variable selection

Multiple variable selection algorithms were employed in this study. The advantages and limitations of each of these algorithms applied are discussed in **Appendix B**. To choose a final model among the variable selection results, BIC was compared (360, 403). Whilst this added additional complexity to the analysis, the selection of the same candidate biomarkers, by several variable selection algorithms increased confidence that true signal variables were being interpreted.

As this was an exploratory analysis and variable selection was employed to reduce the dimension of the data, this could have increased the rate of type II errors (false negatives) (540). Therefore, differentially expressed proteins excluded by variable selection should not be discounted from follow-up studies.

5.7 Future work

Based on the results and potential limitations of the studies reported in this thesis, there are several possible directions for future work

5.7.1 Cell state & donor characteristics

In an extension of the study in **Chapter 3**, I would like to contrast the cell state of immune populations between grafts. Currently, this important biological information is only utilised to characterise cell populations.

Contrasting the expression of immune checkpoint inhibitors, such as PD-1, on monocytes could be informative given the important role of this pathway in T cell suppression (541). The activation status of DC populations could be tested for an association with GvHD (440). Such as by measuring the level of CD86 expression which is upregulated during antigen presentation (542).

The simplest approach would be to contrast the average marker expression of a cell population between grafts. This is a fast reliable strategy for differential state analysis (543).

The study in **Chapter 3** did not statistically test the association between donor characteristics and graft composition. Specifically, donor age, gender, and CMV status should be assessed. Differential abundance analysis could be employed to study these factors (544).

5.7.2 Subgroup analysis

Given the restricted cohort sizes, subgroup analysis was not practical for the reported studies. In **Chapter 3,** survival analysis for compositional data was applied to all PBSC graft groups as a single population. Testing the influence of graft composition within each of the three PBSC graft groups could reveal differential associations dependent on the overall composition.

Moreover, subgroup analysis to contrast different haematological malignancies is required, as the level of heterogeneity in **Chapters 3** and **4** reduces interpretability. The specific malignancy may have influenced protein expression or the impact of particular cell subsets in the graft. Likewise, given the reported effect of TBI on IL6 expression, a subset analysis focusing on different conditioning regimens could unpick these potentially confounding effects (500, 506).

5.7.3 Risk stratification tool for graft composition

In an extension of the study in **Chapter 3,** I would like to explore the use of supervised clustering analysis to define cell populations. Supervised clustering uses prior information to cluster CyTOF events into known cell subsets and would result in clusters that are easily reproducible between datasets.

The combination of supervised clustering to identify cell populations, with clustCoDa to define groups of grafts, could provide a powerful tool stratify allo-HSCT grafts by risk. A related approach was used to stratify COVID-19 patients by risk in a study I published with collaborators during the course of this thesis (**Appendix A**).

Linear Discriminant Analysis (545) and DeepCyTOF (546) are two supervised clustering algorithms shown to accurately classify cytometry data into known cell types (547). However, these require substantial amounts of labelled training data and cannot identify novel cell types.

5.7.4 Higher-dimensional compositions

Conducting the compositional data analysis in **Chapter 3** with high-dimensional compositions could provide additional insights into the effect of individual cell populations (e.g., CD4+ T cells, CD8+ T cells, etc) or clusters of cells.

The current study amalgamated cell populations into lineages to overcome the restricted sample size. Another solution would be to retain all individual parts of the composition and apply multivariate variable selection to create a more parsimonious model. AIC (335) and LASSO selection (548), as applied in **Chapter 4**, have both been applied to simplify regression models with compositional covariates. This could enable survival regression with high-dimensional compositions.

The same approach could be extended to the cluster analysis for compositional data used to identify the PBSC graft groups. For example, by applying variable selection methods for model-based clustering (549), thereby rendering cluster analysis for compositional data compatible with high-dimensional compositions. Alternative compositional clustering algorithms have also been developed specifically for high-dimensional compositions (275, 550).

In microbiome research, it is common to have compositions with a very large number of parts but relatively small sample sizes. The compositions in microbiome data are often very sparse with 50 -90% zero values (551). Survival analysis methods have been developed for this sparse microbiome data

(552, 553). Namely, optimal microbiome-based survival analysis (OMiSA) (552), and microbiome regression-based kernel association test (MiRKAT-S) (553).

Both OMiSA and MiRKAT-S are compatible with log-ratio transformations (552, 553) and rely on aspects of the Cox regression model, so are suitable for survival events. Tools to test the effect of sparse compositions on competing risks (e.g., using the FG model) would be invaluable, but to the best of my knowledge, none have been published.

A combination of these approaches could be explored to test the impact of individual cell populations on prognosis. As well as to define the PBSC graft groups based on high-dimensional compositions.

5.7.5 Absolute protein concentration models

For a subset of candidate biomarkers in **Chapter 4**, Kriti Verma conducted an ELISA assay to quantify absolute protein concentration. This data could be used to estimate hazard ratios per unit change in absolute protein concentration for the candidate prognostic biomarkers. As well as being more biologically interpretable, this would enable validation of the estimates and facilitate risk model development. Both the EASIX and rPAM risk scores discussed in **Section 1.2.6.1**, for example, utilise biomarkers to estimate risk in allo-HSCT recipients (173).

5.7.6 CCL23 / CCR1 interactions and GvHD

The data in **Chapter 4** indicated a potential link between CCL23 and AGvHD. Given the role CCR1 / CCL5 interaction plays in GvHD development (489), the action of CCL23 / CCR1 interaction on allogeneic T cell responses could be investigated for its influence on GvHD and GvL effects.

5.7.7 Integrating CyTOF and OLINK proteomics data

The absence of paired samples between the cohorts in **Chapters 3** and **4**, meant each cohort of allo-HSCT recipients was studied in isolation. Had the data been derived from paired samples, the CyTOF and OLINK proteomics data could have been integrated. This could reveal novel biological insights, such as the influence of graft composition on post-transplant protein expression. Taking a systems-level approach could also identify interactions between these two data sets that are informative to recipient prognosis.

A study by Lakshmikanth et al. 2017, used topological data analysis to integrate CyTOF and OLINK data after allo-HSCT (487). Using this approach, they found a series of immune signatures pertinent to prognosis. For example, the dysregulation of CCL23 in recipients prior to AGvHD diagnosis. That CCL23 was implicated in AGvHD in **Chapter 4**, further highlights the potential utility of this approach.

The UMAP algorithm used for visualisation in **Chapter 3**, relies on topological data analysis for dimension reduction. UMAP could also be applied to integrate the CyTOF, OLINK, and clinical data to identify co-regulation patterns in a similar approach to that of (487), as well as to develop a prognostic model (305). An example of using UMAP to integrate several types of clinical data for risk stratification can be seen in the published study in **Appendix A**.

5.8 Conclusions

Many challenges remain in understanding the cellular composition of PBSC grafts, and how this influences prognosis following allo-HSCT. Furthermore, the prognostic utility of serum proteins post-transplant requires clarification. This thesis has contributed to our knowledge of graft cellular composition and expanded our understanding of which cell populations and serum proteins are associated with prognosis. These findings will support the development of personalised graft composition strategies and tools for individual risk assessment, early after transplantation. Moreover, it has demonstrated the benefit of applied data science for discovery in allo-HSCT research.

Comprehensive characterisation of immune cell populations improved our understanding of PBSC graft composition and how this varies between donations. Statistical models compared the composition between grafts to test for association with prognosis and the presence of distinct graft types. This indicated a clear association between non-CD34+ content of grafts and prognosis and the presence of three distinct types of graft composition that stratified recipients by risk.

Established associations were reinforced, such as graft T cell content being predictive of chronic GvHD risk. It also raised several important questions such as why the proportion of monocyte and DNAM-1^{+} lymphoid progenitor cells in grafts conferred improvements in GRFS and TRM, respectively. As these cell populations did not influence the risk of relapse, they could be studied for their potential to impede GvHD without obstructing GvL effects.

Valuable insights have been obtained into the impact of post-transplant protein concentrations on prognosis, highlighting the utility of variable selection algorithms in biomarker discovery. A previously validated prognostic biomarker, IL6, was shown to be predictive of NRM incidence. For the first time, the concentration of PD-L1, CCL23, and CDCP1 in the serum 14 days post-transplant were identified as potential prognostic biomarkers for OS, GRFS, and RM, respectively. If validated, these could help advance the development of biomarker-based risk scores.

Furthermore, PhenoGraph was shown to be capable of identifying cell populations from millions of CyTOF events, derived from heterogeneous graft samples. Both well-defined and less-studied populations were identified based purely on patterns within the CyTOF data. This emphasises how unsupervised cluster analysis could facilitate the interrogation of large-scale single-cell data for clinical research.

Bibiliography

1. Kindt TJ, Goldsby RA, Osborne BA, Kuby J. Kuby immunology. 6th ed. / Thomas J. Kindt, Richard A. Goldsby, Barbara A. Osborne. ed. New York: W.H. Freeman; 2007.
2. McDuffie FC. Antigens and Immunogens. In: Safai B, Good RA, editors. Immunodermatology. Boston, MA: Springer US; 1981. p. 35-45.
3. Marshall JS, Warrington R, Watson W, Kim HL. An introduction to immunology and immunopathology. Allergy, Asthma & Clinical Immunology. 2018;14(2):49.
4. Conti B, Tabarean I, Andrei C, Bartfai T. Cytokines and fever. Frontiers in bioscience : a journal and virtual library. 2004;9:1433-49.
5. Sprague A, Khalil R. Inflammatory Cytokines in Vascular Dysfunction and Vascular Disease. Biochemical pharmacology. 2009;78:539-52.
6. Roh J, Sohn D. Damage-Associated Molecular Patterns in Inflammatory Diseases. Immune Network. 2018;18.
7. Uribe-Querol E, Rosales C. Phagocytosis: Our Current Understanding of a Universal Biological Process. Frontiers in Immunology. 2020;11.
8. Guillerey C, Huntington ND, Smyth MJ. Targeting natural killer cells in cancer immunotherapy. Nature Immunology. 2016;17(9):1025-36.
9. Vivier E, Ugolini S, Blaise D, Chabannon C, Brossay L. Targeting natural killer cells and natural killer T cells in cancer. Nature Reviews Immunology. 2012;12(4):239-52.
10. Chiossone L, Dumas P-Y, Vienne M, Vivier E. Natural killer cells and other innate lymphoid cells in cancer. Nature Reviews Immunology. 2018;18(11):671-88.
11. Wu S-Y, Fu T, Jiang Y-Z, Shao Z-M. Natural killer cells in cancer biology and therapy. Molecular Cancer. 2020;19(1):120.
12. Stone KD, Prussin C, Metcalfe DD. IgE, mast cells, basophils, and eosinophils. Journal of Allergy and Clinical Immunology. 2010;125(2, Supplement 2):S73-S80.
13. Wen T, Rothenberg M. The Regulatory Function of Eosinophils. Microbiology spectrum. 2016;4.
14. Bonilla FA, Oettgen HC. Adaptive immunity. Journal of Allergy and Clinical Immunology. 2010;125(2, Supplement 2):S33-S40.
15. Uzman A. Molecular biology of the cell (4th ed.): Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. Biochemistry and Molecular Biology Education. 2003;31(4):212-4.
16. Nemazee D. Mechanisms of central tolerance for B cells. Nature Reviews Immunology. 2017;17(5):281-94.
17. Iwanami S, Iwami S. Quantitative Immunology by Data Analysis Using Mathematical Models. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of Bioinformatics and Computational Biology. Oxford: Academic Press; 2019. p. 984-92.
18. Ng IC, Pawijit P, Tan J, Yu H. Anatomy and Physiology for Biomaterials Research and Development. In: Narayan R, editor. Encyclopedia of Biomedical Engineering. Oxford: Elsevier; 2019. p. 225-36.
19. Crotty S, Felgner P, Davies D, Glidewell J, Villarreal L, Ahmed R. Cutting Edge: Long-Term B Cell Memory in Humans after Smallpox Vaccination. Journal of immunology (Baltimore, Md : 1950). 2003;171:4969-73.
20. Thau L, Asuka E, Mahajan K. Physiology, Opsonization. 2020.
21. Lo Nigro C, Macagno M, Sangiolo D, Bertolaccini L, Aglietta M, Merlano M. NK-mediated antibody-dependent cell-mediated cytotoxicity in solid tumors: biological evidence and clinical perspectives. Annals of Translational Medicine. 2019;7.
22. Yoshimura S, Bondeson J, Foxwell BMJ, Brennan F, Feldmann M. Effective antigen presentation by dendritic cells is NF-??B dependent: Coordinate regulation of MHC, co-stimulatory molecules and cytokines. International immunology. 2001;13:675-83.
23. Klein L, Kyewski B, Allen PM, Hogquist KA. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). Nature Reviews Immunology. 2014;14(6):377-91.
24. He Q, Jiang X, Zhou X, Weng J. Targeting cancers through TCR-peptide/MHC interactions. Journal of Hematology & Oncology. 2019;12(1):139.
25. Wo J, Zhang F, Li Z, Sun C, Zhang W, Sun G. The Role of Gamma-Delta T Cells in Diseases of the Central Nervous System. Frontiers in Immunology. 2020;11.
26. Gunsilius E, Gastl G, Petzer AL. Hematopoietic stem cells. Biomedicine & Pharmacotherapy. 2001;55(4):186-94.
27. Wolf AA, Yáñez A, Barman PK, Goodridge HS. The Ontogeny of Monocyte Subsets. Frontiers in Immunology. 2019;10.

28. Brignole C, Pastorino F, Perri P, Amoroso L, Bensa V, Calarco E, et al. Bone Marrow Environment in Metastatic Neuroblastoma. Cancers [Internet]. 2021; 13(10).
29. Kondo M. Lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors. Immunological Reviews. 2010;238(1):37-46.
30. Banerjee S, Parasramka MA, Paruthy SB. Chapter 15 - Garcinol: Preclinical Perspective Underpinning Chemo- and Radiosensitization of Cancer. In: Bharti AC, Aggarwal BB, editors. Role of Nutraceuticals in Cancer Chemosensitization. 2: Academic Press; 2018. p. 297-324.
31. Rankin A. Latest statistics released on number of blood cancer cases in England (25th Jan 2018): Blood cancer UK

2018 [Available from: https://bloodcancer.org.uk/news/latest-statistics-released-number-blood-cancer-cases-england/.
32. Randolph BV, Ciurea SO. What the Intensivist Needs to Know About Hematopoietic Stem Cell Transplantation? In: Nates JL, Price KJ, editors. Oncologic Critical Care. Cham: Springer International Publishing; 2020. p. 1531-46.
33. Hamilton BK, Copelan EA. Concise Review: The Role of Hematopoietic Stem Cell Transplantation in the Treatment of Acute Myeloid Leukemia. STEM CELLS. 2012;30(8):1581-6.
34. Copelan EA. Hematopoietic Stem-Cell Transplantation. New England Journal of Medicine. 2006;354(17):1813-26.
35. Verma K, Croft W, Pearce H, Zuo J, Stephens C, Nunnick J, et al. Early expression of CD94 and loss of CD96 on CD8+ T cells after allogeneic SCT is predictive of subsequent relapse and survival. Haematologica. 2022.
36. Nagler A, Shimoni A. Conditioning. In: Carreras E, Dufour C, Mohty M, Kröger N, editors. The EBMT Handbook: Hematopoietic Stem Cell Transplantation and Cellular Therapies. Cham: Springer International Publishing; 2019. p. 99-107.
37. Bleakley M, Riddell SR. Molecules and mechanisms of the graft-versus-leukaemia effect. Nature Reviews Cancer. 2004;4(5):371-80.
38. Chang Y-J, Zhao X-Y, Huang X-J. Strategies for Enhancing and Preserving Anti-leukemia Effects Without Aggravating Graft-Versus-Host Disease. Frontiers in Immunology. 2018;9.
39. Hsu KC, Keever-Taylor CA, Wilton A, Pinto C, Heller G, Arkun K, et al. Improved outcome in HLA-identical sibling hematopoietic stem-cell transplantation for acute myelogenous leukemia predicted by KIR and HLA genotypes. Blood. 2005;105(12):4878-84.
40. Dermime S, Mavroudis D, Jiang YZ, Hensel N, Molldrem J, Barrett AJ. Immune escape from a graft-versus-leukemia effect may play a role in the relapse of myeloid leukemias following allogeneic bone marrow transplantation. Bone Marrow Transplantation. 1997;19(10):989-99.
41. Osawa M, Hanada K-i, Hamada H, Nakauchi H. Long-Term Lymphohematopoietic Reconstitution by a Single CD34-Low/Negative Hematopoietic Stem Cell. Science. 1996;273(5272):242-5.
42. Lorenz E, Congdon C, Uphoff D. Modification of Acute Irradiation Injury in Mice and Guinea-Pigs by Bone Marrow Injections. Radiology. 1952;58(6):863-77.
43. Storb R, Epstein RB, Graham TC, Thomas ED. METHOTREXATE REGIMENS FOR CONTROL OF GRAFT-VERSUS-HOST DISEASE IN DOGS WITH ALLOGENEIC MARROW GRAFTS. Transplantation. 1970;9(3).
44. Loke J, Malladi R, Moss P, Craddock C. The role of allogeneic stem cell transplantation in the management of acute myeloid leukaemia: a triumph of hope and experience. British Journal of Haematology. 2020;188(1):129-46.
45. Mattsson J, Ringdén O, Storb R. Graft Failure after Allogeneic Hematopoietic Cell Transplantation. Biology of Blood and Marrow Transplantation. 2008;14(1, Supplement):165-70.
46. McSweeney PA, Niederwieser D, Shizuru JA, Sandmaier BM, Molina AJ, Maloney DG, et al. Hematopoietic cell transplantation in older patients with hematologic malignancies: replacing high-dose cytotoxic therapy with graft-versus-tumor effects. Blood. 2001;97(11):3390-400.
47. Horowitz M, Schreiber H, Elder A, Heidenreich O, Vormoor J, Toffalori C, et al. Epidemiology and biology of relapse after stem cell transplantation. Bone Marrow Transplantation. 2018;53(11):1379-89.
48. Appelbaum FR. The current status of hematopoietic cell transplantation. Annual Review of Medicine. 2003;54:491.
49. Malagola M, Greco R, Peccatori J, Isidori A, Romee R, Mohty M, et al. Editorial: Strengths and Challenges of Allo-SCT in the Modern Era. Frontiers in Oncology. 2022;12.
50. Styczyński J, Tridello G, Koster L, Iacobelli S, van Biezen A, van der Werf S, et al. Death after hematopoietic stem cell transplantation: changes over calendar year time, infections and associated factors. Bone Marrow Transplantation. 2020;55(1):126-36.

51. Atilla E, Atilla P, Demirer T. A Review of Myeloablative vs Reduced Intensity/Non-Myeloablative Regimens in Allogeneic Hematopoietic Stem Cell Transplantations. Balkan Medical Journal. 2017;34:1-9.
52. Wang C, Xiong C, Hsu Y-C, Wang X, Chen L. Human leukocyte antigen (HLA) and cancer immunotherapy: HLA-dependent and -independent adoptive immunotherapies. Annals of Blood. 2020;5.
53. Ferrara JLM, Levine JE, Reddy P, Holler E. Graft-versus-host disease. The Lancet. 2009;373(9674):1550-61.
54. Vago L, Perna SK, Zanussi M, Mazzi B, Barlassina C, Stanghellini MTL, et al. Loss of Mismatched HLA in Leukemia after Stem-Cell Transplantation. New England Journal of Medicine. 2009;361(5):478-88.
55. Barrett A, Battiwalla M. Relapse after allogeneic stem cell transplantation. Expert review of hematology. 2010;3:429-41.
56. Jang JE, Hwang DY, Chung H, Kim S-J, Eom J-I, Jeung H-K, et al. Early Cytomegalovirus Reactivation and Expansion of CD56brightCD16dim/−DNAM1+ Natural Killer Cells Are Associated with Antileukemia Effect after Haploidentical Stem Cell Transplantation in Acute Leukemia. Biology of Blood and Marrow Transplantation. 2019;25(10):2070-8.
57. Kenyon M, Babic A. The European Blood and Marrow Transplantation Textbook for Nurses: Under the Auspices of EBMT2018.
58. Segal E, Martens M, Wang H-L, Brazauskas R, Weisdorf D, Sandmaier B, et al. Comparing outcomes of matched related donor and matched unrelated donor hematopoietic cell transplants in adults with B-Cell acute lymphoblastic leukemia: Allogeneic Transplant Donors in ALL. Cancer. 2017;123.
59. Ayuk F, Balduzzi A. Donor selection for adults and pediatrics. 2018. p. 87-97.
60. Shaw BE, Ball L, Beksac M, Bengtsson M, Confer D, Diler S, et al. Donor safety: the role of the WMDA in ensuring the safety of volunteer unrelated donors: clinical and ethical considerations. Bone Marrow Transplantation. 2010;45(5):832-8.
61. Saber W, Opie S, Rizzo J, Zhang M-J, Horowitz M, Schriber J. Outcomes after matched unrelated donor versus identical sibling hematopoietic cell transplantation in adults with acute myelogenous leukemia. Blood. 2012;119:3908-16.
62. Woolfrey A, Lee S, Gooley T, Malkki M, Martin P, Pagel J, et al. HLA-Allele Matched Unrelated Donors Compared to HLA-Matched Sibling Donors: Role of Cell Source and Disease Risk Category. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2010;16:1382-7.
63. Gupta V, Tallman M, He W, Logan B, Copelan E, Gale R, et al. Comparable survival after HLA-well-matched unrelated or matched sibling donor transplantation for acute myeloid leukemia in first remission with unfavorable cytogenetics at diagnosis. Blood. 2010;116:1839-48.
64. Richard C. Is unrelated donor or haploidentical hematopoietic transplantation preferred for patients with acute myeloid leukemia in remission? Haematologica. 2020;105(2):252-4.
65. Al Malki MM, Tsai N-C, Palmer J, Mokhtari S, Tsai W, Cao T, et al. Posttransplant cyclophosphamide as GVHD prophylaxis for peripheral blood stem cell HLA-mismatched unrelated donor transplant. Blood Advances. 2021;5(12):2650-9.
66. Sweeney C, Vyas P. The Graft-Versus-Leukemia Effect in AML. Frontiers in Oncology. 2019;9.
67. Baumeister SH, Rambaldi B, Shapiro RM, Romee R. Key Aspects of the Immunobiology of Haploidentical Hematopoietic Cell Transplantation. Frontiers in Immunology. 2020;11.
68. Aversa F, Pierini A, Ruggeri L, Martelli MF, Velardi A. The Evolution of T Cell Depleted Haploidentical Transplantation. Frontiers in Immunology. 2019;10.
69. Aversa F. T-cell depletion: from positive selection to negative depletion in adult patients. Bone Marrow Transplantation. 2015;50:S11-S3.
70. Zeiser R, Blazar BR. Acute Graft-versus-Host Disease - Biologic Process, Prevention, and Therapy. The New England journal of medicine. 2017;377 22:2167-79.
71. Petersdorf EW. Role of major histocompatibility complex variation in graft-versus-host disease after hematopoietic cell transplantation. F1000Research. 2017;6.
72. Kanda J. Effect of HLA mismatch on acute graft-versus-host disease. International Journal of Hematology. 2013;98:300-8.
73. Holler E, Greinix H, Zeiser R. Acute Graft-Versus-Host Disease. In: Carreras E, Dufour C, Mohty M, Kröger N, editors. The EBMT Handbook: Hematopoietic Stem Cell Transplantation and Cellular Therapies. Cham: Springer International Publishing; 2019. p. 323-30.
74. Choi SW, Reddy P. Current and emerging strategies for the prevention of graft-versus-host disease. Nature Reviews Clinical Oncology. 2014;11:536-47.
75. Gooptu M, Antin JH. GVHD Prophylaxis 2020. Frontiers in Immunology. 2021;12.

76. Wolff D, Lawitschka A. Chronic Graft-Versus-Host Disease. In: Carreras E, Dufour C, Mohty M, Kröger N, editors. The EBMT Handbook: Hematopoietic Stem Cell Transplantation and Cellular Therapies. Cham: Springer International Publishing; 2019. p. 331-45.
77. Cooke KR, Luznik L, Sarantopoulos S, Hakim FT, Jagasia MH, Fowler DH, et al. The Biology of Chronic Graft-versus-Host Disease: A Task Force Report from the National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2017;23 2:211-34.
78. Baird K, Cooke KR, Schultz KR. Chronic graft-versus-host disease (GVHD) in children. Pediatric clinics of North America. 2010;57 1:297-322.
79. Fukushi N, Arase H, Wang B, Ogasawara K, Gotohda T, Good RA, et al. Thymus: a direct target tissue in graft-versus-host reaction after allogeneic bone marrow transplantation that results in abrogation of induction of self-tolerance. Proceedings of the National Academy of Sciences. 1990;87(16):6301-5.
80. Blazar BR, MacDonald KPA, Hill GR. Immune regulatory cell infusion for graft-versus-host disease prevention and therapy. Blood. 2018;131 24:2651-60.
81. Safarini O, Keshavamurthy C, Patel P. Calcineurin Inhibitors. 2020.
82. Hamilton BK. Current approaches to prevent and treat GVHD after allogeneic stem cell transplantation. Hematology. 2018;2018(1):228-35.
83. Zheng X, Tian Z. Which is better, HLA-matched sibling or haploidentical transplantation? Cellular & Molecular Immunology. 2021;18(5):1347-.
84. Kuzmina Z, Eder S, Böhm A, Pernicka E, Vormittag L, Kalhs P, et al. Significantly worse survival of patients with NIH-defined chronic graft-versus-host disease and thrombocytopenia or progressive onset type: results of a prospective study. Leukemia. 2012;26:746-56.
85. Hutt D. Engraftment, Graft Failure, and Rejection. 2018. p. 259-70.
86. Seita J, Weissman IL. Hematopoietic stem cell: self-renewal versus differentiation. Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2010;2.
87. Valcárcel D, Sureda A. Graft Failure. The EBMT Handbook. 2018.
88. Ayas M, Eapen M, Le-Rademacher JG, Carreras J, Abdel-Azim H, Alter BP, et al. Second Allogeneic Hematopoietic Cell Transplantation for Patients with Fanconi Anemia and Bone Marrow Failure. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2015;21 10:1790-5.
89. Wolff S. Second hematopoietic stem cell transplantation for the treatment of graft failure, graft rejection or relapse after allogeneic transplantation. Bone Marrow Transplantation. 2002;29:545-52.
90. Pham SM, Mitruka SN, Youm W, Li S, Kawaharada N, Yousem SA, et al. Mixed hematopoietic chimerism induces donor-specific tolerance for lung allografts in rodents. American journal of respiratory and critical care medicine. 1999;159 1:199-205.
91. Nikolic B, Sykes M. Mixed hematopoietic chimerism and transplantation tolerance. Immunologic Research. 1997;16:217-28.
92. Bader P. Documentation of Engraftment and Chimerism After HSCT. The EBMT Handbook. 2018.
93. Blouin AG, Ye F, Williams JD, Askar M. A practical guide to chimerism analysis: Review of the literature and testing practices worldwide. Human immunology. 2021.
94. Champlin RE, Schmitz N, Horowitz MM, Chapuis B, Chopra R, Cornelissen JJ, et al. Blood stem cells compared with bone marrow as a source of hematopoietic cells for allogeneic transplantation. IBMTR Histocompatibility and Stem Cell Sources Working Committee and the European Group for Blood and Marrow Transplantation (EBMT). Blood. 2000;95 12:3702-9.
95. Ruggeri A, Labopin M, Sormani MP, Sanz GF, Sanz J, Volt F, et al. Engraftment kinetics and graft failure after single umbilical cord blood transplantation using a myeloablative conditioning regimen. Haematologica. 2014;99:1509 - 15.
96. Bensinger WI. Allogeneic transplantation: peripheral blood vs. bone marrow. Current Opinion in Oncology. 2012;24:191–6.
97. Anasetti CM, Logan BR, Lee SJ, Waller EK, Weisdorf DJ, Wingard JR, et al. Peripheral-blood stem cells versus bone marrow from unrelated donors. The New England journal of medicine. 2012;367 16:1487-96.
98. Kekre N, Antin JH. Hematopoietic stem cell transplantation donor sources in the 21st century: choosing the ideal donor when a perfect match does not exist. Blood. 2014;124 3:334-43.
99. Du W, Liu W, Mizukawa BE, Shang XZ, Sipple J, Wunderlich M, et al. A non-myeloablative conditioning approach for long-term engraftment of human and mouse hematopoietic stem cells. Leukemia. 2018;32:2041-6.
100. Zulu S, Kenyon MO, editors. Principles of Conditioning Therapy and Cell Infusion2018.

101. Sabloff M, Tisseverasinghe S, Babadağli ME, Samant RS. Total Body Irradiation for Hematopoietic Stem Cell Transplantation: What Can We Agree on? Current Oncology. 2021;28:903 - 17.
102. Willemsen L, Jol-van der Zijde CM, Admiraal R, Putter H, Jansen-Hoogendijk AM, Ostaijen-ten Dam MM, et al. Impact of serotherapy on immune reconstitution and survival outcomes after stem cell transplantations in children: thymoglobulin versus alemtuzumab. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2015;21 3:473-82.
103. Hoegh-Petersen M, Amin MA, Liu Y, Ugarte-Torres A, Williamson TS, Podgorny PJ, et al. Anti-thymocyte globulins capable of binding to T and B cells reduce graft-vs-host disease without increasing relapse. Bone Marrow Transplantation. 2013;48:105-14.
104. Stauch D, Dernier A, Sarmiento Marchese E, Kunert K, Volk H-D, Pratschke J, et al. Targeting of Natural Killer Cells by Rabbit Antithymocyte Globulin and Campath-1H: Similar Effects Independent of Specificity. PLoS ONE. 2009;4.
105. Shah AJ, Kapoor N, Crooks GM, Weinberg K, Azim HA-, Killen R, et al. The effects of Campath 1H upon graft-versus-host disease, infection, relapse, and immune reconstitution in recipients of pediatric unrelated transplants. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2007;13 5:584-93.
106. Bacigalupo A, Lamparelli T, Bruzzi P, Guidi S, Alessandrino PE, Di Bartolomeo P, et al. Antithymocyte globulin for graft-versus-host disease prophylaxis in transplants from unrelated donors: 2 randomized studies from Gruppo Italiano Trapianti Midollo Osseo (GITMO). Blood. 2001;98 10:2942-7.
107. Krok-Schoen JL, Fisher JL, Stephens JA, Mims AS, Ayyappan S, Woyach JA, et al. Incidence and survival of hematological cancers among adults ages ≥75 years. Cancer Medicine. 2018;7:3425 - 33.
108. Bacigalupo A, Ballen K, Rizzo D, Giralt SA, Lazarus HM, Ho VT, et al. Defining the intensity of conditioning regimens: working definitions. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2009;15 12:1628-33.
109. Mielcarek M, Martin PJ, Leisenring WM, Flowers ME, Maloney DG, Sandmaier BM, et al. Graft-versus-host disease after nonmyeloablative versus conventional hematopoietic stem cell transplantation. Blood. 2003;102 2:756-62.
110. Locatelli F, Lucarelli B, Merli P. Current and future approaches to treat graft failure after allogeneic hematopoietic stem cell transplantation. Expert Opinion on Pharmacotherapy. 2014;15:23 - 36.
111. Giralt SA, Estey EH, Albitar M, van Besien K, Rondón G, Anderlini PN, et al. Engraftment of allogeneic hematopoietic progenitor cells with purine analog-containing chemotherapy: harnessing graft-versus-leukemia without myeloablative therapy. Blood. 1997;89 12:4531-6.
112. Greaves M, Ford AM, editors. Chromosome Translocations, Cancer Initiation and Clonal Evolution2015.
113. Vago L, Toffalori C, Ciceri F, Fleischhauer K. Genomic loss of mismatched human leukocyte antigen and leukemia immune escape from haploidentical graft-versus-leukemia. Seminars in oncology. 2012;39 6:707-15.
114. Reshef R, Hexner EO, Loren AW, Frey NV, Stadtmauer EA, Luger SM, et al. Early donor chimerism levels predict relapse and survival after allogeneic stem cell transplantation with reduced-intensity conditioning. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2014;20 11:1758-66.
115. Chen S, Zeiser R. Novel Biomarkers for Outcome After Allogeneic Hematopoietic Stem Cell Transplantation. Frontiers in Immunology. 2020;11.
116. Boatsman EE, Fu CH, Song SX, Moore TB. Graft-Versus-Leukemia Effect on Infant Lymphoblastic Leukemia Relapsed After Sibling Hematopoietic Stem Cell Transplantation. Journal of Pediatric Hematology/Oncology. 2010;32:e57-e60.
117. Rezvani K, Mielke S, Ahmadzadeh M, Kilical Y, Savani BN, Zeilah J, et al. High donor FOXP3-positive regulatory T-cell (Treg) content is associated with a low risk of GVHD following HLA-matched allogeneic SCT. Blood. 2006;108 4:1291-7.
118. Maggs L, Kinsella FAM, Chan YLT, Eldershaw SE, Murray D, Nunnick J, et al. The number of CD56dim NK cells in the graft has a major impact on risk of disease relapse following allo-HSCT. Blood advances. 2017;1 19:1589-97.
119. Hübel K. Mobilization and Collection of HSC. The EBMT Handbook. 2018.
120. Wuchter P. Processing, Cryopreserving and Controlling the Quality of HSCs. The EBMT Handbook. 2018.
121. Schumm MA, Lang P, Handgretinger R. Graft Manipulation. The EBMT Handbook. 2018.
122. Zaucha JM, Gooley TA, Bensinger WI, Heimfeld S, Chauncey TR, Zaucha R, et al. CD34 cell dose in granulocyte colony-stimulating factor-mobilized peripheral blood mononuclear cell grafts affects engraftment kinetics and development of extensive chronic graft-versus-host disease after human leukocyte antigen-identical sibling transplantation. Blood. 2001;98 12:3221-7.

123. Gauntner TD, Brunstein CG, Cao Q, Weisdorf D, Warlick ED, Jurdi NE, et al. Association of CD34 cell dose with 5-year overall survival after peripheral blood allogeneic hematopoietic cell transplantation in adults with hematologic malignancies. Transplantation and cellular therapy. 2021.
124. Pasquini MC, Devine SM, Mendizabal AM, Baden LR, Wingard JR, Lazarus HM, et al. Comparative outcomes of donor graft CD34+ selection and immune suppressive therapy as graft-versus-host disease prophylaxis for patients with acute myeloid leukemia in complete remission undergoing HLA-matched sibling allogeneic hematopoietic cell transplantation. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2012;30 26:3194-201.
125. Lang P, Handgretinger R, Niethammer D, Schlegel PG, Schumm MA, Greil J, et al. Transplantation of highly purified CD34+ progenitor cells from unrelated donors in pediatric leukemia. Blood. 2003;101 4:1630-6.
126. Remberger M, Grønvold BC, Ali M, Mattsson J, Egeland T, Lundin KU, et al. The CD34+ Cell Dose Matters in Hematopoietic Stem Cell Transplantation with Peripheral Blood Stem Cells from Sibling Donors. Clinical Hematology International. 2020;2:74 - 81.
127. Gómez-Almaguer D, Gómez-Peña Á, Jaime-Pérez JC, Gomez-Guijosa M-A, Cantú-Rodríguez OG, Gutiérrez-Aguirre H, et al. Higher doses of CD34+ progenitors are associated with improved overall survival without increasing GVHD in reduced intensity conditioning allogeneic transplant recipients with clinically advanced disease. Journal of Clinical Apheresis. 2013;28.
128. Yokoyama Y, Maie K, Fukuda T, Uchida N, Mukae J, Sawa M, et al. A high CD34+ cell dose is associated with better disease-free survival in patients with low-risk diseases undergoing peripheral blood stem cell transplantation from HLA-matched related donors. Bone Marrow Transplantation. 2020;55:1726 - 35.
129. Martin PS, Li S, Nikiforow S, Alyea EP, Antin JH, Armand P, et al. Infused total nucleated cell dose is a better predictor of transplant outcomes than CD34+ cell number in reduced-intensity mobilized peripheral blood allogeneic hematopoietic cell transplantation. Haematologica. 2016;101:499 - 505.
130. Pérez-Simón JA, Díez-Campelo M, Martino R, Sureda A, Caballero D, Cañizo Cd, et al. Impact of CD34+ cell dose on the outcome of patients undergoing reduced-intensity-conditioning allogeneic peripheral blood stem cell transplantation. Blood. 2003;102 3:1108-13.
131. Gorin NC, Labopin M, Boiron JM, Theorin N, Littlewood TJ, Slavin S, et al. Results of genoidentical hemopoietic stem cell transplantation with reduced intensity conditioning for acute myelocytic leukemia: higher doses of stem cells infused benefit patients receiving transplants in second remission or beyond--the Acute Leukemia Working Party of the European Cooperative Grou. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2006;24 24:3959-66.
132. Mohty M, Bilger K, Jourdan E, Kuentz M, Michallet M, Bourhis J, et al. Higher doses of CD34+ peripheral blood stem cells are associated with increased mortality from chronic graft-versus-host disease after allogeneic HLA-identical sibling transplantation. Leukemia. 2003;17:869-75.
133. Busca A, Aversa F. In-vivo or ex-vivo T cell depletion or both to prevent graft-versus-host disease after hematopoietic stem cell transplantation. Expert Opinion on Biological Therapy. 2017;17:1401 - 15.
134. Finke J, Bethge WA, Schmoor C, Ottinger HD, Stelljes M, Zander AR, et al. Standard graft-versus-host disease prophylaxis with or without anti-T-cell globulin in haematopoietic cell transplantation from matched unrelated donors: a randomised, open-label, multicentre phase 3 trial. The Lancet Oncology. 2009;10 9:855-64.
135. Mohty M, Malard F. Antithymocyte Globulin for Graft-Versus-Host Disease Prophylaxis After Allogeneic Hematopoietic Stem-Cell Transplantation. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2017;35 36:3993-5.
136. Biernacki MA, Sheth V, Bleakley M. T cell optimization for graft-versus-leukemia responses. JCI insight. 2020;5 9.
137. Handgretinger R, Schilbach K. The potential role of γδ T cells after allogeneic HCT for leukemia. Blood. 2018;131 10:1063-72.
138. Chaleff S, Otto M, Barfield RC, Leimig T, Iyengar R, Martín J, et al. A large-scale method for the selective depletion of αβ T lymphocytes from PBSC for allogeneic transplantation. Cytotherapy. 2007;9:746-54.
139. Lang P, Feuchtinger T, Teltschik HM, Schwinger W, Schlegel P, Pfeiffer MM, et al. Improved immune recovery after transplantation of TCRαβ/CD19-depleted allografts from haploidentical donors in pediatric patients. Bone Marrow Transplantation. 2015;50:S6-S10.
140. Li Pira G, Malaspina D, Girolami E, Biagini S, Cicchetti E, Conflitti G, et al. Selective Depletion of αβ T Cells and B Cells for Human Leukocyte Antigen-Haploidentical Hematopoietic Stem Cell Transplantation. A Three-Year Follow-Up of Procedure Efficiency. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2016;22 11:2056-64.

141. Enok Bonong PR, Zahreddine M, Buteau C, Duval M, Laporte L, Lacroix JR, et al. Factors Associated with Post-Transplant Active Epstein-Barr Virus Infection and Lymphoproliferative Disease in Hematopoietic Stem Cell Transplant Recipients: A Systematic Review and Meta-Analysis. Vaccines. 2021;9.
142. Waller EK, Rosenthal H, Jones TW, Peel J, Lonial S, Langston AA, et al. Larger numbers of CD4bright dendritic cells in donor bone marrow are associated with increased relapse after allogeneic bone marrow transplantation. Blood. 2001;97:2948-56.
143. Xu Z, Huang X-j. Optimizing allogeneic grafts in hematopoietic stem cell transplantation. Stem Cells Translational Medicine. 2021;10:S41 - S7.
144. Holtan SG, DeFor TE, Lazaryan A, Bejanyan N, Arora M, Brunstein CG, et al. Composite end point of graft-versus-host disease-free, relapse-free survival after allogeneic hematopoietic cell transplantation. Blood. 2015;125 8:1333-8.
145. Iacobelli S, Wreede LCd. Statistical Methods in HSCT and Cellular Therapies. The EBMT Handbook. 2018.
146. McDonald GB, Sandmaier BM, Mielcarek M, Sorror ML, Pergam SA, Cheng G-S, et al. Survival, Nonrelapse Mortality, and Relapse-Related Mortality After Allogeneic Hematopoietic Cell Transplantation: Comparing 2003–2007 Versus 2013–2017 Cohorts. Annals of Internal Medicine. 2020;172:229 - 39.
147. Valderas JM, Starfield BH, Sibbald B, Salisbury C, Roland M. Defining Comorbidity: Implications for Understanding Health and Health Services. The Annals of Family Medicine. 2009;7:357 - 63.
148. Sorror ML, Storer BE, Sandmaier BM, Maloney DG, Chauncey TR, Langston AA, et al. Hematopoietic cell transplantation-comorbidity index and Karnofsky performance status are independent predictors of morbidity and mortality after allogeneic nonmyeloablative hematopoietic cell transplantation. Cancer. 2008;112.
149. Galgano L, Hutt D, editors. HSCT: How Does It Work?2018.
150. Piccirillo JF, Tierney RM, Costas I, Grove LJ, Spitznagel EL. Prognostic importance of comorbidity in a hospital-based cancer registry. JAMA. 2004;291 20:2441-7.
151. Duarte RF, Sánchez-Ortega I. HSCT in Elderly Patients. The EBMT Handbook. 2018.
152. Extermann M. Measuring comorbidity in older cancer patients. European journal of cancer. 2000;36 4:453-71.
153. Kollman C, Spellman SR, Zhang M-J, Hassebroek AM, Anasetti CM, Antin JH, et al. The effect of donor characteristics on survival after unrelated donor transplantation for hematologic malignancy. Blood. 2016;127 2:260-7.
154. Canaani J, Savani BN, Labopin M, Huang X-j, Ciceri F, Arcese W, et al. Donor age determines outcome in acute leukemia patients over 40 undergoing haploidentical hematopoietic cell transplantation. American Journal of Hematology. 2018;93:246 - 53.
155. Galamidi-Cohen E, Joyce A, Simantov R. Impact of Donor Age on Outcomes in Hematopoietic Stem Cell Transplantation. STEM CELLS Translational Medicine. 2020;9.
156. DeZern AE, Franklin CE, Tsai H-l, Imus P, Cooke KR, Varadhan R, et al. Relationship of donor age and relationship to outcomes of haploidentical transplantation with posttransplant cyclophosphamide. Blood advances. 2021;5 5:1360-8.
157. Nakasone H, Remberger M, Tian L, Brodin P, Sahaf B, Wu F, et al. Risks and benefits of sex-mismatched hematopoietic cell transplantation differ according to conditioning strategy. Haematologica. 2015;100:1477 - 85.
158. Randolph SS, Gooley TA, Warren EH, Appelbaum FR, Riddell SR. Female donors contribute to a selective graft-versus-leukemia effect in male recipients of HLA-matched, related hematopoietic stem cell transplants. Blood. 2004;103 1:347-52.
159. Hefazi M, Litzow MR, Hogan WJ, Gastineau DA, Jacob EK, Damlaj M, et al. ABO blood group incompatibility as an adverse risk factor for outcomes in patients with myelodysplastic syndromes and acute myeloid leukemia undergoing HLA-matched peripheral blood hematopoietic cell transplantation after reduced-intensity conditioning. Transfusion. 2016;56.
160. Worel N. ABO-Mismatched Allogeneic Hematopoietic Stem Cell Transplantation. Transfusion Medicine and Hemotherapy. 2015;43:3 - 12.
161. Gratwohl A. The EBMT risk score. Bone Marrow Transplantation. 2012;47:749-56.
162. Jakharia N, Howard DS, Riedel DJ. CMV Infection in Hematopoietic Stem Cell Transplantation: Prevention and Treatment Strategies. Current Treatment Options in Infectious Diseases. 2021;13:123 - 40.
163. Manicklal S, Emery VC, Lazzarotto T, Boppana SB, Gupta RK. The “Silent” Global Burden of Congenital Cytomegalovirus. Clinical Microbiology Reviews. 2013;26:86 - 102.
164. Styczyński J. Who Is the Patient at Risk of CMV Recurrence: A Review of the Current Scientific Evidence with a Focus on Hematopoietic Cell Transplantation. Infectious Diseases and Therapy. 2017;7:1 - 16.

165. Ljungman P, Brand R, Hoek J, de la Cámara R, Cordonnier C, Einsele H, et al. Donor cytomegalovirus status influences the outcome of allogeneic stem cell transplant: a study by the European group for blood and marrow transplantation. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2014;59 4:473-81.
166. Abello V, Rosales C, Rosales M, Figueroa J, Córdoba I, Molina AAB, et al. Validation of the EBMT Risk Score for Patients with Hematologic Malignancies Receiving Allogeneic Hematopoietic Stem Cell Transplant in a Transplant Center in Colombia over 20 Years. Blood. 2015;126:5542-.
167. Pitombeira BS, Paz AA, Pezzi A, Amorin B, Valim VdS, Laureano ÁM, et al. Validation of the EBMT Risk Score for South Brazilian Patients Submitted to Allogeneic Hematopoietic Stem Cell Transplantation. Bone Marrow Research. 2013;2013.
168. Sorror ML, Maris MB, Storb RF, Baron F, Sandmaier BM, Maloney DG, et al. Hematopoietic cell transplantation (HCT)-specific comorbidity index: a new tool for risk assessment before allogeneic HCT. Blood. 2005;106 8:2912-9.
169. Gratwohl A, Stern M, Brand R, Apperley JF, Baldomero H, de Witte TJM, et al. Risk score for outcome after allogeneic hematopoietic stem cell transplantation. Cancer. 2009;115.
170. Shouval R, Fein JA, Shouval A, Danylesko I, Shem-Tov N, Zlotnik M, et al. External validation and comparison of multiple prognostic scores in allogeneic hematopoietic stem cell transplantation. Blood advances. 2019;3 12:1881-90.
171. Elsawy M, Storer BE, Pulsipher MA, Maziarz RT, Bhatia S, Maris MB, et al. Multi-centre validation of the prognostic value of the haematopoietic cell transplantation- specific comorbidity index among recipient of allogeneic haematopoietic cell transplantation. British Journal of Haematology. 2015;170.
172. Cordes S, Mokhtari Z, Bartosova M, Mertlitz S, Riesner K, Shi Y, et al. Endothelial damage and dysfunction in acute graft-versus-host disease. Haematologica. 2020;106:2147 - 60.
173. Polipalli SK, Karra VK, Jindal AK, Puppala M, Singh P, Rawat K, et al. Cytogenetic Analysis for Suspected Chromosomal Abnormalities; A Five Years Experience. Journal of clinical and diagnostic research : JCDR. 2016;10 9:GC01-GC5.
174. Armand P, Gibson CJ, Cutler CS, Ho VT, Koreth J, Alyea EP, et al. A disease risk index for patients undergoing allogeneic stem cell transplantation. Blood. 2012;120 4:905-13.
175. Strimbu K, Tavel JA. What are biomarkers? Current Opinion in HIV and AIDS. 2010;5:463–6.
176. Carlomagno N, Incollingo P, Tammaro V, Peluso G, Rupealta N, Chiacchio G, et al. Diagnostic, Predictive, Prognostic, and Therapeutic Molecular Biomarkers in Third Millennium: A Breakthrough in Gastric Cancer. BioMed Research International. 2017;2017.
177. Leptak CL, Menetski JP, Wagner JA, Aubrecht J, Brady L, Brumfield MA, et al. What evidence do we need for biomarker qualification? Science Translational Medicine. 2017;9.
178. Malone FR, Leisenring WM, Storer BE, Lawler RL, Stern JM, Aker SN, et al. Prolonged anorexia and elevated plasma cytokine levels following myeloablative allogeneic hematopoietic cell transplant. Bone Marrow Transplantation. 2007;40:765-72.
179. McDonald GB, Tabellini L, Storer BE, Lawler RL, Martin PJ, Hansen JA. Plasma biomarkers of acute GVHD and nonrelapse mortality: predictive value of measurements before GVHD onset and treatment. Blood. 2015;126 1:113-20.
180. Ahmed SS, Wang XN, Norden J, Pearce KF, El-Gezawy EM, Atarod S, et al. Identification and validation of biomarkers associated with acute and chronic graft versus host disease. Bone Marrow Transplantation. 2015;50:1563-71.
181. Hansen JA, Hanash SM, Tabellini L, Baik C, Lawler RL, Grogan BM, et al. A novel soluble form of Tim-3 associated with severe graft-versus-host disease. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2013;19 9:1323-30.
182. August KJ, Chiang KY, Bostick RM, Flanders WD, Waller EK, Langston AA, et al. Biomarkers of immune activation to screen for severe, acute GVHD. Bone Marrow Transplantation. 2011;46:601-4.
183. Nakamura H, Komatsu K, Ayaki M, Kawamoto S, Murakami M, Uoshima N, et al. Serum levels of soluble IL-2 receptor, IL-12, IL-18, and IFN-γ in patients with acute graft-versus-host disease after allogeneic bone marrow transplantation. The Journal of Allergy and Clinical Immunology. 2000;106.
184. von Bahr Greenwood T, Palmkvist-Kaijser K, Chiang SC, Tesi B, Bryceson YT, Hjelmqvist H, et al. Elevated ferritin and soluble CD25 in critically ill are associated with parameters of (hyper)inflammation and lymphocyte cytotoxicity. Minerva anestesiologica. 2019.
185. Paczesny S, Krijanovski OI, Braun TM, Choi SW, Clouthier SG, Kuick R, et al. A biomarker panel for acute graft-versus-host disease. Blood. 2009;113 2:273-8.

186. David JM, Dominguez CL, Hamilton DH, Palena C. The IL-8/IL-8R Axis: A Double Agent in Tumor Immune Resistance. Vaccines. 2016;4.
187. Wajant H, Siegmund D. TNFR1 and TNFR2 in the Control of the Life and Death Balance of Macrophages. Frontiers in Cell and Developmental Biology. 2019;7.
188. Harris AC, Ferrara JLM, Braun TM, Holler E, Teshima T, Levine JE, et al. Plasma biomarkers of lower gastrointestinal and liver acute GVHD. Blood. 2012;119 12:2960-3.
189. Ferrara JLM, Harris AC, Greenson J, Braun TM, Holler E, Teshima T, et al. Regenerating islet-derived 3-alpha is a biomarker of gastrointestinal graft-versus-host disease. Blood. 2011;118 25:6702-8.
190. Rotz SJ, Dandoy CE, Davies SM. ST2 and Endothelial Injury as a Link between GVHD and Microangiopathy. The New England journal of medicine. 2017;376 12:1189-90.
191. Vander Lugt MT, Braun TM, Hanash SM, Ritz J, Ho VT, Antin JH, et al. ST2 as a marker for risk of therapy-resistant graft-versus-host disease and death. The New England journal of medicine. 2013;369 6:529-39.
192. Major-Monfried H, Renteria AS, Pawarode A, Reddy P, Ayuk FA, Holler E, et al. MAGIC biomarkers predict long-term outcomes for steroid-resistant acute GVHD. Blood. 2018;131 25:2846-55.
193. Hartwell MJ, Özbek U, Holler E, Renteria AS, Major-Monfried H, Reddy P, et al. An early-biomarker algorithm predicts lethal graft-versus-host disease and survival. JCI insight. 2017;2 3:e89798.
194. Fiore PF, Di Matteo S, Tumino N, Mariotti FR, Pietra G, Ottonello S, et al. Interleukin-15 and cancer: some solved and many unsolved questions. Journal for Immunotherapy of Cancer. 2020;8.
195. Wu X-j, Xie Y, Wang C, Han Y, Bao X-b, Ma S, et al. Prediction of acute GVHD and relapse by metabolic biomarkers after allogeneic hematopoietic stem cell transplantation. JCI insight. 2018;3 9.
196. Wäsch R, Bertz H, Kunzmann R, Finke J. Incidence of mixed chimaerism and clinical outcome in 101 patients after myeloablative conditioning regimens and allogeneic stem cell transplantation. British Journal of Haematology. 2000;109.
197. Lamba R, Abella E, Kukuruga DL, Klein JL, Savaşan S, Abidi MH, et al. Mixed hematopoietic chimerism at day 90 following allogenic myeloablative stem cell transplantation is a predictor of relapse and survival. Leukemia. 2004;18:1681-6.
198. Terwey TH, Hemmati PG, Nagy M, Pfeifer H, Gökbuget N, Brüggemann M, et al. Comparison of chimerism and minimal residual disease monitoring for relapse prediction after allogeneic stem cell transplantation for adult acute lymphoblastic leukemia. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2014;20 10:1522-9.
199. Bornhäuser M, Oelschlaegel U, Platzbecker U, Bug G, Lutterbeck K, Kiehl M, et al. Monitoring of donor chimerism in sorted CD34+ peripheral blood cells allows the sensitive detection of imminent relapse after allogeneic stem cell transplantation. Haematologica. 2009;94:1613 - 7.
200. Lee HC, Saliba RM, Rondón G, Chen JJ, Charafeddine Y, Medeiros LJ, et al. Mixed T Lymphocyte Chimerism after Allogeneic Hematopoietic Transplantation Is Predictive for Relapse of Acute Myeloid Leukemia and Myelodysplastic Syndromes. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2015;21 11:1948-54.
201. Walter RB, Gooley TA, Wood BL, Milano F, Fang M, Sorror ML, et al. Impact of pretransplantation minimal residual disease, as detected by multiparametric flow cytometry, on outcome of myeloablative hematopoietic cell transplantation for acute myeloid leukemia. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2011;29 9:1190-7.
202. Craddock CF, Hoelzer D, Komanduri KV. Current status and future clinical directions in the prevention and treatment of relapse following hematopoietic transplantation for acute myeloid and lymphoblastic leukemia. Bone Marrow Transplantation. 2018;54:6-16.
203. Schnittger S, Kern W, Tschulik C, Weiss T, Dicker F, Falini B, et al. Minimal residual disease levels assessed by NPM1 mutation-specific RQ-PCR provide important prognostic information in AML. Blood. 2009;114 11:2220-31.
204. Shayegi N, Kramer MS, Bornhäuser M, Schaich M, Schetelig J, Platzbecker U, et al. The level of residual disease based on mutant NPM1 is an independent prognostic factor for relapse and survival in AML. Blood. 2013;122 1:83-92.
205. Hrnjica B, Softic S, editors. The Survival Analysis for a Predictive Maintenance in Manufacturing. APMS; 2021.
206. Buehler S, Kaiser C, Jaeger F. The geographic determinants of bankruptcy: evidence from Switzerland. Small Business Economics. 2012;39:231-51.
207. Harrison T, Ansell J. Customer retention in the insurance industry: Using survival analysis to predict cross-selling opportunities. Journal of Financial Services Marketing. 2002;6:229-39.
208. Cox DR, editor Regression models and life tables (with discussion1972.

209. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958;53:457-81.
210. Gray RJ. A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. Annals of Statistics. 1988;16:1141-54.
211. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;94:496-509.
212. Prinja S, Gupta N, Verma RK. Censoring in Clinical Trials: Review of Survival Analysis Techniques. Indian Journal of Community Medicine : Official Publication of Indian Association of Preventive & Social Medicine. 2010;35:217 - 21.
213. Leung KM, Elashoff RM, Afifi AA. Censoring issues in survival analysis. Annual review of public health. 1997;18:83-104.
214. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016;133:601 - 9.
215. Zwiener I, Blettner M, Hommel G. Survival analysis: part 15 of a series on evaluation of scientific publications. Deutsches Arzteblatt international. 2011;108 10:163-9.
216. Andersen PK, Geskus RB, de Witte TJM, Putter H. Competing risks in epidemiology: possibilities and pitfalls. International journal of epidemiology. 2012;41 3:861-70.
217. Clark TG, Bradburn M, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer. 2003;89:232 - 8.
218. Schober P, Vetter TR. Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. Anesthesia and Analgesia. 2018;127:792 - 8.
219. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Controlled clinical trials. 1996;17 4:343-6.
220. Schuster NA, Hoogendijk EO, Kok AAL, Twisk JWR, Heymans MW. Ignoring competing events in the analysis of survival data may lead to biased results: a non-mathematical illustration of competing risk analysis. Journal of clinical epidemiology. 2020.
221. Feakins BG, editor Competing risks methodology in the evaluation of cardiovascular and cancer mortality as a consequence of albuminuria in type 2 diabetes2016.
222. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. American journal of epidemiology. 2009;170 2:244-56.
223. Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? Statistics in Medicine. 2012;31:1089 - 97.
224. Lima BA, Henriques T, Reis F, Alves H. Transplant Open Registry Initiative. Portuguese Journal of Nephrology & Hypertension. 2022.
225. Sanchez-Pinto LN, Luo Y, Churpek MM. Big Data and Data Science in Critical Care. Chest. 2018;154 5:1239-48.
226. Adderley NJ, Taverner T, Price MJ, Sainsbury C, Greenwood D, Chandan JS, et al. Development and external validation of prognostic models for COVID-19 to support risk stratification in secondary care. BMJ Open. 2022;12(1):e049506.
227. Greenwood D, Taverner T, Adderley NJ, Price M, Gokhale KM, Sainsbury C, et al. Machine learning of COVID-19 clinical data identifies population structures with therapeutic potential. iScience. 2022;25.
228. O'Donnell EA, Ernst DN, Hingorani R. Multiparameter Flow Cytometry: Advances in High Resolution Analysis. Immune Network. 2013;13:43 - 54.
229. Rosa SCD, Herzenberg LA, Herzenberg L, Roederer M. 11-color, 13-parameter flow cytometry: Identification of human naive T cells by phenotype, function, and T-cell receptor diversity. Nature Medicine. 2001;7:245-8.
230. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. Analytical chemistry. 2009;81 16:6813-22.
231. Iyer A, Hamers AAJ, Pillai AB. CyTOF® for the Masses. Frontiers in Immunology. 2022;13.
232. Gadalla R, Noamani B, Macleod BL, Dickson RJ, Guo M, Xu W, et al. Validation of CyTOF Against Flow Cytometry for Immunological Studies and Monitoring of Human Cancer Clinical Trials. Frontiers in Oncology. 2019;9.
233. Mair F, Hartmann FJ, Mrdjen D, Tosevski V, Krieg C, Becher B. The end of gating? An introduction to automated analysis of high dimensional cytometry data. European Journal of Immunology. 2016;46.
234. McKinnon K. Flow Cytometry: An Overview. Current Protocols in Immunology. 2018;120:5.1. - 5.1.11.

235. R Core Team. R: A language and environment for statistical computing. In: R Foundation for Statistical Computing, editor. Version 4022020.
236. Meur NL, Hahne F, Haaland B, editors. flowCore: data structures package for flow cytometry data2007.
237. Dowle M, Srinivasan A. data.table: Extension of `data.frame`. 2021.
238. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani R, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proceedings of the National Academy of Sciences. 2014;111:E2770 - E7.
239. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202.
240. Saxena AK, Prasad M, Gupta A, Bharill N, Patel OP, Tiwari A, et al. A review of clustering techniques and developments. Neurocomputing. 2017;267:664-81.
241. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir E-aD, Tadmor MD, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015;162:184-97.
242. Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry Part A. 2016;89.
243. Liu X, Song W, Wong BY, Zhang T, Yu S-y, Lin GN, et al. A comparison framework and guideline of clustering methods for mass cytometry data. Genome Biology. 2019;20.
244. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowledge and Information Systems. 2007;14:1-37.
245. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008;2008:P10008.
246. Girvan M, Newman MEJ. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America. 2002;99:7821 - 6.
247. Aumüller M, Bernhardsson E, Faithfull A, editors. ANN-Benchmarks: A Benchmarking Tool for Approximate Nearest Neighbor Algorithms. Similarity Search and Applications; 2017.
248. Li W, Zhang Y, Sun Y, Wang W, Zhang W, Lin X. Approximate Nearest Neighbor Search on High Dimensional Data — Experiments, Analyses, and Improvement. IEEE Transactions on Knowledge and Data Engineering. 2020;32:1475-88.
249. Fortunato S. Community detection in graphs. ArXiv. 2009;abs/0906.0612.
250. Stuchly J, Novak D. Rphenoannoy: R implementation of the phenograph algorithm - approximate KNN

modification, based on Rphenograph package. 2020.
251. Guyon I, Gunn SR, Nikravesh M, Zadeh LA, editors. Feature Extraction - Foundations and Applications. Feature Extraction; 2006.
252. Maaten Lvd, Hinton GE. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9:2579-605.
253. Amir E-aD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nature Biotechnology. 2013;31:545-52.
254. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018;3(29).
255. Moon KR, van Dijk D, Wang Z, Gigante SA, Burkhardt DB, Chen WS, et al. Visualizing structure and transitions in high-dimensional biological data. Nature Biotechnology. 2019;37:1482 - 92.
256. Moon KR, Dijk Dv, Wang Z, Chen WC, Hirn MJ, Coifman RR, et al. PHATE: A Dimensionality Reduction Method for Visualizing Trajectory Structures in High-Dimensional Biological Data. bioRxiv. 2017:120378.
257. Nanga S, Bawah AT, Acquaye B, Billa M-I, Baeta F, Odai NA, et al. Review of Dimension Reduction Methods. Journal of Data Analysis and Information Processing. 2021.
258. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology. 2019;37(1):38-44.
259. Krijthe JH. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation. 2022.
260. Chan D, Rao R, Huang F, Canny JF. T-SNE-CUDA: GPU-Accelerated T-SNE and its Applications to Modern Data. 2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD). 2018:330-8.
261. Watcham S, Kucinski I, Gottgens B. New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing. Blood. 2019;133 13:1415-26.
262. Kucinski I, Wilson NK, Hannah R, Kinston S, Cauchy P, Lenaerts A, et al. Interactions between lineage-associated transcription factors govern haematopoietic progenitor states. The EMBO Journal. 2020;39.

263. Abbas HA, Hao D, Tomczak K, Barrodia P, Im JS, Reville PK, et al. Single cell T cell landscape and T cell receptor repertoire profiling of AML in context of PD-1 blockade therapy. Nature Communications. 2021;12.
264. Grollemund V, Chat GL, Secchi-Buhour MS, Delbot F, Pradat-Peyre JF, Bede P, et al. Development and validation of a 1-year survival prognosis estimation model for Amyotrophic Lateral Sclerosis using manifold learning algorithm UMAP. Sci Rep. 2020;10(1):13378.
265. Weijler L, Kowarsch F, Wödlinger M, Reiter M, Maurer-Granofszky M, Schumich A, et al. UMAP Based Anomaly Detection for Minimal Residual Disease Quantification within Acute Myeloid Leukemia. Cancers. 2022;14.
266. Melville J. uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction. R package version 0192020.
267. Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. A review of UMAP in population genetics. Journal of Human Genetics. 2020;66:85-91.
268. Morcos MNF, Li C, Munz CM, Greco A, Dressel N, Reinhardt S, et al. Fate mapping of hematopoietic stem cells reveals two pathways of native thrombopoiesis. Nature Communications. 2022;13.
269. Connolly KA, Kuchroo M, Venkat A, Khatun A, Wang J, William I, et al. A reservoir of stem-like CD8+ T cells in the tumor-draining lymph node preserves the ongoing antitumor immune response. Science Immunology. 2021;6.
270. Srinivasan K, editor PHATE - Potential of Heat-Diffusion for Affinity-Based Transition Embedding [R package phateR version 1.0.4]2020.
271. Wickham H. Ggplot2: Elegant graphics for data analysis. 2 ed. Cham, Switzerland: Springer International Publishing; 2016 2016/6/8. 260 p.
272. Diggins KE, Ferrell PB, Irish J. Methods for discovery and characterization of cell subsets in high dimensional mass cytometry data. Methods. 2015;82:55-63.
273. Diggins KE, Greenplate AR, Leelatian N, Wogsland CE, Irish J. Characterizing cell subsets in heterogeneous tissues using marker enrichment modeling. Nature methods. 2017;14:275 - 8.
274. Daly JC, Cohen J, editors. Statistical Power Analysis For The Behavioral Sciences Revised Edition1987.
275. Diener MJ, editor Cohen's d2010.
276. Calin-Jageman RJ. esci: Effect Sizes and Confidence Intervals. 2022.
277. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nature Communications. 2018;10.
278. Rudra P, Baxter RM, Hsieh EWY, Ghosh D. Compositional Data Analysis using Kernels in mass cytometry data. Bioinformatics Advances. 2021;2.
279. Templ M, Filzmoser P, Reimann C, editors. Wahrscheinlichkeitstheorie Cluster analysis applied to regional geochemical data : Problems and possibilities2006.
280. Templ M, Hron K, Filzmoser P, editors. Compositional Data Analysis [R package robCompositions version 2.2.1]2020.
281. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J. 2016;8(1):289-317.
282. Fraley C, Raftery AE. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association. 2002;97:611 - 31.
283. Dempster AP, Laird NM, Rubin DB, editors. Maximum likelihood from incomplete data via the EM - algorithm plus discussions on the paper1977.
284. McLachlan GJ, Peel D, editors. Finite Mixture Models. Wiley Series in Probability and Statistics; 2000.
285. Schwarz G. Estimating the Dimension of a Model. Annals of Statistics. 1978;6:461-4.
286. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53-65.
287. Dudek A, editor Silhouette Index as Clustering Evaluation Tool2020.
288. Shutaywi M, Kachouie NN. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. Entropy. 2021;23.
289. Xu R, Wunsch DC. Clustering Algorithms in Biomedical Research: A Review. IEEE Reviews in Biomedical Engineering. 2010;3:120-54.
290. Luxburg Uv, Williamson RC, Guyon I, editors. Clustering: Science or Art? ICML Unsupervised and Transfer Learning; 2012.
291. Johnsson K, editor Structures in High-Dimensional Data: Intrinsic Dimension and Cluster Analysis2016.
292. Therneau TM, Grambsch PM. Modeling survival data : extending the Cox model

Modeling Survival Data: Extending the Cox Model. New York

New York, NY: Springer

Springer New York : Imprint: Springer; 2000.
293. Scrucca L, Santucci A, Aversa F. Competing risk analysis using R: an easy guide for clinicians. Bone Marrow Transplantation. 2007;40:381-7.
294. Ozenne B, Sorensen A, Scheike TH, Torp-Pedersen C, Gerds TA. riskRegression: Predicting the Risk of an Event using Cox Regression Models. R J. 2017;9:440.
295. Bland JM, Altman DG. The logrank test. BMJ : British Medical Journal. 2004;328:1073.
296. Kassambara A, Kosinski M, Biecek P, editors. Drawing Survival Curves using 'ggplot2' [R package survminer version 0.4.8]2020.
297. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011;30.
298. Iacobelli S. Suggestions on the use of statistical methodologies in studies of the European Group for Blood and Marrow Transplantation. Bone Marrow Transplantation. 2013;48:S1-S37.
299. Lau B, Lesko CR. Missingness in the Setting of Competing Risks: from Missing Values to Missing Potential Outcomes. Current Epidemiology Reports. 2018;5:153-9.
300. Rubin DB, editor Multiple imputation for nonresponse in surveys1987.
301. White IR, Daniel RM, Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. Computational Statistics & Data Analysis. 2010;54:2267 - 75.
302. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research. 2011;20.
303. Buuren Sv, Groothuis-Oudshoorn KGM. MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45:1-67.
304. Kleinke K. Multiple Imputation by Predictive Mean Matching When Sample Size Is Small. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences. 2018;14:3–15.
305. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81:515-26.
306. Li J, Scheike TH, Zhang M-J. Checking Fine and Gray subdistribution hazards model with cumulative sums of residuals. Lifetime Data Analysis. 2015;21:197-217.
307. McGregor DB, Palarea-Albaladejo J, Dall PM, Hron K, Chastin SFM. Cox regression survival analysis with compositional covariates: Application to modelling mortality risk from 24-h physical activity patterns. Statistical Methods in Medical Research. 2019;29:1447 - 65.
308. Greenacre MJ. Compositional Data Analysis. Annual Review of Statistics and Its Application. 2021.
309. Jones MC, Aitchison JW, editors. The Statistical Analysis of Compositional Data1982.
310. Greenacre MJ, Martínez-Álvaro M, Blasco A. Compositional Data Analysis of Microbiome and Any-Omics Datasets: A Validation of the Additive Logratio Transformation. Frontiers in Microbiology. 2021;12.
311. Lovell DR, Chua X-Y, McGrath A. Counts: an outstanding challenge for log-ratio analysis of compositional data in the molecular biosciences. NAR Genomics and Bioinformatics. 2020;2.
312. Quinn T, Erb I, Gloor GB, Notredame C, Richardson MF, Crowley TM. A field guide for the compositional analysis of any-omics data. GigaScience. 2018;8.
313. Pearson K. Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London.60:489 - 98.
314. Smith PF, Renner RM, Haslett S. Compositional data in neuroscience: If you've got it, log it! Journal of Neuroscience Methods. 2016;271:154-9.
315. Wang H-Y, Yang Q, Qin H, Zha H, editors. Dirichlet component analysis: feature extraction for compositional data. ICML '08; 2008.
316. Coenders G, Greenacre MJ. Three approaches to supervised learning for compositional data with pairwise logratios. ArXiv. 2022;abs/2111.08953.
317. Lu J, Shi P, Li H. Generalized linear models with linear constraints for microbiome compositional data. Biometrics. 2019;75.
318. Walmsley R, Chan S, Smith-Byrne K, Ramakrishnan R, Woodward M, Rahimi K, et al. Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. British Journal of Sports Medicine. 2021;56:1008 - 17.
319. Walmsley R. epicoda: Supports epidemiological analyses using compositional exposure variables. 2022.
320. Greenwood D. Fine-Gray subdistribution hazard model for compositional data. 0.9 ed. GitHub 2021.

321. Assarsson E, Lundberg M, Holmquist G, Björkesten J, Bucht Thorsen S, Ekman D, et al. Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. PLoS ONE. 2014;9.
322. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57(1):289-300.
323. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Family Medicine and Community Health. 2020;8.
324. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. Feature selection for high-dimensional data. Progress in Artificial Intelligence. 2015;5:65-75.
325. Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in medicine. 1997;16 4:385-95.
326. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the royal statistical society series b-methodological. 1996;58:267-88.
327. Simon N, Friedman JH, Hastie TJ, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of statistical software. 2011;39 5:1-13.
328. Heinze G, Wallisch C, Dunkler D. Variable selection – A review and recommendations for the practicing statistician. Biometrical Journal Biometrische Zeitschrift. 2018;60:431 - 49.
329. Akaike H, editor Information Theory and an Extension of the Maximum Likelihood Principle1973.
330. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. Wiley StatsRef: Statistics Reference Online. 2019.
331. Wang H, Li G. A Selective Review on Random Survival Forests for High Dimensional Data. Quantitative bio-science. 2017;36 2:85-96.
332. Pasin C, Moy RH, Reshef R, Yates AJ. Variable selection methods for predicting clinical outcomes following allogeneic hematopoietic cell transplantation. Scientific Reports. 2021;11.
333. Adom D, Rowan CM, Adeniyan T, Yang J, Paczesny S. Biomarkers for Allogeneic HCT Outcomes. Frontiers in Immunology. 2020;11.
334. Okamura H, Nakamae M, Koh S, Nanno S, Nakashima Y, Koh H, et al. Interactive web application for plotting personalized prognosis prediction curves in allogeneic hematopoietic cell transplantation using machine learning. Transplantation. 2019.
335. Ungaro RC, Hu L, Ji J, Nayar S, Kugathasan S, Denson LA, et al. Machine learning identifies novel blood protein predictors of penetrating and stricturing complications in newly diagnosed paediatric Crohn's disease. Alimentary Pharmacology & Therapeutics. 2020;53:281 - 90.
336. Drake I, Hindy G, Almgren P, Engström G, Nilsson J, Melander O, et al. Methodological considerations for identifying multiple plasma proteins associated with all-cause mortality in a population-based prospective cohort. Scientific Reports. 2021;11.
337. Fu Z, Parikh CR, Zhou B. Penalized variable selection in competing risks regression. Lifetime Data Analysis. 2017;23:353-76.
338. Kuk D, Varadhan R. Model selection in competing risks regression. Statistics in Medicine. 2013;32.
339. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau B. Random survival forests for competing risks. Biostatistics. 2014;15 4:757-73.
340. Venables WN, Ripley BD, editors. Modern Applied Statistics with S Fourth edition2002.
341. Mogensen UB, Ishwaran H, Gerds TA. Evaluating Random Forests for Survival Analysis using Prediction Error Curves. Journal of statistical software. 2012;50 11:1-23.
342. Ishwaran H, Lu M, Kogalur UB, editors. randomForestSRC: Getting Started with randomForestSRC Vignette2021.
343. Fabozzi FJ, Focardi S, Rachev ST, Arshanapalli BG, editors. The basics of financial econometrics : tools, concepts, and asset management applications2014.
344. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. American journal of epidemiology. 2007;165 6:710-8.
345. Danby RD, Zhang WJ, Medd PG, Littlewood TJ, Peniket A, Rocha V, et al. High proportions of regulatory T cells in PBSC grafts predict improved survival after allogeneic haematopoietic SCT. Bone Marrow Transplantation. 2016;51:110 - 8.
346. Peng L-s, Mao F-y, Zhao Y, Wang T-t, Chen N, Zhang J-y, et al. Altered phenotypic and functional characteristics of CD3+CD56+ NKT-like cells in human gastric cancer. Oncotarget. 2016;7:55222 - 30.
347. Kapellos TS, Bonaguro L, Gemünd ID, Reusch N, Saglam A, Hinkley ER, et al. Human Monocyte Subsets and Phenotypes in Major Chronic Inflammatory Diseases. Frontiers in Immunology. 2019;10.

348. Li H, Zhang G-X, Chen YH, Xu H, Fitzgerald DC, Zhao Z, et al. CD11c+CD11b+ Dendritic Cells Play an Important Role in Intravenous Tolerance and the Suppression of Experimental Autoimmune Encephalomyelitis1. The Journal of Immunology. 2008;181:2483 - 93.
349. Bozzano F, Marras F, De Maria A. Natural Killer Cell Development and Maturation Revisited: Possible Implications of a Novel Distinct Lin−CD34+DNAM-1brightCXCR4+ Cell Progenitor. Frontiers in Immunology. 2017;8.
350. Bozzano F, Marras F, Ascierto ML, Cantoni C, Cenderello G, Dentone C, et al. 'Emergency exit' of bone-marrow-resident CD34+DNAM-1brightCXCR4+-committed lymphoid precursors during chronic infection and inflammation. Nature Communications. 2015;6.
351. Bozzano F, Perrone C, Moretta L, De Maria A. NK Cell Precursors in Human Bone Marrow in Health and Inflammation. Frontiers in Immunology. 2019;10.
352. Challen GA, Boles NC, Lin K-YK, Goodell MA. Mouse hematopoietic stem cell identification and analysis. Cytometry Part A. 2009;75A.
353. Kumar R, Fossati V, Israel MA, Snoeck H-W. Lin−Sca1+Kit− Bone Marrow Cells Contain Early Lymphoid-Committed Precursors That Are Distinct from Common Lymphoid Progenitors1. The Journal of Immunology. 2008;181:7507 - 13.
354. Burel JG, Pomaznoy M, Lindestam Arlehamn CS, Weiskopf D, da Silva Antunes R, Jung Y, et al. Circulating T cell-monocyte complexes are markers of immune perturbations. eLife. 2019;8.
355. Greenacre MJ, editor Towards a pragmatic approach to compositional data analysis2017.
356. Sanz I, Wei C, Jenks SA, Cashman KS, Tipton CM, Woodruff MC, et al. Challenges and Opportunities for Consistent Classification of Human B Cell and Plasma Cell Populations. Frontiers in Immunology. 2019;10.
357. Amu S, Brisslert M. Phenotype and Function of CD25-Expressing B Lymphocytes Isolated from Human Umbilical Cord Blood. Clinical and Developmental Immunology. 2011;2011.
358. Pinto BF, Medeiros NI, Teixeira-Carvalho A, Elói-Santos SM, Fontes-Cal TCM, Rocha DA, et al. CD86 Expression by Monocytes Influences an Immunomodulatory Profile in Asymptomatic Patients with Chronic Chagas Disease. Frontiers in Immunology. 2018;9.
359. Ziegler-Heitbrock L, Ancuta P, Crowe SM, Dalod M, Grau V, Hart DNJ, et al. Nomenclature of monocytes and dendritic cells in blood. Blood. 2010;116 16:e74-80.
360. Marimuthu R, Francis H, Dervish S, Li SCH, Medbury HJ, Williams H. Characterization of Human Monocyte Subsets by Whole Blood Flow Cytometry Analysis. Journal of Visualized Experiments : JoVE. 2018.
361. Nagasawa T, Kobayashi H, Aramaki M, Kiji M, Oda S, Izumi Y. Expression of CD14, CD16 and CD45RA on monocytes from periodontitis patients. Journal of periodontal research. 2004;39 1:72-8.
362. Kling JC, Mack M, Körner H. The Absence of CCR7 Results in Dysregulated Monocyte Migration and Immunosuppression Facilitating Chronic Cutaneous Leishmaniasis. PLoS ONE. 2013;8.
363. Collin M, Bigley V. Human dendritic cell subsets: an update. Immunology. 2018;154:3 - 20.
364. Patente TA, Pinho MP, Oliveira AA, Evangelista GCM, Bergami-Santos PC, Barbuto JAM. Human Dendritic Cells: Their Heterogeneity and Clinical Application Potential in Cancer Immunotherapy. Frontiers in Immunology. 2019;9.
365. López-Botet M, Vilches C, Redondo-Pachón D, Muntasell A, Pupuleku A, Yélamos J, et al. Dual Role of Natural Killer Cells on Graft Rejection and Control of Cytomegalovirus Infection in Renal Transplantation. Frontiers in Immunology. 2017;8.
366. Poli A, Michel T, Thérésine M, Andrès E, Hentges Fo, Zimmer J. CD56bright natural killer (NK) cells: an important NK cell subset. Immunology. 2009;126.
367. Kondo M, Weissman IL, Akashi K. Identification of Clonogenic Common Lymphoid Progenitors in Mouse Bone Marrow. Cell. 1997;91:661-72.
368. Bühring H-JD, Kuçi S, Conze T, Rathke G, Bartolović K, Grünebach F, et al. CDCP1 Identifies a Broad Spectrum of Normal and Malignant Stem/Progenitor Cell Subsets of Hematopoietic and Nonhematopoietic Origin. STEM CELLS. 2004;22.
369. Schönrich G, Raftery MJ. The PD-1/PD-L1 Axis and Virus Infections: A Delicate Balance. Frontiers in Cellular and Infection Microbiology. 2019;9.
370. León-Flores AMd, Del Rio Estrada PM, Alvarez-García LX, Pitén-Isidro E, Reyes-Terán G. Increased levels of soluble co-stimulatory molecule PD-L1 (B7-H1) in the plasma of viraemic HIV-1+ individuals. Immunology letters. 2018;203:70-9.
371. Yoon S-i, Logsdon NJ, Sheikh FG, Donnelly RP, Walter MR. Conformational Changes Mediate Interleukin-10 Receptor 2 (IL-10R2) Binding to IL-10 and Assembly of the Signaling Complex*. Journal of Biological Chemistry. 2006;281:35088 - 96.

372. Sun Z, Ren Z, Yang K, Liu Z, Cao S, Deng S, et al. A next-generation tumor-targeting IL-2 preferentially promotes tumor-infiltrating CD8+ T-cell response and effective tumor control. Nature Communications. 2019;10.
373. Zhao X-Y, Zhao X-S, Wang Y-t, Chen Y-H, Xu L, Zhang X-h, et al. Prophylactic use of low-dose interleukin-2 and the clinical outcomes of hematopoietic stem cell transplantation: A randomized study. OncoImmunology. 2016;5.
374. Tie R, Li H, Cai S, Liang Z, Shan W, Wang B, et al. Interleukin-6 signaling regulates hematopoietic stem cell emergence. Experimental & Molecular Medicine. 2019;51.
375. Greco R, Lorentino F, Nitti R, Lupo Stanghellini MT, Giglio F, Clerici D, et al. Interleukin-6 as Biomarker for Acute GvHD and Survival After Allogeneic Transplant With Post-transplant Cyclophosphamide. Frontiers in Immunology. 2019;10.
376. Ramirez-Carrozzi V, Sambandam A, Luis E, Lin Z-m, Jeet S, Lesch J, et al. IL-17C regulates the innate immune function of epithelial cells in an autocrine manner. Nature Immunology. 2011;12:1159-66.
377. Justilien V, Regala RP, Tseng I-C, Walsh M, Batra J, Radisky ES, et al. Matrix Metalloproteinase-10 Is Required for Lung Cancer Stem Cell Maintenance, Tumor Initiation and Metastatic Potential. PLoS ONE. 2012;7.
378. Liu M, Guo S, Stiles JK. The emerging role of CXCL10 in cancer (Review). Oncology letters. 2011;2 4:583-9.
379. Moehler T, Neben K, Ho AD, Goldschmidt H. Angiogenesis in hematologic malignancies. Annals of Hematology. 2001;80:695-705.
380. Hwang J, Son K-N, Kim CW, Ko J, Na DS, Kwon BS, et al. Human CC chemokine CCL23, a ligand for CCR1, induces endothelial cell migration and promotes angiogenesis. Cytokine. 2005;30 5:254-63.
381. Yue X, Wu L, Hu W. The regulation of leukemia inhibitory factor. Cancer cell & microenvironment. 2015;2 3.
382. He S, Chu J, Vasu S, Deng Y-P, Yuan S, Zhang J, et al. FLT3L and plerixafor combination increases hematopoietic stem cell mobilization and leads to improved transplantation outcome. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2014;20 3:309-13.
383. Clave E, Busson M, Douay C, Peffault de Latour R, Berrou J, Rabian C, et al. Acute graft-versus-host disease transiently impairs thymic output in young patients after allogeneic hematopoietic stem cell transplantation. Blood. 2009;113 25:6477-84.
384. Shaw BE, Lee F, Krishnamurthy S, Byrne JL, Seedhouse CH, Mayor NP, et al. Caspase-8 polymorphisms result in reduced Alemtuzumab-induced T-cell apoptosis and worse survival after transplantation. Bone Marrow Transplantation. 2015;50:237-43.
385. OlinkProteomics. What is NPX? 2016 [Available from: https://olink.com/faq/what-is-npx/.
386. Raftery AE. Bayesian Model Selection in Social Research. Sociological Methodology. 1995;25:111.
387. Impola U, Larjo A, Salmenniemi U, Putkonen M, Itälä-Remes M, Partanen J. Graft Immune Cell Composition Associates with Clinical Outcome of Allogeneic Hematopoietic Stem Cell Transplantation in Patients with AML. Frontiers in Immunology. 2016;7.
388. Waller EK, Logan BR, Harris WAC, Devine SM, Porter DL, Mineishi S, et al. Improved survival after transplantation of more donor plasmacytoid dendritic or naïve T cells from unrelated-donor marrow grafts: results from BMTCTN 0201. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2014;32 22:2365-72.
389. Saraceni F, Shem-Tov N, Olivieri A, Nagler A. Mobilized peripheral blood grafts include more than hematopoietic stem cells: the immunological perspective. Bone Marrow Transplantation. 2015;50:886-91.
390. Podgorny PJ, Liu Y, Dharmani-Khan P, Pratt LM, Jamani K, Luider JM, et al. Immune cell subset counts associated with graft-versus-host disease. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2014;20 4:450-62.
391. Vasu S, Geyer SM, Bingman AK, Auletta JJ, Jaglowski SM, Elder PJ, et al. Granulocyte Colony-Stimulating Factor-Mobilized Allografts Contain Activated Immune Cell Subsets Associated with Risk of Acute and Chronic Graft-versus-Host Disease. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2016;22 4:658-68.
392. Reshef R. Peripheral blood stem cell grafts in allogeneic hematopoietic cell transplantation: It is not all about the CD34+ cell dose. Transfusion and apheresis science : official journal of the World Apheresis Association : official journal of the European Society for Haemapheresis. 2021:103081.
393. Yaman A, Çetiner S, Kibar F, Taşova Y, Seydaoglu G, Dündar İ. Reference Ranges of Lymphocyte Subsets of Healthy Adults in Turkey. Medical Principles and Practice. 2005;14:189 - 93.

394. Choi J, Lee SJ, Lee YA, Maeng HG, Lee JK, Kang YW. Reference Values for Peripheral Blood Lymphocyte Subsets in a Healthy Korean Population. Immune Network. 2014;14:289 - 95.
395. Pabst C, Schirutschke H, Ehninger G, Bornhäuser M, Platzbecker U. The Graft Content of Donor T Cells Expressing γδTCR+ and CD4+foxp3+ Predicts the Risk of Acute Graft versus Host Disease after Transplantation of Allogeneic Peripheral Blood Stem Cells from Unrelated Donors. Clinical Cancer Research. 2007;13:2916 - 22.
396. Gaballa A, Stikvoort A, Önfelt B, Mattsson J, Sundin M, Watz E, et al. T-cell frequencies of CD8+ γδ and CD27+ γδ cells in the stem cell graft predict the outcome after allogeneic hematopoietic cell transplantation. Bone Marrow Transplantation. 2019:1-13.
397. Mak TW, Saunders ME, Jett BD. Chapter 11 - NK, γδ T and NKT Cells. In: Mak TW, Saunders ME, Jett BD, editors. Primer to the Immune Response (Second Edition). Boston: Academic Cell; 2014. p. 247-68.
398. Morbach H, Eichhorn E-M, Liese JG, Girschick H. Reference values for B cell subpopulations from infancy to adulthood. Clinical & Experimental Immunology. 2010;162.
399. Rundgren IM, Ersvær E, Ahmed A, Ryningen A, Bruserud Ø. A Pilot Study of Circulating Monocyte Subsets in Patients Treated with Stem Cell Transplantation for High-Risk Hematological Malignancies. Medicina. 2020;56.
400. Wittenbecher F, Lesch S, Kolling S, Blau IW, Vuong LG, Borchert F, et al. Paired Donor and Recipient Immunophenotyping in Allogeneic Hematopoietic Stem Cell Transplantation: A Cellular Network Approach. Frontiers in Immunology. 2022;13.
401. Gonçalves MV, Yamamoto M, Kimura EYS, Colturato VAR, de Souza MP, Mauad MA, et al. Low Counts of Plasmacytoid Dendritic Cells after Engraftment Are Associated with High Early Mortality after Allogeneic Stem Cell Transplantation. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2015;21 7:1223-9.
402. Yu H, Tian Y, Wang Y, Mineishi S, Zhang Y. Dendritic Cell Regulation of Graft-Vs.-Host Disease: Immunostimulation and Tolerance. Frontiers in Immunology. 2019;10.
403. Banović T, Markey KA, Kuns RD, Olver SD, Raffelt NC, Don ALJ, et al. Graft-versus-Host Disease Prevents the Maturation of Plasmacytoid Dendritic Cells. The Journal of Immunology. 2009;182:912 - 20.
404. Talarn C, Urbano-Ispizua Á, Martino R, Batlle M, Fernández-Avilés F, Herrera C, et al. G-CSF increases the number of peripheral blood dendritic cells CD16+ and modifies the expression of the costimulatory molecule CD86+. Bone Marrow Transplantation. 2006;37:873-9.
405. Bozzano F, Della Chiesa M, Pelosi A, Antonini F, Ascierto ML, Del Zotto G, et al. HCMV-controlling NKG2C+ NK cells originate from novel circulating inflammatory precursors. The Journal of allergy and clinical immunology. 2021.
406. Quatrini L, Chiesa MD, Sivori S, Mingari MC, Pende D, Moretta L. Human NK cells, their receptors and function. European Journal of Immunology. 2021;51:1566 - 79.
407. Cichocki F, Grzywacz BJ, Miller JS. Human NK Cell Development: One Road or Many? Frontiers in Immunology. 2019;10.
408. Jiang H, Fu D, Bidgoli A, Paczesny S. T Cell Subsets in Graft Versus Host Disease and Graft Versus Tumor. Frontiers in Immunology. 2021;12.
409. Arruda LCM, Gaballa A, Uhlin M. Impact of γδ T cells on clinical outcome of hematopoietic stem cell transplantation: systematic review and meta-analysis. Blood advances. 2019;3 21:3436-48.
410. Chen W, Ford MS, Young KJ, Cybulsky MI, Zhang L. Role of Double-Negative Regulatory T Cells in Long-Term Cardiac Xenograft Survival. The Journal of Immunology. 2003;170:1846 - 53.
411. Young KJ, DuTemple B, Phillips MJ, Zhang L. Inhibition of Graft-Versus-Host Disease by Double-Negative Regulatory T Cells. The Journal of Immunology. 2003;171:134 - 41.
412. Chen X, Wang D, Zhu X. Application of double-negative T cells in haematological malignancies: recent progress and future directions. Biomarker Research. 2022;10.
413. Körbling M, Anderlini PN. Peripheral blood stem cell versus bone marrow allotransplantation: does the source of hematopoietic stem cells matter? Blood. 2001;98 10:2900-8.
414. Mielcarek M, Martin PJ, Torok-Storb B. Suppression of alloantigen-induced T-cell proliferation by CD14+ cells derived from granulocyte colony-stimulating factor-mobilized peripheral blood mononuclear cells. Blood. 1997;89 5:1629-34.
415. Turcotte LM, Cao Q, Cooley S, Curtsinger JM, Holtan SG, Luo X, et al. Monocyte Subpopulation Recovery as Predictors of Hematopoietic Cell Transplantation Outcomes. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2019;25 5:883-90.
416. Thoma MD, Huneke TJ, Decook L, Johnson ND, Wiegand RA, Litzow MR, et al. Peripheral blood lymphocyte and monocyte recovery and survival in acute leukemia postmyeloablative allogeneic hematopoietic

stem cell transplant. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2012;18 4:600-7.
417. Decook L, Thoma MD, Huneke TJ, Johnson N, Wiegand RA, Patnaik MM, et al. Impact of lymphocyte and monocyte recovery on the outcomes of allogeneic hematopoietic SCT with fludarabine and melphalan conditioning. Bone Marrow Transplantation. 2013;48:708-14.
418. Zhao H-Y, Lyu Z-S, Duan C, Song Y, Han T, Mo X, et al. An unbalanced monocyte macrophage polarization in the bone marrow microenvironment of patients with poor graft function after allogeneic haematopoietic stem cell transplantation. British Journal of Haematology. 2018;182.
419. Iori AP, Torelli GF, de Propris MS, Milano F, Pupella S, Gozzer M, et al. B-Cell Concentration in the Apheretic Product Predicts Acute Graft-Versus-Host Disease and Treatment-Related Mortality of Allogeneic Peripheral Blood Stem Cell Transplantation. Transplantation. 2008;85:386-90.
420. Schultz KR, Paquet J, Bader S, Hayglass KT. Requirement for B cells in T cell priming to minor histocompatibility antigens and development of graft-versus-host disease. Bone marrow transplantation. 1995;16 2:289-95.
421. Zhang C, Todorov I, Zhang Z, Liu Y, Kandeel FR, Forman SJ, et al. Donor CD4+ T and B cells in transplants induce chronic graft-versus-host disease with autoimmune manifestations. Blood. 2006;107 7:2993-3001.
422. Shimabukuro-Vornhagen A, Hallek M, Storb RF, von Bergwelt-Baildon M. The role of B cells in the pathogenesis of graft-versus-host disease. Blood. 2009;114 24:4919-27.
423. Yakoub-Agha I, Saule P, Depil S, Micol J-B, Grutzmacher C, Boulanger-Villard F, et al. A high proportion of donor CD4+ T cells expressing the lymph node-homing chemokine receptor CCR7 increases incidence and severity of acute graft-versus-host disease in patients undergoing allogeneic stem cell transplantation for hematological malignancy. Leukemia. 2006;20:1557-65.
424. Banović T, Markey KA, Kuns RD, Olver SD, Raffelt NC, Don ALJ, et al. Graft-versus-Host Disease Prevents the Maturation of Plasmacytoid Dendritic Cells1. The Journal of Immunology. 2009;182:912 - 20.
425. Lau J, Sartor MM, Bradstock KF, Vuckovic S, Munster DJ, Hart DNJ. Activated Circulating Dendritic Cells After Hematopoietic Stem Cell Transplantation Predict Acute Graft-Versus-Host Disease. Transplantation. 2007;83:839-46.
426. Wen Q, Zhao H-Y, Yao W-L, Zhang Y-y, Fu H-x, Wang Y, et al. Monocyte subsets in bone marrow grafts may contribute to a low incidence of acute graft-vs-host disease for young donors. Journal of Cellular and Molecular Medicine. 2020;24:9204 - 16.
427. Yamasaki S, Henzan H, Ohno Y, Yamanaka T, Iino T, Itou Y, et al. Influence of transplanted dose of CD56+ cells on development of graft-versus-host disease in patients receiving G-CSF-mobilized peripheral blood progenitor cells from HLA-identical sibling donors. Bone Marrow Transplantation. 2003;32:505-10.
428. Kariminia A, Ivison SM, Ng B, Rozmus J, Sung S, Varshney A, et al. CD56bright natural killer regulatory cells in filgrastim primed donor blood or marrow products regulate chronic graft-versus-host disease: the Canadian Blood and Marrow Transplant Group randomized 0601 study results. Haematologica. 2017;102:1936 - 46.
429. Huntington ND, Martinet L, Smyth MJ. DNAM-1: would the real natural killer cell please stand up! Oncotarget. 2015;6:28537 - 8.
430. Muthukrishnan R, Rohini R. LASSO: A feature selection technique in predictive modeling for machine learning. 2016 IEEE International Conference on Advances in Computer Applications (ICACA). 2016:18-20.
431. Frigola XB, Inman BA, Krco CJ, Liu X, Harrington S, Bulur PA, et al. Soluble B7-H1: differences in production between dendritic cells and T cells. Immunology letters. 2012;142 1-2:78-82.
432. Avendaño-Ortiz J, Rubio-Garrido M, Lozano-Rodríguez R, del Romero J, Rodriguez C, Moreno S, et al. Soluble PD-L1: a potential immune marker for HIV-1 infection and virological failure. Medicine. 2020;99.
433. Wherry EJ, Kurachi M. Molecular and cellular insights into T cell exhaustion. Nature Reviews Immunology. 2015;15:486-99.
434. Zhang Z, Liu S, Zhang B, Qiao L, Zhang Y, Zhang Y. T Cell Dysfunction and Exhaustion in Cancer. Frontiers in Cell and Developmental Biology. 2020;8.
435. Jiang Y, Li Y, Zhu B. T-cell exhaustion in the tumor microenvironment. Cell Death & Disease. 2015;6.
436. Barber DL, Wherry EJ, Masopust D, Zhu B, Allison JP, Sharpe AH, et al. Restoring function in exhausted CD8 T cells during chronic viral infection. Nature. 2006;439:682-7.
437. Ribas A, Hu-Lieskovan S. What does PD-L1 positive or negative mean? The Journal of Experimental Medicine. 2016;213:2835 - 40.
438. Ilcus C, Bagacean C, Tempescul A, Popescu C, Parvu A, Cenariu M, et al. Immune checkpoint blockade: the role of PD-1-PD-L axis in lymphoid malignancies. OncoTargets and therapy. 2017;10:2349 - 63.

439. Niu M, Liu Y, Yi M, Jiao D, Wu K. Biological Characteristics and Clinical Significance of Soluble PD-1/PD-L1 and Exosomal PD-L1 in Cancer. Frontiers in Immunology. 2022;13.
440. Rao M, Valentini D, Dodoo E, Zumla A, Maeurer M. Anti-PD-1/PD-L1 therapy for infectious diseases: learning from the cancer paradigm. International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases. 2017;56:221-8.
441. Mahoney KM, Ross-Macdonald P, Yuan L, Song L, Veras E, Wind-Rotolo M, et al. Soluble PD-L1 as an early marker of progressive disease on nivolumab. Journal for Immunotherapy of Cancer. 2022;10.
442. Freeman GJ, Long AJ, Iwai Y, Bourque K, Chernova T, Nishimura H, et al. Engagement of the Pd-1 Immunoinhibitory Receptor by a Novel B7 Family Member Leads to Negative Regulation of Lymphocyte Activation. The Journal of Experimental Medicine. 2000;192:1027 - 34.
443. Tremblay-Lemay R, Rastgoo N, Chang H. Modulating PD-L1 expression in multiple myeloma: an alternative strategy to target the PD-1/PD-L1 pathway. Journal of Hematology & Oncology. 2018;11.
444. Hude I, Sasse S, Engert A, Bröckelmann PJ. The emerging role of immune checkpoint inhibition in malignant lymphoma. Haematologica. 2017;102:30 - 42.
445. Udall M, Rizzo MC, Kenny J, Doherty JP, Dahm S, Robbins P, et al. PD-L1 diagnostic tests: a systematic literature review of scoring algorithms and test-validation metrics. Diagnostic Pathology. 2018;13.
446. Cho I, Lee H, Yoon SE, Ryu KJ, Ko YH, Kim WS, et al. Serum levels of soluble programmed death-ligand 1 (sPD-L1) in patients with primary central nervous system diffuse large B-cell lymphoma. BMC Cancer. 2020;20.
447. Daassi D, Mahoney KM, Freeman GJ. The importance of exosomal PDL1 in tumour immune evasion. Nature Reviews Immunology. 2020;20:209-15.
448. Simonetta F, Pradier A, Bosshard C, Masouridi-Levrat S, Dantin C, Koutsi A, et al. Dynamics of Expression of Programmed Cell Death Protein-1 (PD-1) on T Cells After Allogeneic Hematopoietic Stem Cell Transplantation. Frontiers in Immunology. 2019;10.
449. Hu Y, Chen W, Yan Z, Ma J-Y, Zhu F, Huo J. Prognostic value of PD-L1 expression in patients with pancreatic cancer. Medicine. 2019;98.
450. Guo L, Li W, Zhu X, Ling Y, Qiu T, Dong L, et al. PD-L1 expression and CD274 gene alteration in triple-negative breast cancer: implication for prognostic biomarker. SpringerPlus. 2016;5.
451. Cheng Y, Wang C, Wang Y, Dai L. Soluble PD-L1 as a predictive biomarker in lung cancer: a systematic review and meta-analysis. Future oncology. 2021.
452. Tuminello S, Sikavi DR, Veluswamy R, Gamarra C, Lieberman-Cribbin W, Flores RM, et al. PD-L1 as a prognostic biomarker in surgically resectable non-small cell lung cancer: a meta-analysis. Translational Lung Cancer Research. 2020;9:1343 - 60.
453. Boussiotis VA. Molecular and Biochemical Aspects of the PD-1 Checkpoint Pathway. The New England journal of medicine. 2016;375 18:1767-78.
454. Saha A, O'Connor RS, Thangavelu G, Lovitch SB, Dandamudi DB, Wilson CB, et al. Programmed death ligand-1 expression on donor T cells drives graft-versus-host disease lethality. The Journal of clinical investigation. 2016;126 7:2642-60.
455. Lun Y, Borjini N, Miura NN, Ohno N, Singer NG, Lin F. CDCP1 on Dendritic Cells Contributes to the Development of a Model of Kawasaki Disease. The Journal of Immunology. 2021;206:2819 - 27.
456. Casar B, Rimann I, Kato H, Shattil SJ, Quigley JP, Deryugina EI. In vivo cleaved CDCP1 promotes early tumor dissemination via complexing with activated β1 integrin and induction of FAK/PI3K/Akt motility signaling. Oncogene. 2014;33:255-68.
457. Wright HJ, Hou J, Xu B, Cortez M, Potma EO, Tromberg BJ, et al. CDCP1 drives triple-negative breast cancer metastasis through reduction of lipid-droplet abundance and stimulation of fatty acid oxidation. Proceedings of the National Academy of Sciences. 2017;114:E6556 - E65.
458. Spassov DS, Wong C-h, Wong SY, Reiter JF, Moasser MM. Trask loss enhances tumorigenic growth by liberating integrin signaling and growth factor receptor cross-talk in unanchored cells. Cancer research. 2013;73 3:1168-79.
459. Frändberg S, Asp J, Waldner B, Holgersson J, Palmqvist L. Concentration of the CDCP1 protein in human cord plasma may serve as a predictor of hematopoietic stem and progenitor cell content. Stem cell research. 2018;29:24-7.
460. Dagnino S, Bodinier B, Guida F, Smith-Byrne K, Petrovic D, Whitaker MD, et al. Prospective Identification of Elevated Circulating CDCP1 in Patients Years before Onset of Lung Cancer. Cancer Research. 2021;81:3738 - 48.
461. Scherl-Mostageer M, Sommergruber W, Abseher R, Hauptmann R, Ambros PF, Schweifer N. Identification of a novel gene, CDCP1, overexpressed in human colorectal cancer. Oncogene. 2001;20:4402-8.

462. Heitmann JS, Hagelstein I, Hinterleitner C, Roerden M, Jung G, Salih HR, et al. Identification of CD318 (CDCP1) as novel prognostic marker in AML. Annals of Hematology. 2020;99:477 - 86.
463. Kollmorgen G, Niederfellner G, Lifke A, Spohn GJ, Rieder N, Vega Harring S, et al. Antibody mediated CDCP1 degradation as mode of action for cancer targeted therapy. Molecular Oncology. 2013;7.
464. Gioia R, Leroy C, Drullion C, Lagarde V, Étienne G, Dulucq S, et al. Quantitative phosphoproteomics revealed interplay between Syk and Lyn in the resistance to nilotinib in chronic myeloid leukemia cells. Blood. 2011;118 8:2211-21.
465. Boyer AP, Collier TS, Vidavsky I, Bose R. Quantitative Proteomics with siRNA Screening Identifies Novel Mechanisms of Trastuzumab Resistance in HER2 Amplified Breast Cancers. Molecular & Cellular Proteomics. 2012;12:180 - 93.
466. Miura S, Hamada S, Masamune A, Satoh K, Shimosegawa T. CUB-domain containing protein 1 represses the epithelial phenotype of pancreatic cancer cells. Experimental cell research. 2014;321 2:209-18.
467. Korbecki J, Kojder K, Simińska D, Bohatyrewicz R, Gutowska I, Chlubek D, et al. CC Chemokines in a Tumor: A Review of Pro-Cancer and Anti-Cancer Properties of the Ligands of Receptors CCR1, CCR2, CCR3, and CCR4. International Journal of Molecular Sciences. 2020;21.
468. Lakshmikanth T, Olin A, Chen Y, Mikes J, Fredlund E, Remberger M, et al. Mass Cytometry and Topological Data Analysis Reveal Immune Parameters Associated with Complications after Allogeneic Stem Cell Transplantation. Cell reports. 2017;20 9:2238-50.
469. Nardelli B, Tiffany HL, Bong GW, Yourey PA, Morahan DK, Li Y, et al. Characterization of the signal transduction pathway activated in human monocytes and dendritic cells by MPIF-1, a specific ligand for CC chemokine receptor 1. Journal of immunology. 1999;162 1:435-44.
470. Choi SW, Hildebrandt GC, Olkiewicz KM, Hanauer DA, Chaudhary MN, Silva IA, et al. CCR1/CCL5 (RANTES) receptor-ligand interactions modulate allogeneic T-cell responses and graft-versus-host disease following stem-cell transplantation. Blood. 2007;110 9:3447-55.
471. Karan D. CCL23 in Balancing the Act of Endoplasmic Reticulum Stress and Antitumor Immunity in Hepatocellular Carcinoma. Frontiers in Oncology. 2021;11.
472. Meng J, Wang L, Hou J, Yang X-f, Lin K, Nan H, et al. CCL23 suppresses liver cancer progression through the CCR1/AKT/ESR1 feedback loop. Cancer Science. 2021;112:3099 - 110.
473. Nishida N, Yano H, Nishida T, Kamura T, Kojiro M. Angiogenesis in Cancer. Vascular Health and Risk Management. 2006;2:213 - 9.
474. Han KY, Kim CW, Lee TH, Son Y, Kim J. CCL23 up-regulates expression of KDR/Flk-1 and potentiates VEGF-induced proliferation and migration of human endothelial cells. Biochemical and biophysical research communications. 2009;382 1:124-8.
475. Krishnan V, Tallapragada S, Schaar BT, Kamat K, Chanana AM, Zhang Y, et al. Omental macrophages secrete chemokine ligands that promote ovarian cancer colonization of the omentum via CCR1. Communications Biology. 2020;3.
476. Kamat K, Krishnan V, Dorigo O. Macrophage-derived CCL23 upregulates expression of T-cell exhaustion markers in ovarian cancer. British journal of cancer. 2022.
477. Scheller J, Garbers C, Rose-John S. Interleukin-6: from basic biology to selective blockade of pro-inflammatory activities. Seminars in immunology. 2014;26 1:2-12.
478. Narazaki M, Kishimoto T. The Two-Faced Cytokine IL-6 in Host Defense and Diseases. International Journal of Molecular Sciences. 2018;19.
479. Tvedt THA, Ersvaer E, Tveita AA, Bruserud Ø. Interleukin-6 in Allogeneic Stem Cell Transplantation: Its Possible Importance for Immunoregulation and As a Therapeutic Target. Frontiers in Immunology. 2017;8.
480. Tvedt THA, Lie SA, Reikvam H, Rye KP, Lindås R, Gedde-Dahl T, et al. Pretransplant Levels of CRP and Interleukin-6 Family Cytokines; Effects on Outcome after Allogeneic Stem Cell Transplantation. International Journal of Molecular Sciences. 2016;17.
481. Kennedy GA, Varelias A, Vuckovic S, Le Texier L, Gartlan KH, Zhang P, et al. Addition of interleukin-6 inhibition with tocilizumab to standard graft-versus-host disease prophylaxis after allogeneic stem-cell transplantation: a phase 1/2 trial. The Lancet Oncology. 2014;15 13:1451-9.
482. Kishimoto T. Interleukin-6: from basic science to medicine--40 years in immunology. Annual review of immunology. 2005;23:1-21.
483. Chen X, Das R, Komorowski RA, Beres AJ, Hessner MJ, Mihara M, et al. Blockade of interleukin-6 signaling augments regulatory T-cell reconstitution and attenuates the severity of graft-versus-host disease. Blood. 2009;114 4:891-900.

484. Tawara I, Koyama M, Liu C, Toubai T, Thomas DG, Evers R, et al. Interleukin-6 Modulates Graft-versus-Host Responses after Experimental Allogeneic Bone Marrow Transplantation. Clinical Cancer Research. 2010;17:77 - 88.
485. Drobyski WR, Szabo A, Zhu F, Keever-Taylor CA, Hebert KM, Dunn R, et al. Tocilizumab, tacrolimus and methotrexate for the prevention of acute graft-versus-host disease: low incidence of lower gastrointestinal tract disease. Haematologica. 2018;103:717 - 27.
486. Wilkinson AN, Chang K, Kuns RD, Henden AS, Minnie SA, Ensbey KS, et al. IL-6 dysregulation originates in dendritic cells and initiates graft-versus-host disease via classical signaling. Blood. 2019.
487. Hill GR, Koyama M. Cytokines and Co-stimulation in Acute Graft-versus-Host Disease. Blood. 2020.
488. Geleff S, Draganovici D, Jaksch P, Segerer S. The role of chemokine receptors in acute lung allograft rejection. European Respiratory Journal. 2009;35:167 - 75.
489. Shino MY, Weigt SS, Li N, Palchevskiy V, Derhovanessian A, Saggar R, et al. CXCR3 ligands are associated with the continuum of diffuse alveolar damage to chronic lung allograft dysfunction. American journal of respiratory and critical care medicine. 2013;188 9:1117-25.
490. Agostini C, Calabrese F, Rea F, Facco M, Tosoni A, Loy M, et al. Cxcr3 and its ligand CXCL10 are expressed by inflammatory cells infiltrating lung allografts and mediate chemotaxis of T cells at sites of rejection. The American journal of pathology. 2001;158 5:1703-11.
491. Lee EY, Lee ZH, Song Y-W. CXCL10 and autoimmune diseases. Autoimmunity reviews. 2009;8 5:379-83.
492. Wang L, Chen P, Luo S, Li J, Liu K, Hu H-z, et al. CXC-chemokine-ligand-10 gene therapy efficiently inhibits the growth of cervical carcinoma on the basis of its anti-angiogenic and antiviral activity. Biotechnology and Applied Biochemistry. 2009;53.
493. Enderlin M, Kleinmann EV, Struyf S, Buracchi C, Vecchi A, Kinscherf R, et al. TNF-α and the IFN-γ-inducible protein 10 (IP-10/CXCL-10) delivered by parvoviral vectors act in synergy to induce antitumor effects in mouse glioblastoma. Cancer Gene Therapy. 2009;16:149-60.
494. Sallusto F, Lenig D, Mackay CR, Lanzavecchia A. Flexible Programs of Chemokine Receptor Expression on Human Polarized T Helper 1 and 2 Lymphocytes. The Journal of Experimental Medicine. 1998;187:875 - 83.
495. Loetscher M, Loetscher P, Brass N, Meese E, Moser B. Lymphocyte-specific chemokine receptor CXCR3: regulation, chemokine binding and gene localization. European Journal of Immunology. 1998;28.
496. Persano L, Crescenzi M, Indraccolo S. Anti-angiogenic gene therapy of cancer: current status and future prospects. Molecular aspects of medicine. 2007;28 1:87-114.
497. Pellegrino A, Antonaci F, Russo FY, Merchionne F, Ribatti D, Vacca A, et al. CXCR3-binding chemokines in multiple myeloma. Cancer letters. 2004;207 2:221-7.
498. Jones DM, Benjamin RJ, Shahsafaei A, Dorfman DM. The chemokine receptor CXCR3 is expressed in a subset of B-cell lymphomas and is a marker of B-cell chronic lymphocytic leukemia. Blood. 2000;95 2:627-32.
499. Datta D, Flaxenburg JA, Laxmanan S, Geehan CS, Grimm MR, Waaga-Gasser AM, et al. Ras-induced modulation of CXCL10 and its receptor splice variant CXCR3-B in MDA-MB-435 and MCF-7 cells: relevance for the development of human breast cancer. Cancer research. 2006;66 19:9509-18.
500. Sempowski GD, Hale LP, Sundy JS, Massey JM, Koup RA, Douek DC, et al. Leukemia Inhibitory Factor, Oncostatin M, IL-6, and Stem Cell Factor mRNA Expression in Human Thymus Increases with Age and Is Associated with Thymic Atrophy1. The Journal of Immunology. 2000;164:2180 - 7.
501. Gearing DP, Gough NM, King JA, Hilton DJ, Nicola NA, Simpson RJY, et al. Molecular cloning and expression of cDNA encoding a murine myeloid leukaemia inhibitory factor (LIF). The EMBO Journal. 1987;6.
502. Li X, Yang Q, Yu H, Wu L, Zhao Y, Zhang C, et al. LIF promotes tumorigenesis and metastasis of breast cancer through the AKT-mTOR pathway. Oncotarget. 2014;5:788 - 801.
503. Fitzgerald JS, Tsareva SA, Poehlmann TG, Berod L, Meissner A, Corvinus FM, et al. Leukemia inhibitory factor triggers activation of signal transducer and activator of transcription 3, proliferation, invasiveness, and altered protease expression in choriocarcinoma cells. The international journal of biochemistry & cell biology. 2005;37 11:2284-96.
504. Wysoczynski M, Miekus K, Jankowski K, Wanzeck J, Bertolone SJ, Janowska-Wieczorek A, et al. Leukemia inhibitory factor: a newly identified metastatic factor in rhabdomyosarcomas. Cancer research. 2007;67 5:2131-40.
505. Liu S-C, Tsang N-M, Chiang W-C, Chang K, Hsueh C, Liang Y, et al. Leukemia inhibitory factor promotes nasopharyngeal carcinoma progression and radioresistance. The Journal of clinical investigation. 2013;123 12:5269-83.
506. Yu H, Yue X, Zhao Y, Li X, Wu L, Zhang C, et al. LIF negatively regulates tumour-suppressor p53 through Stat3/ID1/MDM2 in colorectal cancers. Nature Communications. 2014;5(1):5218.

507. D'aveni M, Rossignol J, Coman T, Sivakumaran S, Henderson S, Manzo T, et al. G-CSF mobilizes CD34+ regulatory monocytes that inhibit graft-versus-host disease. Science Translational Medicine. 2015;7:281ra42 - ra42.
508. Orlova D, Meehan S, Parks D, Moore W, Meehan C, Zhao Q, et al. QFMatch: multidimensional flow and mass cytometry samples alignment. Scientific Reports. 2018;8.
509. Li YH, Li D, Samusik N, Wang X, Guan L, Nolan GP, et al. Scalable multi-sample single-cell data analysis by Partition-Assisted Clustering and Multiple Alignments of Networks. PLoS Computational Biology. 2017;13.
510. Schuyler RP, Jackson CL, Garcia-Perez JE, Baxter RM, Ogolla SO, Rochford R, et al. Minimizing Batch Effects in Mass Cytometry Data. Frontiers in Immunology. 2019;10.
511. Pedersen CB, Dam SH, Barnkob MB, Leipold MD, Purroy N, Rassenti L, et al. cyCombine allows for robust integration of single-cell cytometry datasets within and across technologies. Nature Communications. 2022;13.
512. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? International Journal of Epidemiology. 2014;43:1336 - 9.
513. White IR, Royston P. Imputing missing covariate values for the Cox model. Statistics in Medicine. 2009;28:1982 - 98.
514. Bartlett JW, Taylor JM. Missing covariates in competing risks analysis. Biostatistics (Oxford, England). 2016;17:751 - 63.
515. Weber LM, Nowicka M, Soneson C, Robinson MD. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. Communications Biology. 2019;2.
516. Quinn T, Richardson MF, Lovell DR, Crowley TM. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. Scientific Reports. 2017;7.
517. Greenacre MJ, Grunsky EC, Bacon-Shone J, Erb I, Quinn T, editors. Aitchison's Compositional Data Analysis 40 Years On: A Reappraisal2022.
518. Filzmoser P, Hron K, Reimann C. Principal component analysis for compositional data with outliers. Environmetrics. 2009;20.
519. Godichon-Baggioni A, Maugis-Rabusseau C, Rau A. Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. Journal of Applied Statistics. 2017;46:47 - 65.
520. Aitchison JW, editor The single principle of compositional data analysis, continuing fallacies, confusionsand misunderstandings and some suggested remedies2008.
521. Kjærgaard S, Ergemen YE, Kallestrup-Lamb M, Oeppen J, Lindahl-Jacobsen R. Forecasting causes of death by using compositional data analysis: the case of cancer deaths. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2019;68.
522. Bergeron-Boucher M-P, Canudas-Romo V, Oeppen J, Vaupel JW. Coherent Forecasts of Mortality with Compositional Data Analysis. Demographic Research. 2017;37:527-66.
523. Oeppen J, editor Coherent forecasting of multiple-decrement life tables: a test using Japanese cause of death data2008.
524. Navarro-Lopez C, Linares-Mustarós S, Mulet-Forteza C. "The Statistical Analysis of Compositional Data" by John Aitchison (1986): A Bibliometric Overview. SAGE Open. 2022;12.
525. Pal S, Heumann C. Clustering compositional data using Dirichlet mixture model. PLoS ONE. 2022;17.
526. Jeng XJ, Chen X. Variable selection via adaptive false negative control in linear regression. Electronic Journal of Statistics. 2019.
527. Yasuoka H, Asai A, Ohama H, Tsuchimoto Y, Fukunishi S, Higuchi K. Increased both PD–L1 and PD–L2 expressions on monocytes of patients with hepatocellular carcinoma was associated with a poor prognosis. Scientific Reports. 2020;10.
528. Mbongue JC, Nieves HA, Torrez TW, Langridge WHR. The Role of Dendritic Cell Maturation in the Induction of Insulin-Dependent Diabetes Mellitus. Frontiers in Immunology. 2017;8.
529. Arend L, Bernett J, Manz Q, Klug M, Lazareva OV, Baumbach J, et al. A systematic comparison of novel and existing differential analysis methods for CyTOF data. Briefings in bioinformatics. 2022.
530. Zhou H, He K, Chen J, Zhang X. LinDA: linear models for differential abundance analysis of microbiome compositional data. Genome Biology. 2022;23.
531. Abdelaal T, van Unen V, Höllt T, Koning F, Reinders MJT, Mahfouz A. Predicting Cell Populations in Single Cell Mass Cytometry Data. Cytometry. 2019;95:769 - 81.
532. Li H, Shaham U, Stanton KP, Yao Y, Montgomery RR, Kluger Y. Gating mass cytometry data by deep learning. Bioinformatics. 2017;33(21):3423-30.
533. Liu P, Liu S, Fang Y, Xue X, Zou J, Tseng GC, et al. Recent Advances in Computer-Assisted Algorithms for Cell Subtype Identification of Cytometry Data. Frontiers in Cell and Developmental Biology. 2020;8.

534. Lin W, Shi P, Feng R, Li H. Variable selection in regression with compositional covariates. Biometrika. 2014;101:785-97.
535. Raftery AE, Dean N. Variable Selection for Model-Based Clustering. Journal of the American Statistical Association. 2006;101:168 - 78.
536. Wang X, Wang H, Wang S, Yuan J. Convex clustering method for compositional data via sparse group lasso. Neurocomputing. 2021;425:23-36.
537. Hu Y, Li Y, Satten GA, Hu Y. Testing microbiome associations with survival times at both the community and individual taxon levels. PLoS Computational Biology. 2022;18.
538. Koh H, Livanos AE, Blaser MJ, Li H. A highly adaptive microbiome-based association test for survival traits. BMC Genomics. 2018;19.
539. Plantinga AM, Zhan X, Zhao N, Chen J, Jenq RR, Wu MC. MiRKAT-S: a community-level test of association between the microbiota and survival times. Microbiome. 2017;5.
540. Micheel CM, Nass SJ, Omenn GS, editors. Evolution of Translational Omics: Lessons Learned and the Path Forward2013.
541. Benner A, Zucknick M, Hielscher T, Ittrich C, Mansmann UR. High-Dimensional Cox Models: The Choice of Penalty as Part of the Model Building Process. Biometrical Journal. 2010;52.
542. Sheikh M, Coolen ACC. Analysis of overfitting in the regularized Cox model. Journal of Physics A: Mathematical and Theoretical. 2019;52.
543. Xu J. High-Dimensional Cox Regression Analysis in Genetic Studies with Censored Survival Outcomes. Journal of Probability and Statistics. 2012;2012:1-14.
544. Chen L, editor Curse of Dimensionality. Encyclopedia of Database Systems; 2009.
545. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics. 2022;2.
546. Emura T, Matsui S, Chen H-Y. compound.Cox: Univariate feature selection and compound covariate for predicting survival. Computer methods and programs in biomedicine. 2019;168:21-37.
547. Etz A. Introduction to the Concept of Likelihood and Its Applications. Advances in Methods and Practices in Psychological Science. 2018;1:60 - 9.
548. Petrov BN, Csaki F, editors. 2nd international symposium on information theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971. Budapest, Hungary: Akademiai Kiado; 1973.
549. Lee J, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the lasso. arXiv: Statistics Theory. 2013.
550. Breiman L. Random Forests. Machine Learning. 2004;45:5-32.
551. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001;29:1189-232.
552. Barraza JF, Droguett EL, Martins MR. Towards Interpretable Deep Learning: A Feature Selection Framework for Prognostics and Health Management Using Deep Neural Networks. Sensors (Basel, Switzerland). 2021;21.
553. Wang P, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. ArXiv. 2017;abs/1708.04649.
554. Spooner A, Chen E, Sowmya A, Sachdev PS, Kochan NA, Trollor JN, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Scientific Reports. 2020;10.
555. Tapak L, Saidijam M, Sadeghifar M, Poorolajal J, Mahjub H. Competing Risks Data Analysis with High-dimensional Covariates: An Application in Bladder Cancer. Genomics, Proteomics & Bioinformatics. 2015;13:169 - 76.
556. Morris EL, He K, Li Y, Li Y, Kang J. SurvBoost: An R Package for High-Dimensional Variable Selection in the Stratified Proportional Hazards Model via Gradient Boosting. R J. 2020;12 1:105-17.
557. Bin RD. Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. Comput Stat. 2016;31:513-31.
558. Binder H, Allignol A, Schumacher M, Beyersmann J. Boosting for high-dimensional time-to-event data with competing risks. Bioinformatics. 2009;25 7:890-6.
559. Schapire RE, editor The Strength of Weak Learnability (Extended Abstract). FOCS 1989; 1989.
560. Zhou Z-H, editor Ensemble Learning. Encyclopedia of Biometrics; 2009.
561. Berk RA. An Introduction to Ensemble Methods for Data Analysis. Sociological Methods & Research. 2004;34:263 - 95.
562. Rezaei M, Tapak L, Alimohammadian M, Sadjadi A, Yaseri M, editors. Review of Random Survival Forest method2020.

563. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2006;8:25 -
564. Bühlmann P, Yu B. Boosting. WIREs Computational Statistics. 2010;2(1):69-74.
565. Buhlmann P, Hothorn T. BOOSTING ALGORITHMS: REGULARIZATION, PREDICTION AND MODEL FITTING. Statistical Science. 2007;22:477-505.
566. CRAN. Package 'CoxBoost' was removed from the CRAN repository. The Comprehensive R Archive Network2020 [Available from: https://cran.r-project.org/web/packages/CoxBoost/index.html.

Appendix A Machine learning of COVID-19 clinical data identifies population structures with therapeutic potential.

Published article

David Greenwood,[1,2] Thomas Taverner,[3] Nicola J. Adderley,[3] Malcolm James Price,[3,4] Krishna Gokhale,[3] Christopher Sainsbury,[3] Suzy Gallier,[5,6] Carly Welch,[5,6] Elizabeth Sapey,[5,6,7] Duncan Murray,[1,5] Hilary Fanning,[5] Simon Ball,[1,5,7] Krishnarajah Nirantharakumar,[3,7] Wayne Croft,[1,2] and Paul Moss[1,5,8]

iScience. 2022 Jul 15;25(7):104480. doi: 10.1016/j.isci.2022.104480.

https://doi.org/10.1016/j.isci. 2022.104480

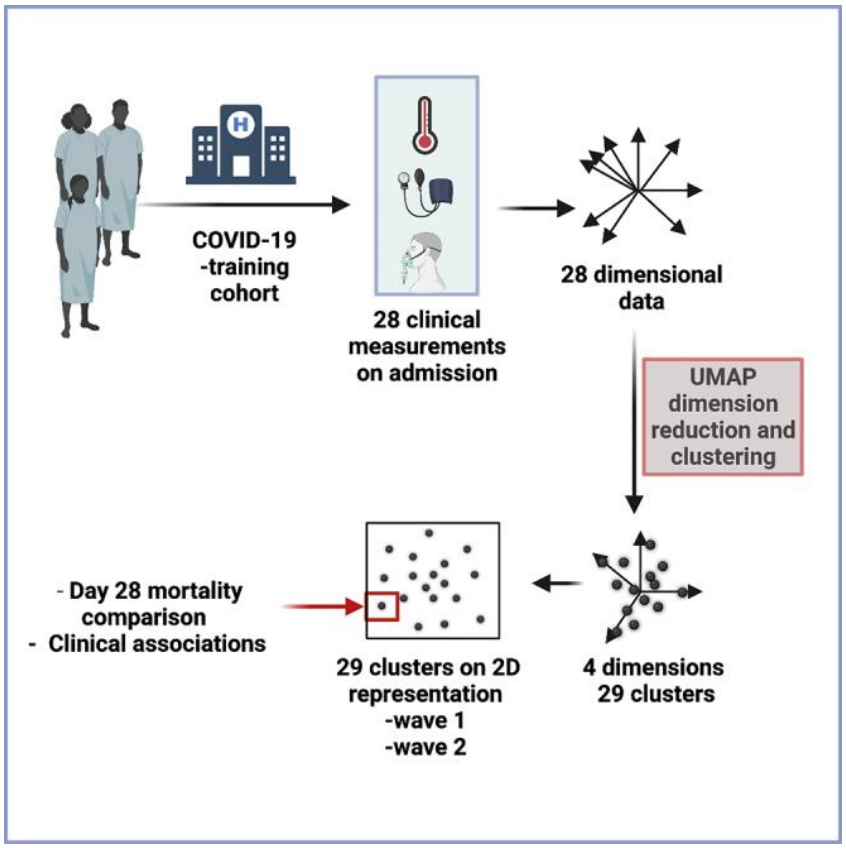
D.G. performed the statistical analysis. D.G., W.C., and P.M. oversaw statistical analyses. D.G., W.C., K.N., N.J.A., T.T., K.G., and P.M. had unrestricted access to all data. D.G., W.C., and P.M. prepared the first draft of the manuscript, reviewed, and edited it. All authors agreed to submit the manuscript, read, and approved the final draft and take full responsibility of its content, including the accuracy of the data, and the fidelity of statistical analysis.

1. Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, UK
2. The Centre for Computational Biology, University of Birmingham, Birmingham, UK
3. Institute of Applied Health Research, University of Birmingham, Birmingham, UK
4. NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK
5. University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
6. Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK
7. Health Data Research, London, UK

A.1 Abstract

Clinical outcomes for patients with COVID-19 are heterogeneous and there is interest in defining subgroups for prognostic modeling and development of treatment algorithms. We obtained 28 demographic and laboratory variables in patients admitted to hospital with COVID-19. These comprised a training cohort (n = 6099) and two validation cohorts during the first and second waves of the pandemic (n = 996; n = 1011). Uniform manifold approximation and projection (UMAP) dimension reduction and Gaussian mixture model (GMM) analysis was used to define patient clusters. 29 clusters were defined in the training cohort and associated with markedly different mortality rates, which were predictive within confirmation datasets. Deconvolution of clinical features within clusters identified unexpected relationships between variables. Integration of large datasets using UMAP-assisted clustering can therefore identify patient subgroups with prognostic information and uncovers unexpected interactions between clinical variables. This application of machine learning represents a powerful approach for delineating disease pathogenesis and potential therapeutic interventions.

A.2 Graphical abstract



A.3 Introduction

The COVID-19 pandemic has led to >4.6 million deaths to date, but the clinical outcome after primary infection is heterogeneous and approaches to predict outcome within individual patients are of value. Several demographic features increase the mortality risk, including age and comorbid conditions, and have been used to define clinical risk scores. The ISARIC4C mortality score is used widely and assesses nine demographic and laboratory values (554).

This clinical heterogeneity of COVID-19 has led to interest in further defining patient subgroups (555, 556), but although inclusion of more demographic and laboratory values can improve accuracy, the associated increase in data dimensionality is challenging for clustering algorithms. As such, integration of outputs from very high dimensional datasets usually requires either feature selection or dimensionality reduction. Although the latter approach loses some information, the associated compression and noise reduction greatly improves utility and is used commonly in biological assessments such as analysis of single-cell RNA sequencing data (557). The reduced dimensional outputs created through principal component analysis (PCA) are a linear combination of the input and struggle to capture complex nonlinearities (242). Nonlinear dimension reduction or manifold learning techniques such as Uniform Manifold Approximation and Projection (UMAP) (295, 558), are based on topological analysis and can define groups without prespecifying a target for the algorithm (305). The combination of UMAP dimensionality reduction and model-based clustering offer a powerful approach to unsupervised machine learning (559, 560).

To identify subgroups of patients with COVID-19, we obtained information on 28 clinical, demographic, and laboratory variables in a training cohort of 6099 patients on the day of their hospital admission with acute COVID-19 with outcomes followed for 28 days including inpatient and community mortality. UMAP dimensionality reduction and clustering were then applied to define clinical subgroups that predicted mortality and were confirmed in two large patient validation cohorts. Besides, it identified unexpected interrelationships between variables that could be valuable for guiding treatment pathways.

A.4 Methods

A.4.1 Patient population

Clinical data were acquired for patients admitted to hospital with COVID-19. The CovidCollab cohort is a multicentre dataset of 6099 patients admitted between January and August 2020 and was used as the training cohort in model development (561). Two cohorts of patients admitted to the University Hospitals NHS Foundation Trust Birmingham (UHBFT) were used for validation. The first cohort ('wave 1') includes 996 patients admitted between January and August 2020. The second cohort ('wave 2') includes 1011 patients admitted between September 2020 and January 2021 (**Table A.1**). Patients tested positive for SARS-CoV-2 by PCR and/or antibody test within 14 days of admission. Mortality at day 28 (including after discharge) was determined for each patient.

Ethical approval was provided by the East Midlands–Derby REC (reference: 20/EM/0158) for the PIONEER Research Database (data from University Hospitals Birmingham). For CovidCollab data, local, regional and national approvals were obtained from all participating sites. In the UK, this study was registered as clinical audit or service evaluation, with approval granted in line with local information governance policies, in line with assessment and guidance by the Health Research Authority. At the lead site (University Hospitals Birmingham NHS Trust), this study was registered as clinical audit (CARMS-15986). In other countries, local principal investigators were responsible for obtaining approvals in line with their local, regional and national guidelines and recommendations. Only routinely collected data were collected and patient care was not altered by this study. Anonymised data were securely transferred to the Birmingham Centre for Prospective and Observational Studies, University of Birmingham via REDCap. All sites were required to confirm that approvals were in place prior to being provided with logins; written data sharing agreements were arranged where requested by individual sites.

A.4.2 Variables and missing data

Data obtained for each patient included demographic, clinical and laboratory variables taken on the day of admission (**Table A.1** and **Table S. 2**). Analysis was conducted in R (562) with RStudio (563). Missing data were quantified by cohort (**Figure S. 1**) and imputed from the observed data by predictive mean matching (564-566). The Multivariate Imputation by Chained Equations (MICE) (567) algorithm was applied independently to each cohort under fully conditional specification. 28 variables (continuous and binary) (**Table A.1**) were used for model development for which the maximum proportion of missing observations was 36% in the training cohort, 43% in wave 1 and 42% in wave 2. A single imputation of the training cohort was used for model development. Multiple imputations of waves 1 and 2 were analysed for validation (n. imputations = 5).

A.4.3 Data processing

UMAP was applied to the training cohort prior to model development (307) (**Figure S. 2**). Binary variables were one-hot encoded (0/1), and all variables standardised prior to transformation (mean 0, variance 1). UMAP transformed the 28 clinical variables onto a lower dimensional embedding. Initial hyper-parameters were selected by visualising a 2-D embedding output (**Figure S. 3**). A Euclidean distance metric was used with a target of 40 nearest neighbours and a minimum distance of 0.25.

A.4.4 Clustering model development

Gaussian mixture model (GMM) clustering was applied to automatically label the distribution of patients after transformation by UMAP (274, 568) (**Figure S. 2**). An Expectation-Maximization (EM) algorithm was used to estimate GMM parameters with unconstrained covariance matrices (274).

The number of UMAP dimensions to output for clustering, *D*, was determined qualitatively. Maximum likelihood estimation of intrinsic dimension (291, 569, 570) and Factor Analysis of Mixed Data (FAMD) (571-573) were applied to determine a range for D based on data complexity. Iterative UMAP dimension reduction was used to test the effect of D on the clustering model. GMMs were fitted to each D-dimensional embedding with between 2 and 50 mixture components. Average silhouette width (284, 574, 575) and Bayesian information criterion (BIC) (274, 576) were measured. The maximum D was selected which did not substantially reduce silhouette width or result in high variability in BIC between models. From this, 4 dimensions (4-D embedding) were selected for onward analysis (**Figure S. 4**).

The number of mixture components, *K*, and therefore clusters of patients, was determined using the optimal BIC, maximum silhouette width, and a qualitative assessment of BIC and silhouette width plots (manual BIC, manual silhouette). If multiple values for K were selected, the diversity of mortality rates was then compared between these models and the mortality rate at day 28 after hospital admission calculated for K clusters. ISARIC4C modelling had shown that mortality rates varied markedly between high risk and low risk groups (**Figure A.1**) and the model with the largest range of mortality rate was therefore retained. Based on this, the final reported model was a GMM fitted with K = 29 (**Figure S. 4**).

A.4.5 Clustering validation data

A two-stage process assigned patients from waves 1 and 2 into clusters detected by the GMM in model development. First, the UMAP transformation learned from the training cohort was applied to embed the validation cohorts onto the same 4-D embedding. Secondly, the trained GMM was applied to predict which clusters the new observations should be assigned to with the highest probability. This was repeated for 5 imputations and a majority vote was taken as the final cluster classification (577). Separation of clusters was assessed for each imputation by silhouette analysis and estimates were pooled with a 95% confidence interval (CI) by applying Rubin's rules (578).

A.4.6 2-D visualisation

For visualisation purposes the UMAP analysis was repeated to transform the 4-D embedding of the training cohort from 4-D to 2-D. The same transformation was applied to each validation cohort and scatter plots (UMAP plots) were created (289).

A.4.7 Quantification and statistical analysis

A.4.7.1 Comparing cohorts

Complete cases of each variable were compared between the training and validation cohorts using a Wilcoxon test for continuous variables or Fisher's test for categorical variables (579, 580). The Benjamini-Hochberg procedure was applied to adjust for multiple testing (339). Mortality rate at day 28 after hospital admission was calculated for individual clusters across each of the three patient cohorts and odds-ratios were calculated (95% CI, 1000 bootstrap resamples) (581).

A.4.7.2 Cluster characterisation analysis

Fisher's tests were applied to detect variables significantly associated with each cluster ($p = 0.001$). Continuous variables were categorised into ordinal factors from literature review and tests applied to each factor level (**Table S. 2**). Only variables where the association was significant in the training cohort and at least one validation cohort were reported here. Results were summarised using word clouds with variables scaled by the strength of the association in the training cohort (582-584). Word clouds were overlaid on to a UMAP plot of the training cohort to aid interpretation. Clusters with fewer than 10 patients in either validation cohort were excluded.

A.4.7.3 Risk classification

ISARIC4C mortality score was calculated for each patient (554) and revealed 0–88% risk of death within 28 days. Scores were divided into low (0–3), intermediate (4–8), high (9–14) and very high risk (15–21) for plotting purposes. The training cohort was assigned a score based off a single imputation, waves 1 and 2 were assigned a score based off the mode across 5 imputations. Kaplan–Meier (KM) (317) curves were constructed to estimate survival rates with 95% CI. A log-rank test was applied to test for differences between groups.

A.4.7.4 Role of the funding source

The work was funded from an NIHR grant to PM. The sponsor of the ethics had no role in decision to publish, collection of data or authorship. The contributions by NA, ES, KN, MP, CS and TT were funded by the Medical Research Council UK Research and Innovation (reference COV0306) during the study. The funder had no role in developing the research question or the study protocol.

A.5 Results

A.5.1 Survival probability between cohorts and risk groups

Significant differences in survival probability were observed between the cohorts ($p < 0.0001$). 28 days after admission, the survival probability in the training cohort was 74% (CI 73–75), whereas this was 71% in the wave 1 cohort (CI 69–74) and increased to 86% (CI 83–88) in wave 2. Patients within each ISARIC4C risk group exhibited markedly different outcomes within all cohorts (training $p < 0.0001$, wave 1 $p < 0.0001$, wave 2 $p < 0.0001$) (**Figure A.1**). The relative frequency of ISARIC4C risk groups was comparable in the training and wave 1 cohorts, whereas a notable increase in the proportion of low and intermediate risk patients was seen in wave 2.

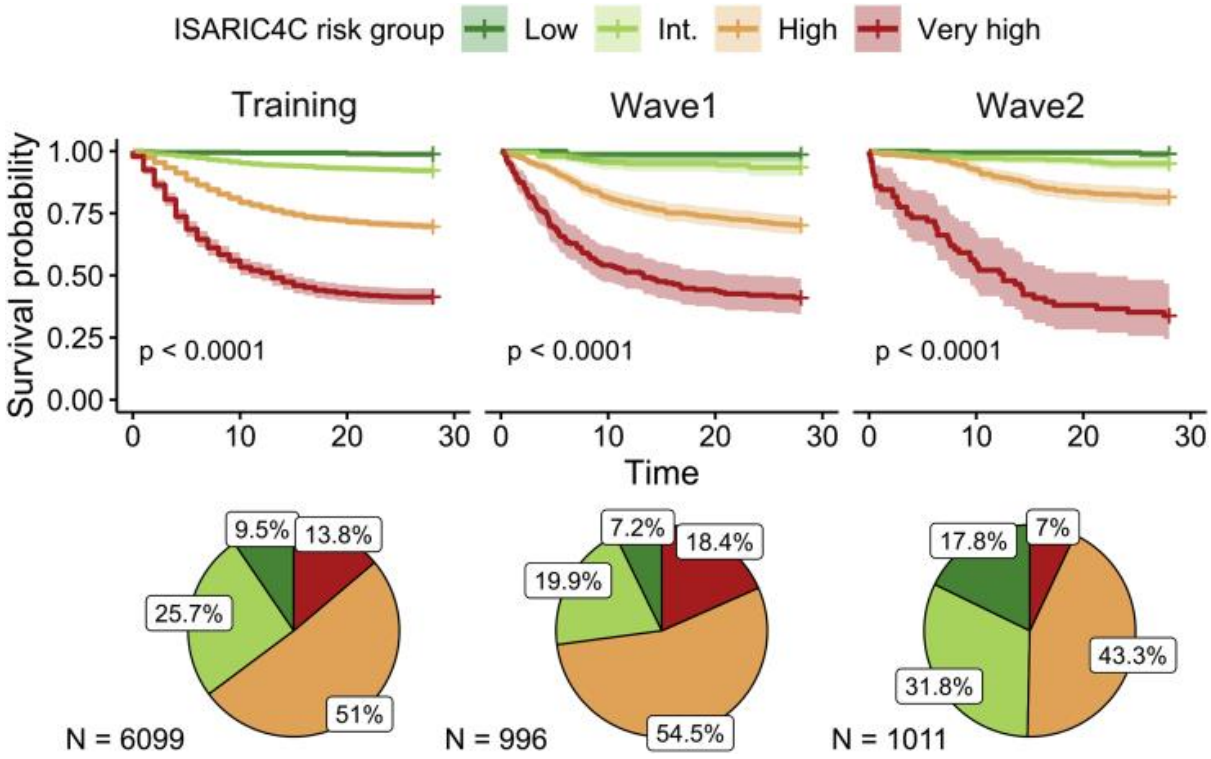


Figure A.1 Survival probability by ISARIC4C risk group. KM survival curves (95% confidence interval) and composition of cohorts by risk group.

Low-risk patients had a consistently high survival probability of 99% (training CI 98–100; wave 1 CI 96–100; wave 2 CI 97–100). For the intermediate risk group, the survival probability was 92% (CI 91–94) in the training cohort, 93% (CI 90–97) in wave 1 and marginally higher in wave 2 at 95% (CI 93–97). High risk patients had a lower survival probability of 70% in both the training (CI 68–71) and wave 1 cohorts (CI 66–74), but this improved in wave 2 to 82% (CI 78–85). However, patients in the very high-risk group had a consistently poor prognosis in each cohort. The survival probability was 41% (CI 38–45) in the training cohort, 40% (CI 34–49) in wave 1, and 34% (CI 23–47) in wave 2.

Taken together, these data indicate that although overall prognosis improved for patients in wave 2, this was driven mainly by improvements in the outcome of intermediate and high-risk patients, with no apparent improvement for those in the very-high risk group.

A.5.2 UMAP transformation and GMM clustering of clinical variables identifies distinct patient groups

28 demographic and laboratory variables were obtained for each patient on the day of entry to hospital and used for inclusion within the UMAP analysis (**Table A.1**). Patient subgroups were identified within the UMAP embedding using GMM clustering (**Figure A.2A**).

Table A.1 Characteristics of patients and clinical variables within the cohorts. Variables summarised by median (continuous) or count (categorical) by cohorts. Variables labelled with a ✓ were included as input into UMAP analysis.

	Cohort			*Adj. P value*		
Variable[1]	**Training**	**Wave 1**	**Wave 2**	**Wave 1 v Training**	**Wave 2 v Training**	**UMAP Input**
Patient number	6099	996	1011			
Age (years) [IQR]	73.00 [58.00, 83.00]	70.00 [56.00, 83.00]	64.00 [50.00, 79.00]	0.04	0.01	✓
Male (%)	3361 (55)	564 (57)	525 (52)	0.97	0.16	✓
BMI (kg/m2) [IQR]	26.84 [23.31, 31.03]	27.78 [24.22, 32.73]	28.02 [23.95, 32.87]	0.01	0.01	✓
28-day deaths (%)	1570 (26)	284 (29)	146 (14)	0.22	0.01	
ISARIC4C Score [IQR]	10.00 [7.00, 13.00]	11.00 [8.00, 14.00]	9.00 [5.00, 11.00]			
Glasgow coma score [IQR]	15.00 [15.00, 15.00]	14.00 [9.00, 15.00]	15.00 [14.00, 15.00]			
Clinical frailty scale [IQR]	4.00 [2.00, 6.00]	4.00 [2.00, 6.00]	3.00 [2.00, 4.00]	0.02	0.01	✓
Cough (%)	4259 (70)	529 (58)	386 (38)	0.01	0.01	✓
Delirium (%)	1160 (20)	78 (9)	28 (3)	0.01	0.01	✓
Fever (%)	3212 (53)	454 (50)	303 (30)	0.41	0.01	✓
Cancer (%)	649 (11)	123 (12)	110 (11)	0.35	1.00	✓
COPD / Sleep Apnoea / Asthma (%)	1579 (26)	297 (30)	271 (27)	0.04	1.00	
CVD (%)	3033 (50)	531 (53)	293 (29)	0.12	0.01	✓
Dementia (%)	934 (15)	343 (34)	236 (23)	0.01	0.01	✓
Diabetes (any) (%)	1794 (29)	356 (36)	318 (32)	0.01	0.42	
Alt (U/L) [IQR]	24.00 [16.00, 41.00]	25.00 [16.00, 39.25]	25.00 [16.00, 41.00]	1.00	1.00	✓
Base excess [IQR]	0.20 [-2.00, 2.30]	-0.30 [-2.50, 2.00]	0.10 [-1.70, 1.70]	0.04	0.13	✓
Chest x-ray imaging (%)				0.01		
Appeared clear	1604 (29)	192 (24)				
Local consolidation	3226 (59)	229 (28)				
GGO / bilateral infiltration	637 (12)	389 (48)				
CRP (mg/L) [IQR]	73.00 [27.00, 146.00]	100.00 [43.00, 176.00]	55.00 [14.00, 118.50]	0.01	0.01	✓
Diastolic BP (mmHg) [IQR]	75.00 [66.00, 84.00]	75.00 [66.75, 84.00]	77.00 [69.00, 84.00]	1.00	0.01	✓
EGFR (mL/min) [IQR]	72.08 [47.41, 90.00]	68.00 [38.00, 90.00]	82.00 [58.00, 90.00]	0.01	0.01	✓
Fraction of inspired O2 (%) [IQR]	0.21 [0.21, 0.36]	0.21 [0.21, 0.21]	0.21 [0.21, 0.21]	0.01	0.01	✓

Hydrogen ion conc.(pH) [IQR]	7.41 [7.36, 7.45]	7.40 [7.36, 7.44]	7.40 [7.37, 7.44]	0.08	0.23	✓
Haemoglobin [IQR]	129.00 [114.00, 142.00]	129.00 [111.00, 144.00]	131.00 [117.00, 143.00]	1.00	0.16	✓
HCO3 [IQR]	24.30 [22.00, 26.50]	24.60 [21.85, 27.30]	24.60 [22.30, 26.70]	0.41	0.15	✓
Heart rate/min. [IQR]	90.00 [78.00, 104.00]	90.00 [79.00, 103.75]	88.00 [76.00, 102.00]	1.00	0.03	✓
Lactate (mmol/L) [IQR]	1.57 [1.13, 2.20]	1.69 [1.27, 2.29]	1.73 [1.36, 2.33]	0.01	0.01	✓
*Lymphocytes (10*9/L) [IQR]*	0.91 [0.62, 1.34]	0.92 [0.64, 1.30]	1.10 [0.73, 1.50]	1.00	0.01	✓
N:L ratio [IQR]	6.00 [3.47, 10.62]	6.00 [3.68, 10.49]	4.81 [2.86, 8.51]	1.00	0.01	✓
*Neutrophils (10*9/L) [IQR]*	5.63 [3.80, 8.50]	5.80 [4.05, 8.35]	5.20 [3.60, 7.97]	0.81	0.02	
O2 saturation (%) [IQR]	96.00 [93.00, 97.00]	96.00 [94.00, 97.00]	96.00 [94.00, 98.00]	0.03	0.01	✓
pCO2 [IQR]	4.63 [4.06, 5.38]	5.40 [4.70, 6.40]	5.40 [4.70, 6.10]	0.01	0.01	
Respiratory rate/min. [IQR]	20.00 [18.00, 25.00]	20.00 [18.00, 25.00]	19.00 [17.00, 23.00]	0.07	0.01	✓
Systolic BP (mmHg) [IQR]	130.00 [115.00, 144.00]	128.00 [114.00, 145.00]	127.00 [115.00, 145.00]	1.00	1.00	✓
Temperature (^{o}C) [IQR]	37.10 [36.50, 38.00]	36.90 [36.20, 37.60]	36.60 [36.10, 37.30]	0.01	0.01	✓
Urea (mmol/L) [IQR]	7.50 [5.00, 12.00]	6.80 [4.60, 12.00]	5.70 [4.00, 8.80]	0.06	0.01	✓

[1]Adj., adjusted; BMI, body mass index; BP, blood pressure; comp., complications; conc., concentration; COPD, chronic obstructive pulmonary disease; CRP, C-reactive protein; EGFR, Estimated Glomerular Filtration Rate; GGO, ground glass occlusion; HCO3, bicarbonate; IQR, interquartile range; L, Litre; min., minute; mmol, Millimoles; N, number; pCO2, partial pressure of carbon dioxide; U, units; umol, micromole

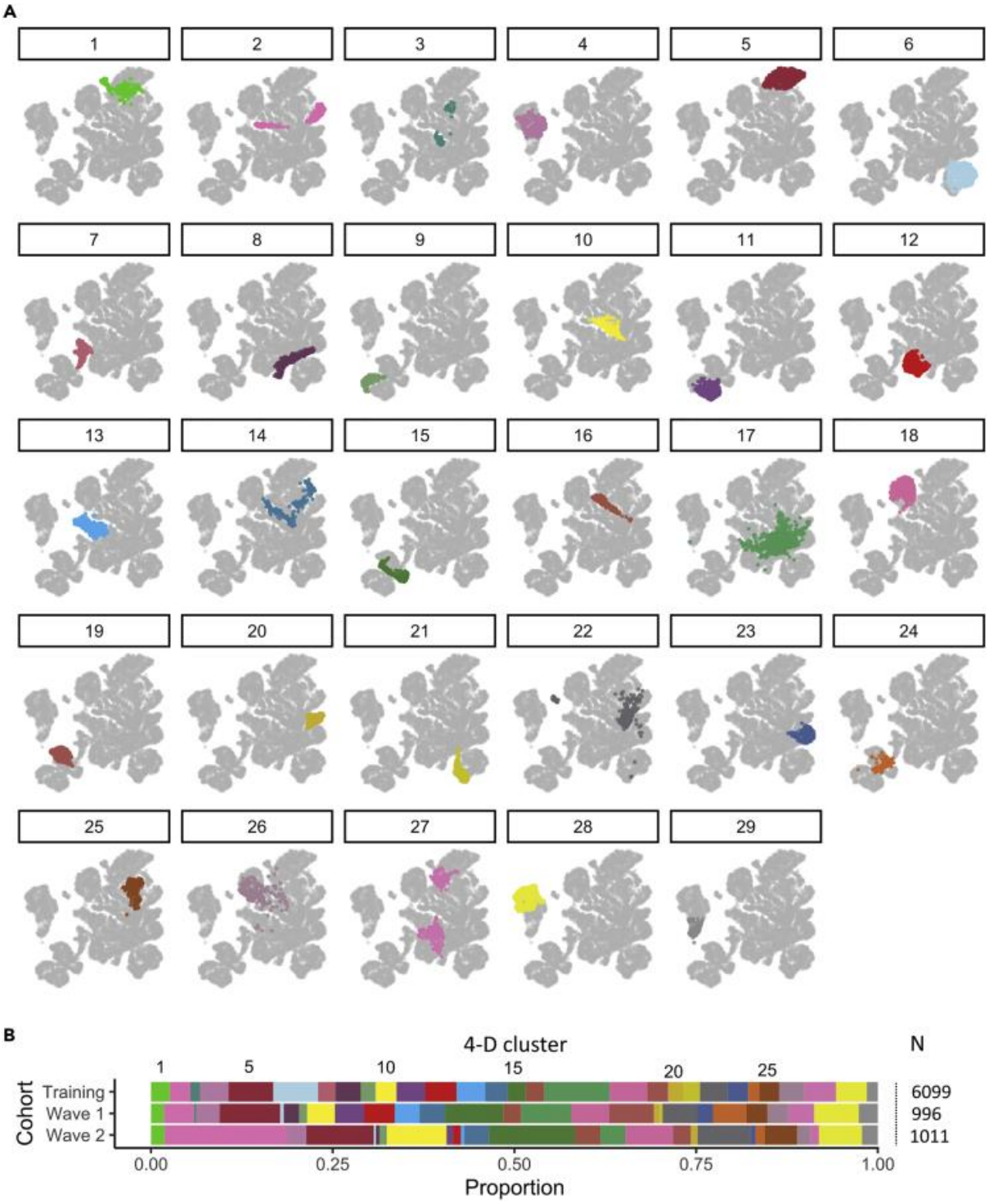


Figure A.2 Visualization of patient clusters within the training cohort. (A) Using a 4-D embedding as input, patients within the training cohort were classified into 29 clusters by fitting a GMM. Each 4-D cluster was then visualised by labelling the training cohort with colours on a 2-D embedding. **(B)** The fitted GMM was then applied to classify patients from waves one and two of the validation cohort into the same set of 4-D clusters. The proportion of patients assigned to each cluster was calculated within each cohort.

A global estimate of intrinsic dimensionality indicated between two and six dimensions in the training cohort (**Figure S. 4A**). The FAMD scree plot indicated an "elbow" point at four or five dimensions (**Figure S. 4B**). Cluster silhouette was relatively stable with increasing dimensionality (**Figure S. 4C**). However, with >4 dimensions BIC comparisons showed greater variance (**Figure S. 4D**). From this, a

UMAP embedding with four dimensions ('4-D embedding') was selected as input for clustering analysis.

The number of mixture components — K — was substantially higher when selected by BIC than by silhouette width (**Figure S. 4E**). As no clear optimal choice was apparent, several models were fitted, and the 28-day mortality rate was compared between clusters to ensure that the range of this value was comparable to that seen with the ISARIC4C score (**Figure S. 4F**). The manual BIC model resulted in the largest range of 28-day mortality (2–65%) where K = 29, in addition to comprising markedly fewer components than the optimal BIC model of K = 49, where the mortality range was 1–63%. In contrast, silhouette models resulted in substantially less diversity in the mortality range between clusters (e.g., 23–38% where K = 3 and 8–45% where K = 9). As such, K = 29 was chosen for onward analysis.

The final model — a GMM with 29 mixture components — was fitted to the training cohort using a 4-D embedding. The median number of patients per cluster was 175 (range 84–548). This model was then applied to patients within the validation datasets (**Figure A.2B**) with comparable representation. Patient distribution within clusters was broadly similar between the training and validation cohorts. Strong cluster separation was observed using silhouette width analysis, although some clusters showed low levels of separation (**Figure S. 5**). In particular, clusters 17, 22, 26, and 14 each had consistently negative silhouette widths whereas clusters 27, 1, 24, and two had a negative silhouette only in the validation cohorts.

A.5.3 Patient clusters defined from the training cohort are predictive of mortality rates within confirmation cohorts

2-D visualization of the 4-D clustering was used to display clusters across the three patient cohorts (**Figure A.3A**). Allocation of sex to the clusters was investigated as this is a significant determinant of mortality. Relative distribution of sex varied between clusters and supported its importance of a key factor in clinical outcome (**Figure A.3B**). We were next interested to assess their potential predictive power for subsequent mortality. The overall 28-day mortality rates for the training cohort and the wave 1 and 2 cohorts were 26%, 29, and 14%, respectively. Mortality at day 28 after admission varied markedly between clusters in the training dataset ranging from 2% within Cluster 18 to 65% in Cluster 24 (**Figure A.3C**).

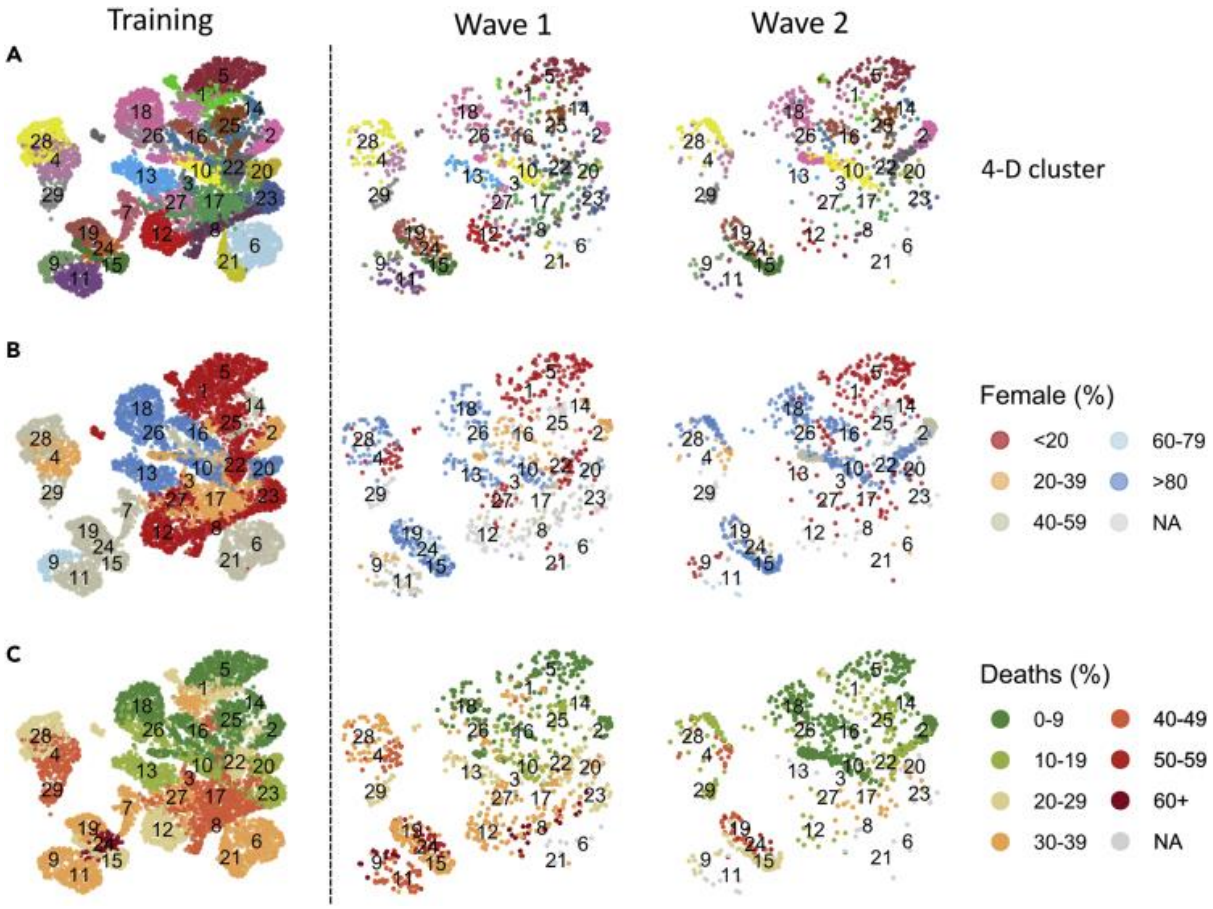


Figure A.3 Sex allocation and Day-28 mortality within patient clusters. (A) Representation of patient clusters in the three patient cohorts. **(B)** Sex distribution within each cluster. **(C)** Relative mortality at day 28 after admission within each cluster. Analysis was performed on clusters which comprised at least 10 patients.

These values were then compared to clinical outcomes for patients assigned to the equivalent cluster within the validation cohorts. Mortality prediction from cluster assignment modeled from the training dataset was found to be highly predictive for patients within the validation cohorts. Indeed, of the 24 clusters within the wave 1 validation cohort, only two had significantly different odds of mortality from the training cohort (**Table A.2**). Specifically, Cluster 12 had a mortality rate of 36% compared to 20% in the training cohort (OR 2.2, p = 0.03), and Cluster 26 had values of 39 vs 19%, respectively (OR 2.8, p = 0.02) (**Figure A.4A**). Cluster-associated mortality rates also showed a high degree of concordance between the wave 2 validation cohort and the training dataset (**Figure A.4B**) with differences observed only for Cluster 2 (2 vs 8%; OR 0.3, p = 0.03) and Cluster 28 (12 vs 25%; OR 0.4, p = 0.04).

These data show a high level of correspondence between the cluster-specific mortality rates in the training cohort and those seen in the validation cohorts indicating confidence that clustering can be generalized to external and temporally independent patient cohorts.

Table A.2 Proportion of Day-28 deaths by cluster. The proportion of deaths by day 28 within each cluster in the training, wave 1 and wave 2 cohorts. Odds ratios are calculated with a Fisher's test of significance. Clusters are ranked by the proportion of deaths observed within the training cohort. N, number; T, training cohort; W1, wave 1 cohort; W2, wave 2 cohort.

Cluster	Rank	N patients			Day-28 deaths (%)			Odds ratio			P value		
		T	W1	W2	T	W1	W2	W1 v T	W2 v T	W2 v W1	W1 v T	W2 v T	W2 v W1
1	13	158	18	19	20	6	21	0.2	1.0	4.4	0.20	1.00	0.34
2	4	171	41	172	8	2	2	0.3	0.3	1.0	0.31	0.03	1.00
3	27	84	3	1	44								
4	26	237	32	24	43	50	42	1.3	1.0	0.7	0.45	1.00	0.60
5	3	375	83	94	5	9	5	1.9	1.2	0.6	0.17	0.78	0.55
6	17	378	5	3	30								
7	19	147	0	0	33								
8	28	210	21	5	49	62		1.7			0.36		
9	20	125	11	10	34	64	20	3.3	0.5	0.2	0.10	0.50	0.08
10	7	175	38	83	15	10	10	0.7	0.6	0.9	0.61	0.33	1.00
11	23	240	40	9	39	45		1.3			0.49		
12	12	264	42	11	20	36	18	2.2	0.9	0.4	0.03	1.00	0.47
13	8	241	34	5	19	24		1.3			0.49		
14	6	188	35	34	9	17	3	2.1	0.3	0.1	0.22	0.32	0.11
15	16	151	80	120	29	39	30	1.5	1.0	0.7	0.14	0.89	0.22
16	2	156	24	35	4	8	6	2.3	1.5	0.7	0.29	0.64	1.00
17	25	548	69	35	43	36	34	0.8	0.7	0.9	0.37	0.38	1.00
18	1	316	52	66	2	6	3	3.8	1.9	0.5	0.09	0.35	0.65
19	22	175	61	25	39	34	43	0.8	1.3	1.5	0.65	0.66	0.47
20	11	131	8	8	19								
21	18	136	4	1	32								
22	14	235	47	74	24	26	13	1.1	0.5	0.5	0.85	0.07	0.15
23	9	170	22	7	19	27		1.6			0.39		
24	29	94	46	13	65	59	54	0.8	0.6	0.8	0.58	0.54	0.76
25	5	166	29	45	8	14	11	1.9	1.5	0.8	0.29	0.55	0.73
26	10	212	28	18	19	39	17	2.8	0.9	0.3	0.02	1.00	0.19
27	21	265	36	12	38	36	16	0.9	0.3	0.4	1.00	0.22	0.29
28	15	259	61	60	25	31	12	1.4	0.4	0.3	0.33	0.04	0.01
29	24	92	26	22	40	27	18	0.6	0.3	0.6	0.26	0.08	0.51

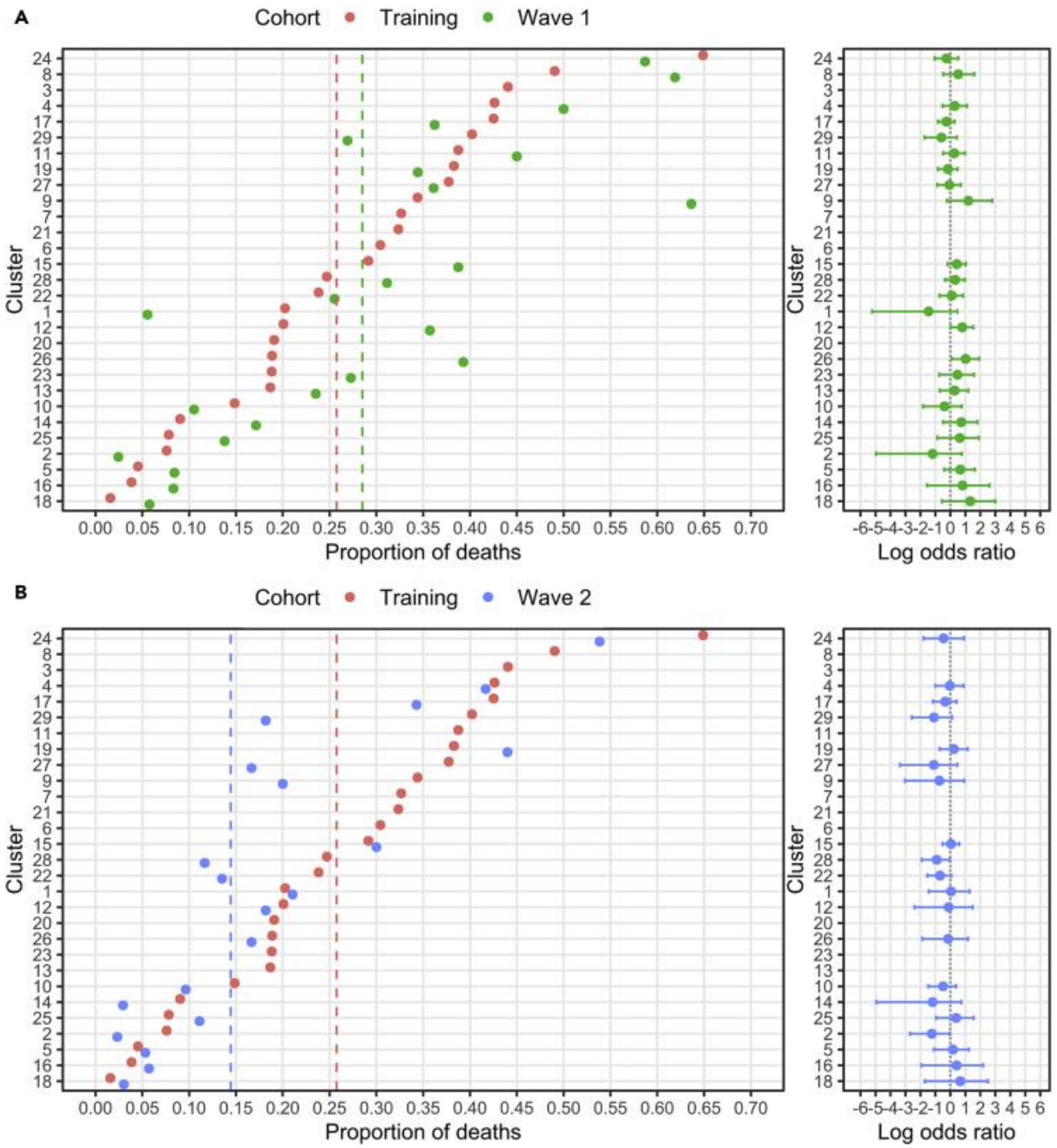


Figure A.4 Day-28 mortality rate within patient clusters across study cohorts. (A) The proportion of deaths by day 28 after admission within each cluster in the training and wave 1 cohorts. **(B)** The proportion of deaths by day 28 after admission within each cluster in the training and wave 2 cohorts. Dashed lines indicate the total proportion of deaths in each cohort. Clusters are ranked by mortality rate observed in the training cohort. Log odds ratios are calculated (95% CI, 1000 bootstraps) such that intervals centred on 0 indicate no difference in the odds of mortality between cohorts. Analysis was performed on clusters that comprised at least 10 patients.

A.5.4 Mortality rates within clusters reveal differential improvement in outcome between wave one and wave two of the COVID-19 pandemic

The period between waves one and two witnessed changes in the clinical management of COVID-19, and the mortality rate of 29% in wave 1 fell to 14% in wave 2. As such, we were interested to determine the utility of cluster modeling to assess relative improvements within different patient subgroups across this period.

Several differences in the demographic profile of patients between the two validation cohorts were apparent. Patients within wave 1 were an average 6 years older than those in wave 2 and had higher rates of comorbidity. Indeed, 53% had cardiovascular disease compared to 29% in wave 2 and rates of dementia or cancer also both fell during wave 2 (34 vs 23%; 12 vs 11%) (**Table A.1**). In addition, the proportion of patients with diabetic complications halved to 5% in wave 2 (**Table S. 1**).

Changes in the profile of laboratory variables were also apparent over time. For instance, the neutrophil:lymphocyte (N:L) ratio fell from 6.0 to 4.8, driven by an increase in lymphocyte count and less marked neutrophilia, whereas estimated glomerular filtration rate (EGFR) and albumin both showed improvements (68 vs. 82 mL/min; 31 g/L vs 34 g/L) (**Figure A.5A**).

We next contrasted cluster-associated mortality rates for patients in wave 1 and wave 2, noting that these were managed in the same hospital. Comparison was possible for 20 clusters, of which, 17 showed a decrease in the odds of mortality, whereas two increased and Cluster 2 remained unchanged (**Figure A.5B**). However, these changes were only significant for Cluster 28, which showed a 70% fall in the odds of mortality within wave 2 (OR 0.3, p = 0.01). Clusters with above average mortality showed little evidence of improvement such as clusters 4, 17, and 24 (OR 0.7, p = 0.6; 0.9, p = 1; OR 0.8, p = 0.76). Strikingly, the number of patients within Cluster 2, which was associated with a very low mortality rate, increased markedly from 4% to 17% between waves 1 and 2.

As such, these findings reveal that the fall in mortality rate in wave 2 was not uniform across clusters. There was a marked improvement in outcomes in the middle of the risk spectrum, whereas high-risk clusters continued to have a bad prognosis. Moreover, low risk patients remained highly likely to survive.

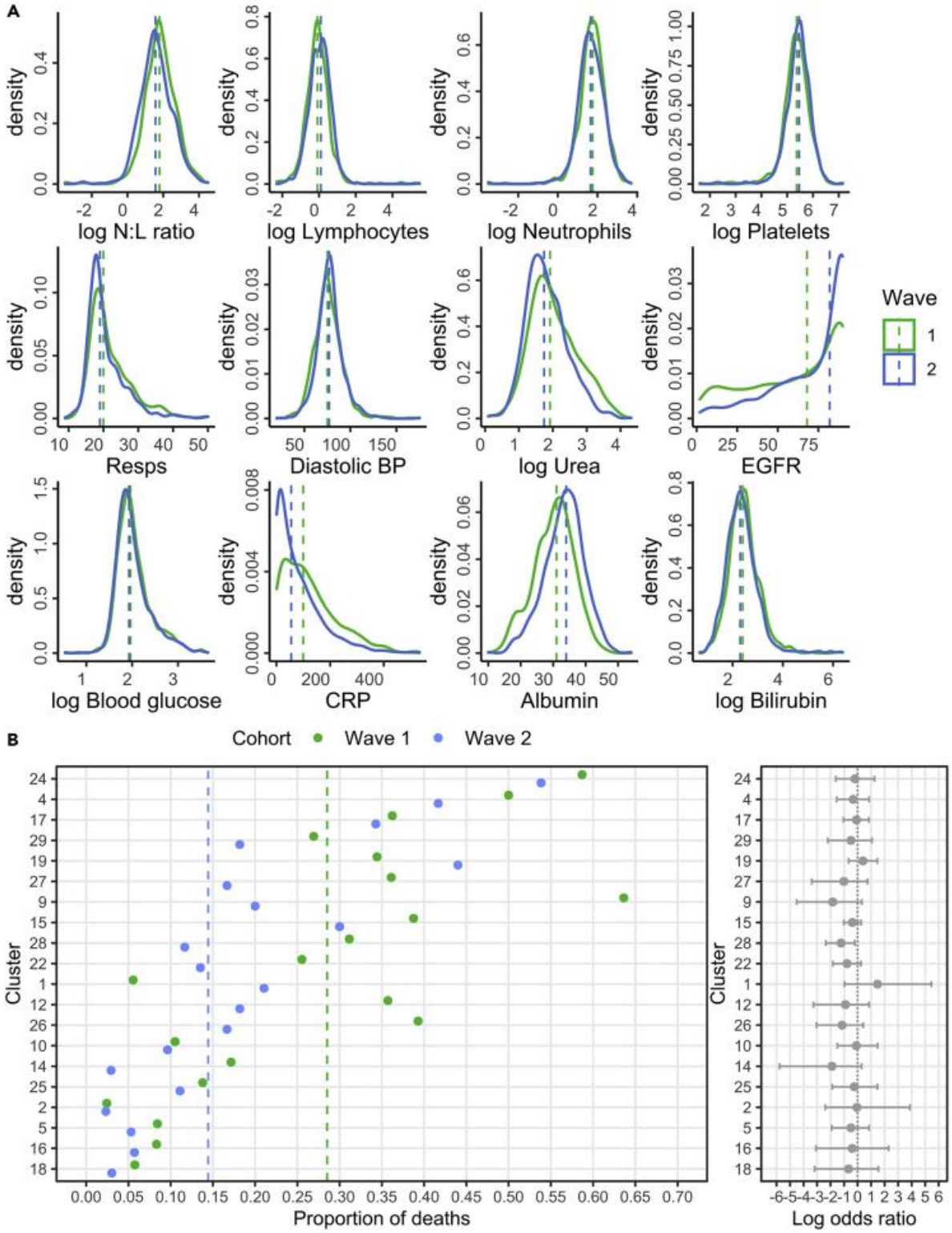


Figure A.5 Distribution of clinical variables and cluster-associated mortality rates within waves one and two of the validation cohorts. (A) Distribution of clinical variables in patients within the validation cohorts. Dashed lines indicate the median values. Missing values were excluded. **(B)** The proportion of deaths by day 28 after admission within the validation cohorts. Dashed lines indicate the total proportion of deaths. Clusters are ranked by mortality rate observed in the training cohort. Log odds ratios are calculated (95% CI, 1000 bootstraps) such that intervals centred on 0 indicate no difference in the odds of mortality between cohorts. Analysis was performed on clusters which comprised at least 10 patients.

A.5.5 Patient clusters reveal interactions between clinical variables that determine patient outcome

As patient clusters varied markedly in relation to 28-day mortality, we were next interested to determine the profile of clinical variables within each subgroup. In particular, the unsupervised nature of the analysis was thought likely to uncover interactions between variables that were not predictable before UMAP transformation and clustering.

Variables that were significantly associated with each cluster were identified and those which could be confirmed in more than one cohort were presented as a word cloud on the UMAP plot (**Figure A.6A**) where the size of the word was scaled according to strength of association with the cluster.

Interesting relationships between variables were observed within the clusters. Increasing age correlated with the clinical severity of cluster-associated outcome, but it was noteworthy that this was not uniform (**Figure A.6B**). Cluster 24, which showed the highest rate of mortality, included many patients with dementia and was associated with low EGFR and high levels of lactate and urea, suggesting a prognostic interaction between impaired renal function and metabolic acidosis (**Figure A.7** and **Figure A.8**). In contrast, Cluster 17 comprised somewhat younger patients with markedly low EGFR and high levels of lactate and urea.

Furthermore, this process of deconvolution of the determinants of clinical risk within clusters does not need to be limited to variables included within the UMAP transformation, allowing additional features associated with each cluster to be assessed in relation to prognostic importance. As an example, the association tests for patients within the Cluster 17 validation cohorts were repeated to include additional variables that were absent in the training cohort (**Table A.3**) and revealed high potassium concentration (K+) as an additional feature of this cluster.

Table A.3 Clinical variables associated with Cluster 17. Clinical variables[1] significantly associated with Cluster 17 by cohort (p=0.001). Variables highlighted in bold were not included within UMAP analysis.

Training	*Wave 1*	*Wave 2*
Base excess↓↓	CVD	Base excess↓↓
CRP↑↑	EGFR↓↓	**Breathlessness**
CVD	H↓↓	**Diabetes (with comp.)**
Diabetes (any)	**K↑↑**	EGFR↓
EGFR↓	pocFiO2↑↑	EGFR↓↓
EGFR↓↓	Ur↑↑	H↓↓
Frailty↕		HCO3↓↓
H↓↓		**Hypertension**
Hb↓↓		Lactate↑↑
HCO3↓↓		O2 sat↓
Lactate↑↑		Ur↑↑
O2 sat↓		
O2 sat↓↓		
pocFiO2↑↑		
Ur↑↑		

[1]Comp., complications; CRP, C-reactive protein; EGFR, estimated glomerular filtration rate; H, hydrogen ion concentration; Hb, haemoglobin; HCO3, bicarbonate; pocFiO2, fraction of inspired O2; O2 sat, O2 saturation; Ur, Urea

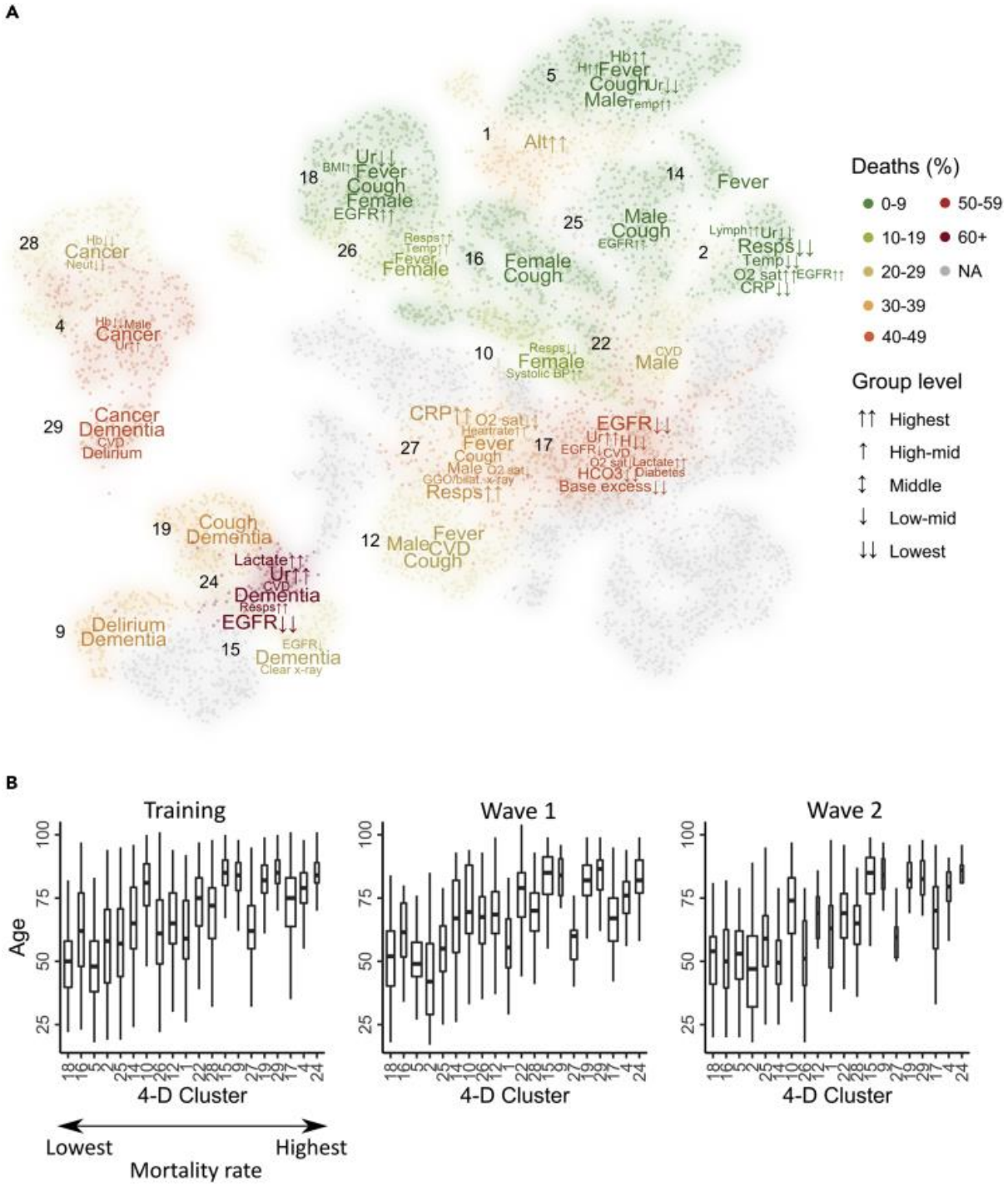


Figure A.6 Cluster characterization reveals clinical variables associated with clusters. (A) Clinical variables associated with patient clusters in the training cohort and at least one validation cohort (p = 0.001). Continuous variables were categorized, and Fisher's tests were applied to each factor level. The strength of the association was used to construct a word cloud with larger words having a stronger association with a cluster in the training cohort. **(B)** Age distribution of patients within clusters. Boxes represent the median and IQR with whiskers extended to the lowest/highest datum within 1.5 IQR of the lower/upper quartile. Clusters are ordered along the x axis in relation to the increasing rate of associated 28-day mortality. Analysis was performed on clusters which comprised at least 10 patients in each cohort.

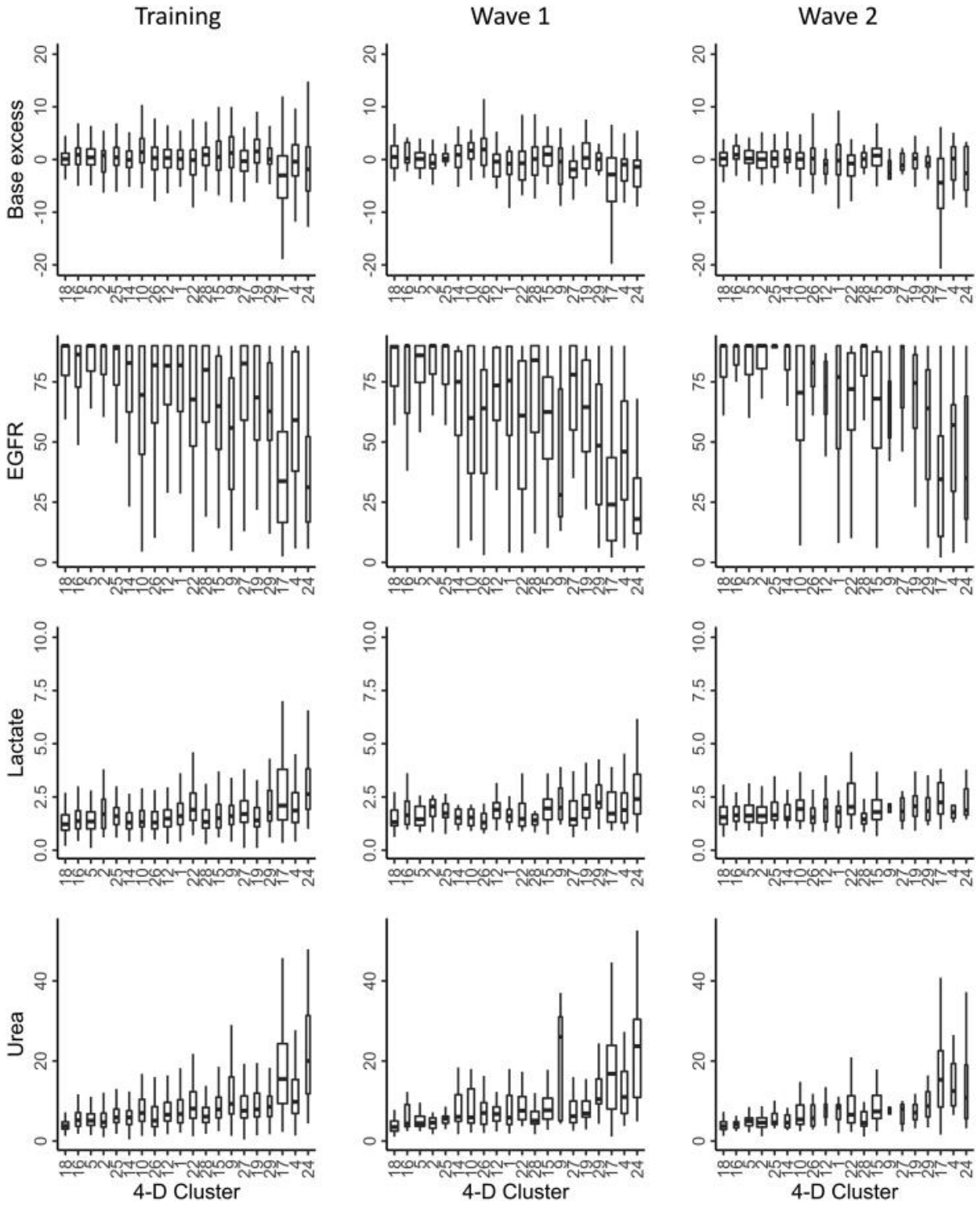


Figure A.7 Distribution of individual clinical variables by cluster. Example of the distribution of four clinical variables by cluster across the three patient cohorts. Base excess, EGFR, lactate, and urea are shown. Boxes represent the median and IQR with whiskers extended to the lowest or highest data within 1.5 IQR of the lower or upper quartiles. Clusters are ordered along the x axis in relation to the increasing rate of associated 28-day mortality. Missing observations were excluded, and the number of observations was used to scale the width of each box. Analysis was performed on clusters which comprised at least 10 patients in each cohort.

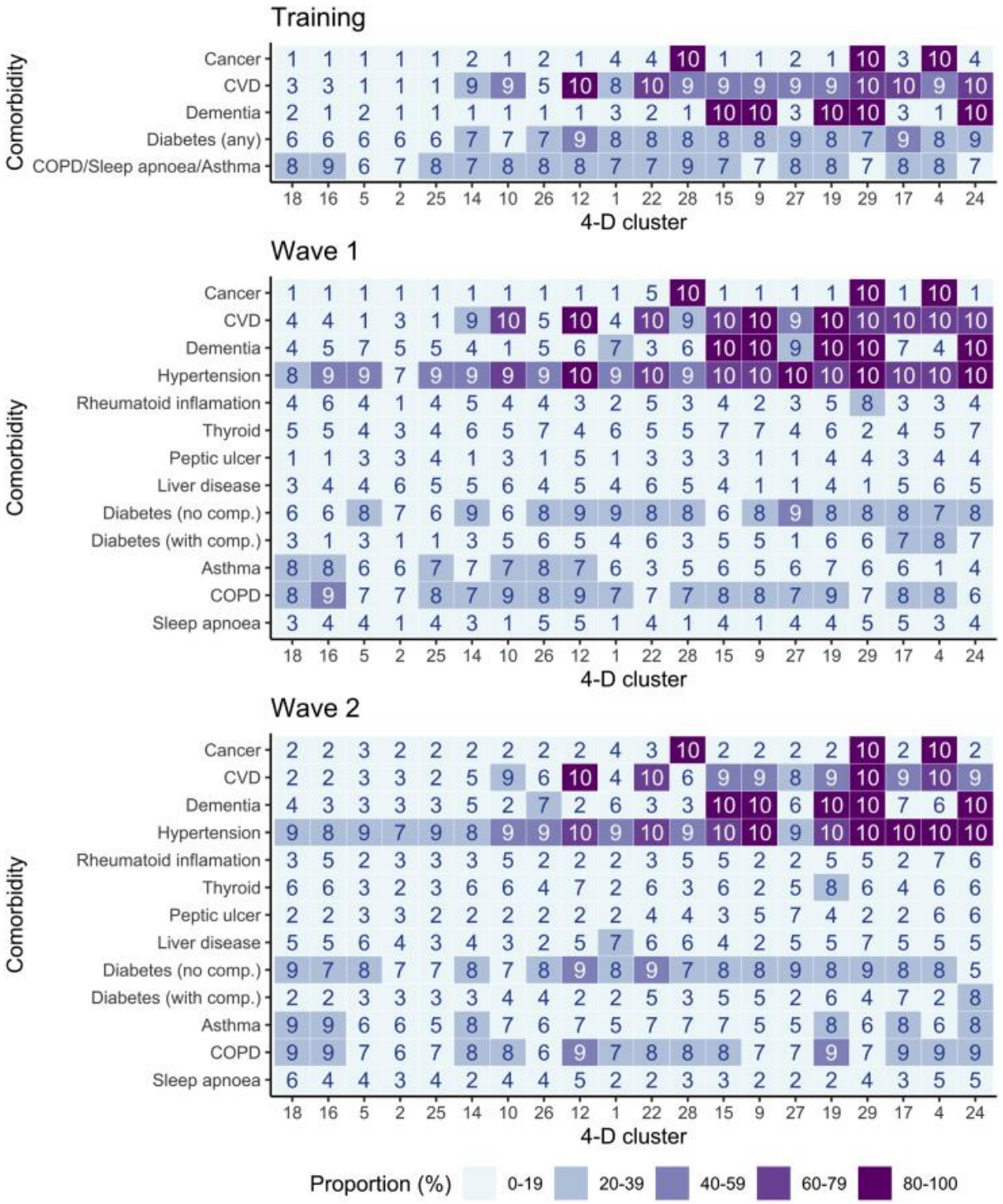


Figure A.8 Proportion of patients with comorbidity within each cluster. The proportion of patients diagnosed with a comorbidity by cluster is represented as a heatmap by decile. Clusters are ordered along the x axis in relation to the increasing rate of associated 28-day mortality. Analysis was performed on clusters which comprised at least 10 patients in each cohort.

A.6 Discussion

Clinical practice is focused increasingly on defining and managing subgroups of patients to facilitate personalized approaches to therapy. Such a process is underway in conditions such as cancer where the use of genetic mutation testing has led to stratification of management (585). However, across most conditions this process is not well advanced. The huge amount of demographic and clinical information that is acquired during patient assessment offers the opportunity to assess discrete subgroups of patients, but a major challenge has been the complexity of using such large datasets with the associated high-level dimensionality that they bring. Here, we used UMAP dimension reduction and GMM clustering to define subgroups of patients with acute COVID-19. We found that clusters were consistent between different populations, are predictive of subsequent mortality, and can be used to uncover unexpected relationships between clinical features.

We chose to use UMAP for dimensionality reduction; this uses topological data analysis and a cross validation optimization process to create a lower dimensional embedding, which retains much of the original data structure. This is an alternative approach to tSNE analysis but is scalable to larger datasets and more effectively captures local as well as global data structure (295). A key advantage of UMAP is that the transformation can be applied to embed new observations into the same latent space, allowing it to be built into a model development pipeline. UMAP has been applied to a range of clinical and biological analyses, including stratification of patients with ALS (305), polygenic risk prediction (586), and single cell analyses (299).

(305) demonstrated that UMAP is capable of stratifying patients into risk groups and used dimension reduction to develop a 1-year mortality prognostic model based on the distribution of patients on a 2-D embedding and comparison of groups based on coordinates on a grid. We extended this approach by dividing the embedding into subgroups through unsupervised clustering. UMAP preprocessing can also improve the results of clustering algorithms (559, 587). The combination of risk-stratification and improved clustering thus makes UMAP-assisted clustering a powerful approach for the detection of novel clinical risk groups.

The default output from UMAP — a 2-D embedding — is ideal for visualization purposes. However, when applying UMAP for nonlinear dimension reduction, the choice is less clear. Our literature review did not indicate an established method to select dimensionality in this context. This decision will depend on both the complexity of the data and the ability of UMAP to represent topological structure on a lower-dimensional embedding. A qualitative approach was therefore undertaken, and a 4-D embedding was selected based on a global estimate of intrinsic dimensionality and assessment of the impact of dimensionality on the clustering model. Other parameters such as the number of nearest neighbours and minimum distance were fixed in our analysis.

Unsupervised clustering analysis was used to label patients by cluster given their distribution in UMAP space. BIC and silhouette width are validated methods for selecting the number of clusters and were adopted in this approach (284, 574, 588, 589). Selecting the number of clusters is a widely debated topic (588, 590-592). A useful property of a GMM is that it can be used to approximate any other probability density function, given a sufficient number of mixture components (593). This allows the model to fit clusters which follow a non-Gaussian distribution. However, in this context the model may require multiple mixture components to fit a single cluster. BIC will optimize the number of components to best fit the underlying data distribution rather than the number of distinct clusters. When the assumption is made that each component corresponds to a single cluster of patients, BIC may lead to overestimates in the true number of clusters (588, 594). Silhouette, on the other hand, measures the compactness within each cluster and the separation between them (284). Using

silhouette to select the number of components therefore ensures the most distinct clustering configuration but can fail to give insight into clustering quality if cluster distributions require multiple components to be fitted (588).

In our analysis, the optimal BIC indicated 49 components but began to plateaux beyond 29. In contrast, the maximum silhouette corresponded to three components but showed a secondary peak with 9. This large discrepancy suggested BIC may have somewhat overestimated the number of clusters, whereas silhouette provided a relative underestimate. We therefore compared the diversity of 28-day mortality risk between clusters, to optimize the potential clinical value of the modeling and found that the model with 29 components resulted in the largest range (2–65%). As such, the decision to use BIC resulted in a model with a greater diversity of mortality risk between clusters, although some clusters had a negative silhouette width, indicative of relatively poor separation. Future work will investigate the use of an entropy criterion to combine components based on the loss of information (594) and examine methods of determining groups from multiple survival curves (595). Other model-based clustering methods such as Hierarchical Density-based Spatial Clustering of Applications with Noise (596) and the use of Dirichlet processes can also be explored (597).

As cluster labels are independent of outcome, unbiased comparisons can be made about the observations assigned to each subgroup. We chose to compare 28-day survival rates but other clinical outcomes, such as subsequent development of long-covid, could be compared without retraining the model. Model-based clustering allows validation of clusters in independent cohorts as the fitted model can be used to cluster new data. Indeed, we observed a concordance of results between the training and validation cohorts, which indicates that this model could potentially be used for analysis of any COVID-19 dataset provided the same predictors were available.

We applied the clustering approach to clinical variables taken on the day of hospital admission to predict subsequent mortality at day 28. This time point was chosen as most patients who die from acute COVID-19 succumb by 4 weeks after hospital entry, and it was felt that it would be challenging to extend prognostic modeling beyond this time. The associated mortality rates of individual patient clusters were very heterogeneous and ranged from 2% for the 316 patients in Cluster 18 through to 65% for 94 patients within Cluster 24. A further striking feature was that values obtained from the training cohort were strongly predictive of outcomes in the validation cohorts. As such, these findings show clustering of demographic, clinical, and laboratory features taken at the time of hospital entry is predictive of medium-term mortality. This is important as this information could be used to guide appropriate triage and clinical management at an early stage of the patient journey.

It was interesting to compare the mortality rates for patients within wave one and wave two of the pandemic treated at the same hospital. Dexamethasone emerged as a standard of care for all patients with severe COVID-19 who required oxygen therapy in July 2020 and as such would have been a default management approach for such patients in wave 2. The overall mortality rate fell by nearly 50% during this period and lower 28-day mortality rates were seen for many clusters within wave 2, although no clear improvement was seen in the very high-risk clusters of 24, 4, and 17 which included many patients with cancer, dementia, and cardiovascular disease. This reveals that cluster analysis could be of value in assessing the differential impact of new therapies on patient groups and identifying those which represent an unmet need.

Furthermore, this use of an unsupervised machine learning approach to define patient subgroups enables identification of new and potentially unexpected interactions between clinical and laboratory variables. Patients in high-risk groups were characterized by medical comorbidity in association with laboratory variables such as impaired renal function and metabolic acidosis. The interaction between

comorbidities within clusters also revealed a number of interesting features (**Figure A.8**). For example, although patients with dementia and cardiovascular disease were enriched within the high-risk Cluster 24, this combination was also seen in several lower risk groups such as 15, 9, 19, and 29. As such, a striking feature from the analysis was that clusters comprised complex combinations of clinical determinants that would not have been easily predictable from clinical assessment. Ethnicity data was not available for the training cohort and in the validation cohorts >50% were of white ethnicity but this was not a key driver of clustering with only Cluster 15 associated with white ethnicity.

In summary, the application of untargeted machine learning to clinical data collected routinely at hospital entry allowed identification of subpopulations of COVID-19 patients with distinct mortality rates and clinical presentation on the day of admission. These were consistent across different clinical datasets and uncovered prognostic interactions between clinical variables that could influence management decisions. Furthermore, such clusters may also reveal activation of different biological pathways and might therefore uncover new mechanisms of COVID-19 susceptibility and a means to stratify therapeutic interventions in phenotype-informed randomized control trials, such as optimized management of renal impairment, metabolic acidosis, and cardiovascular disease for patients in Cluster 24. This approach has application to a wide variety of clinical conditions and contributes to the growing expectation that artificial intelligence systems will transform healthcare delivery and operate in real time to support clinical decision making in both acute and chronic conditions.

A.7 Limitations of the study

Potential limitations of our study include the fact that the training cohort was derived from several clinical sites and could contain site-specific effects with more exposure to reporting errors. The use of unconstrained covariance matrices in model fitting may limit scalability to larger and more complex datasets. In this context, a diagonal covariance matrix may be more appropriate. In addition, missing information was apparent across the three cohorts (**Figure S. 1**).

A.8 Acknowledgments

Funding for D.G. was provided from an NIHR Senior Investigator Award to PM.UHB data was curated by PIONEER, the Health Data Research Hub in Acute Care. Ethical permissions were provided by the East Midlands Research Ethics Commmittee (20EM0158) and with Confidentiality Advisory Group approval (20CAG0084). MJP was supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. We would like to acknowledge the collaborators for the CovidCollab study: Sarah Richardson, Miles Witham, AGE Research Group, NIHR Newcastle Biomedical Research Centre, University of Newcastle and Newcastle-upon-Tyne Hospitals NHS Foundation Trust, UK; Omar A Abdelwahab, Elsayed M Awad, Ahmed Y Azzam, Ahmed Cordie, Ahmed O Elmehrath, Mostafa El-Shazly, Almontacer EB Masood, Alazher University Hospital, Egypt; Osama MAS Abdulhadi, Hazem Ahmed, Muhammed Elhadi, Ahmed KM Hadreiez, Ahmed A Momen, Mosab AA Shaban, Alkhadra Hospital, Libya; Giuseppe Cecere, Aldo Rocca, Antonio Cardarelli Hospital, Italy; Hossam Aldein S Abd Elazeem, Mohammed H Abdelhafez, Islam A Ahmed, Shrouk M Elghazaly, Helal F Hetta, Mohamed Eltaher AA Ibrahim, Soha M Mohamed, Aliae AR Mohamed Hussein, Mohamed M Moustafa, Mariam Albatoul Nageh, Mahmoud M Saad, Alshaimaa M Saad, Omar Zein Elabedeen, Assiut University Hospital, Egypt; Victoria Cox, Danielle Hunsley, Rebecca Ryall, Kathleen T Shakespeare, Thyn Thyn, Rachael Webb, Barnsley Hospital NHS Foundation Trust, UK; Deepthy Hari Madhavan, Nik Sanyal, Birmingham Heartlands Hospital, UK; Bryony Brown, Matthew Hale, Bradford Teaching Hospitals NHS Foundation Trust, UK; Marie Goujon, Cambridge University Hospitals NHS Foundation Trust, UK; Benjamin Jelley, Cardiff University, UK; Laxmi Babar, Tina Doll, Agnieszka Felska, Daniel N Guerero, Sandeep Karthikeyan, Anne Karunatilleke, Helena Lee, Emma Livesey, Amelia Roberts, Charlotte Roberts-Rhodes, City Hospital, Birmingham, UK; Teresa Perra, Alberto Porcu, Cliniche San Pietro, A.O.U. Sassari, Italy; Antonio Buondonno, Giuseppe Cecere, Aldo Rocca, Department of Medicine and Health Science “V. Tiberio”, University of Molise, Italy; Vesna Hogan, Iain Wilkinson, East Surrey Hospital, UK; Ioannis Baloyiannis, Jiannis Hajiioannou, Konstantinos Perivoliotis, George Tzovaras, General University Hospital of Larissa, Greece; Anna Fleck, Aine McGovern, Glasgow Royal Infirmary, UK; Victoria Gaunt, Gloucestershire Hospitals NHS Foundation Trust, UK; Laura Babb, Emily Bailey, Jay Darley, Ioan M Draghita, Alexander Hickman, Jason Kalloo, Akhil Kanzaria, Katy Madden, Wasim Nawaz, Ambreen Sadiq, Good Hope Hospital, UK; Rifa Cardoso, Margherita Faulkner, William Hurst, Ellen James, Aimee Leadbetter, Jordan Mayer, Tanya Robinson, Emma Stratton, Miriam Thake, Hannah Thould, Hannah Watson, Great Western Hospital, UK; Ravindra Belgamwar, Corrina Bentley, Harplands Hospital, UK; Sarah Freshwater, Health Education West Midlands, UK; Sergio DV Ruiz, Nuria M Sanz, Milagros Carrasco Prats, Pedro VF Fernández, Clara G Francés, Esther M Manuel, Miguel R Marin, Pedro L Morales, Patricia P Pérez, María V Soriano, Ismael Mora-Guzmán, Hospital General Reina Sofía, Spain; Ismael Mora-Guzmán, Hospital Santa Bárbara, Spain; Melanie Dani, Imperial College Healthcare NHS Trust, UK; Fabio Barra, Antonella Ferraiolo, Simone Ferrero, Claudio Gustavino, Chiara Kratochwila, IRCCS Ospedale Policlinico San Martino, Italy; Eric W Etchill, Alodia Gabre-Kidan, Joshua H Gray, Elliott R Haut, Harsha Malapati, Sarah F Rapaport, Kent A Stevens, Dominique Vervoort, Johns Hopkins Hospital, USA; Mohammed A Azab, King Abdullah Medical City Specialist Hospital, Saudi Arabia; Catherine Bryant, Hannah Cheney-Lowe, Catrin Cox, Andrew Crowe, Gordon Dick, Sarah Evans, Patrick CP Hogan, Kar Yee Law, Alexandra Richardson, Fabio Speranza, Kathryn Toppley, Julie Whitney, Eirene Yeung, King’s College Hospital, UK; Mary Ni Lochlainn, Claire Steves, King’s College London, UK; Alexandros Charalabopoulos, Spyridon Davakis, Amalia Karapanou, Theodore Liakakos, Eustratia Mpaili, Maria Mpoura, Michail A Sampanis, Nikolaos V Sipsas, Laiko University Hospital, Greece; Lucy Beishon, Elinor Burn, Parveen Doddamani, Victoria Haunton, Shahriar Kabir, Hannah Shaw, Chloe Warner, Leicester Royal Infirmary, UK; Chee Soo, Maidstone and

Tunbridge Wells NHS Trust, UK; Yasmin K NasrEldin, Minia University Hospital, Egypt; Nourhan AA Ghannam, Minya General Hospital, Egypt; Isobel Sleeman, NHS Grampian, UK; Ravindra Belgamwar, Corrina Bentley, North Staffordshire Combined Healthcare NHS Trust, UK; Ali Ali, Sylvia Amini, James Belcher, Marie Giles, Hayley Jarvis, Nathan Jenko, Suvira Madan, Alexander Noar, Favour Nwolu, Jessica Parkin, Lauren C Passby, Jarita Sivam, Michael Surtees, Joanne Wagland, Ruth West, David Williams, Northern General Hospital, UK; Avinash Aujayeb, Lindsey Dew, Catherine Dotchin, James M Dundas, Elinor Edwards, Georgia F Gilbert, Karl Jackson, Sarah H Manning, Dominic Maxfield, Nicholas Moss, Declan C Murphy, Ellen Tullo, Sarah H Welsh, Northumbria NHS Hospital Trust, UK; Tahir Masud, Nottingham University Hospitals NHS Trust, UK; Mustafa Alsahab, Oxford University Hospitals NHS Trust, UK; Antonio Buondonno, Enrico Pinotti, Policlinico San Pietro, Italy; Francesco Alessandri, Gioia Brachini, Giancarlo Ceccarelli, Flavia Ciccarone, Pierfranco M Cicerchia, Bruno Cirillo, Giorgio De Toma, Giulia Duranti, Enrico Fiori, Giovanni B Fonsi, Pierfrancesco Lapolla, Simona Meneghini, Andrea Mingoli, Francesco Pugliese, Paolo Sapienza, Luigi Simonelli, Martina Zambon, Policlinico Umberto I, Sapienza University of Rome, Italy; Caterina Cattel, Laurenny Guzman, Princess Royal Hospital, King's College Hospital Trust, Surrey, UK; Hannah Dowell, Aina Ibukunoluwakitan, Fawsiya Mohamed, Claire Spice, Amanda Stafford, Queen Alexandra Hospital, Portsmouth, UK; Jolene Atia, Catherine Atkin, Hannah Currie, Felicity Evison, Heena Khiroya, Zeinab Majid, Maria Qurashi, Queen Elizabeth Hospital Birmingham, UK; Siobhan Coulter, Claire McDonald, Georgina Muir, Catherine O'Mahony, Caroline Tait, Queen Elizabeth Hospital Gateshead, UK; Rowan Davies, Katie Honney, Laura Winter, Queen Elizabeth Hospital King's Lynn, UK; Olubayode Adewole, Queen's Hospital Romford, UK; Amir Abdelmalak, Mohammed Ahmad, Muhammed H Ansari, Kingsley Appiah, Rajesh Dwivedi, Hope Elrick, Hedra Ghobrial, Rosie Jackson, Sophie Jeffs, Sasha Jeyakumar, Eleanor Lunt, Bushra Muzammil, Sylvia Pytraczyk, Jonathan Sheldrake, Jennifer Smith, Hannah Tobiss, Mark Vettasseri, Ruth H Willott, Hein Zaw, Queens Medical Centre, Nottingham, UK; Katherine Patterson, Queen's University Belfast, UK; Moulinath Bannerjee, Jean Cummings, Barbara Hart, Tom Maughan, Royal Bolton Hospital, UK; Clare Baguneid, Gabrielle Budd, Lizzie Moriarty, Omoteniola Odutola, Hannah Street, Royal Derby Hospital, UK; Alexis Carr, Royal Devon and Exeter Hospital, UK; Jennifer Pigott, Royal Free London NHS Foundation Trust, UK; Sarah Baldwin, Hannah Bashir, Jake Gibbon, Amy Gray, Grace Lewis, Christina Page, Rosanna Varden, Royal Victoria Infirmary, UK; Anthony Grubb, Elizabeth Holmes, Harjinder Kainth, Natalie McNeela, Lara Reilly, Abigail Reynolds, Mark Whitsey, Royal Wolverhampton NHS Trust, UK; Mertcan Akcay, Yesim Akdeniz, Emrah Akin, Fatih Altintoprak, Zülfü Bayhan, Recayi Capoglu, Hakan Demir, Necattin Firat, Emre Gonullu, Tarik Harmantepe, Baris Mantoglu, Ali Muhtaroglu, Merve Yigit, Yasin A Yildiz, Sakarya Faculty of Medicine, Turkey; Lobna Al-Sodani, Nicole Burden, Evelyn Charsley, Thomas Kneen, Angeline Price, Emma Swinnerton, Salford Royal Hospital, UK; Yen Nee J Bo, Hayley R Boden, Reem Bulla, Alison Eastaugh, Helena Lee, Asma Khan, Mohammed Mubin, Amelia Roberts, Anthony Umeadi, Stephanie Wallis, Megan Williamson, Yu Lelt Win, Sandwell General Hospital, UK; Eltayeb A Ahmed, Abdulmoiz Aljafari, Abdulmalek Aljafari, Abdulkader Mohammad, Sharq Alneel Hospital, Sudan; Manpreet Badh, Amy Birchenough, Nick Coulthard, Alice Devaney, Ratnam Gandhi, Katharine Hood, Samuel North, Martha Pinkney, Ellie Shaw, Elisha Whelan, Solihull Hospital, UK; Adam Seed, Southport and Ormskirk Hospital NHS Trust, UK; Gurinder Dogra, Claire Morris, Rebecca Wright, South Tyneside District Hospital, UK; Stephen Lim, Lia Orlando, Harnish Patel, Prabhleen Puri, Sing Yang Sim, Southampton General Hospital, UK; Carolyn Akladious, Gitanjali Amaratungaz, Taha Amir, Cheran Anandarajah, Rachael Anders, Sally Aziz, Anna Barnard, Monica Bawor, Laura Bremner, Hannah Bridgwater, Hejab Butt, Andra Caracostea, Theodore Chevallier, Victoria Comerford, Jack Cullen, Niamh Cunningham, Daniel Curley, Madeleine Daly, Nikhita Dattani, Benyamin Deldar, Arjun Desai, Nirali Desai, Jugdeep Dhesi, Maria Dias, Hannah C Dooley, Samiullah Dost, Hiren Dusara, Alexander Emery, Cassandra Fairhead, Antia Fernandez, Gracie Fisk, Madeleine Garner, Hannah Gerretsen, Andrew Ghobrial, Zaynub Ghufoor, Deirdre Green, Charlotte Greene, Karla

Griffith, Ayushi Gupta, Patrick Harrison, Aidan Haslam, Torben Heinsohn, Lindsay Hennah, Abigail Hobill, Katherine Hopkinson, Lara Howells, Nicole Hrouda, Irem Ishlek, Rishi Iyer, Nuha Kardaman, Mairead Kelly, Nicola I Kelly, Hesham Khalid, Muhammad S Khan, Haris Khan, Matthew King, Li Kok, Aneliya Kuzeva, Rebecca Lau, Gabriel Lee, Gavriella Levinson, Danielle Lis, Baguiasri Mandane, Jamie Mawhinney, Henry Maynard, Sophie Mclachlan, Michelle Metcalf, John Millwood-Hargrave, Kelvin Miu, Aaliya Mohammed, Hamilton Morrin, Stephanie Mulhern, Daniel Muller, Varun Nadkarni, Hanna Nguyen, Alice O'Docherty, Sinead O'Dwyer, Marc Osterdahl, Ismini Panayotidis, Shefali Patel, Rose Pen-fold, Rupini Perinpanathan, Dina Radenkovic, Thurkka Rajeswaran, Tahmina Razzak, Emily Ross-Skinner, Hazel Sanghvi, Ross Sayers, Luca Scott, Sri Sivarajan, Katharine Stambollouian, Jack Stewart, Amybel Taylor, Hrisheekesh Vaidya, Vittoria Vergani, Madiha Virk, Vaishali Vyas, Eleanor Watkins, Catherine Wilcock, Mettha Wimalasundera, Stephanie Worrall, Natalie Yeo, Humza Yusuf, St Thomas' Hospital, UK; Adam H Dyer, Cliona Ni Cheallaigh, Liam Townsend, St. James's Hospital, Ireland; Jocelyn Amer, Emily Lyon, Michael Sen, Sunderland Royal Hospital, UK; Mohammed Al-Sadawi, Adam Budzikoski, Ishmam Ibtida, Yusra Qaiser, SUNY Downstate Brooklyn, USA; Mohammad T Azam, Asad J Choudhry, William Marx, SUNY Upstate University Hospital, USA; Ahmad Bouhuwaish, Ahmed SA Taher, Tobruk Medical Center, Libya; Nikolaos Georgiou, Jade Man, Paul Reynolds, Benjaman To, Tunbridge Wells Hospital, UK; Fatma D Collins, Sharon Budd, Ellanna Griffin, Yue Guan, Deevia Hanji, Lily Lowes, Awolkhier Mohammedseid-Nurhussien, Farhana Moomo, Olebu Ogochukwu, Katie Thin, University Hospitals Coventry and Warwickshire NHS Trust, UK; Elinor Burn, University Hospitals of Leicester NHS Trust; Terry Hughes, Thomas A Jackson, Laura Magill, Lauren McCluskey, Hannah Moorey, Kelvin Okoth, Rita Perry, Michala Petitt, Thomas Pinkney, Daisy Wilson, University of Birmingham; Grace ME Pearson, University of Bristol, UK; Christopher N Osuafor, Kelli Torsney, University of Cambridge, UK; David Strain, Jane Masoli, University of Exeter, UK; Jenni Burton, Terence Quinn, University of Glasgow, UK; Lucy Beishon, University of Leicester, UK; Joanne Taylor, University of Manchester, UK; Adam Gordon, University of Nottingham, UK; Gilda De Paola, Gaetano Gallo, Giuseppe Sammarco, Giuseppina Vescio, University 'Magna Graecia' of Catanzaro, Italy; Shivam Pancholi, University of Nicosia, Cyprus; Natalie Cox, University of Southampton, UK; Rajni Lal, Western Sydney Local Health District, Australia; Rand A Hussein, Zafaraniyah General Hospital, Iraq.

A.9 Supplemental materials

Table S. 1 Additional variables within validation cohorts summarised by median (continuous) or count (categorical).

	Cohort	
Variable[1]	**Wave 1**	**Wave 2**
N patients	996	1011
Age (years) [IQR]	70.00 [56.00, 83.00]	64.00 [50.00, 79.00]
Male (%)	564 (57)	525 (52)
Ethnicity (%)		
Black	63 (6)	45 (4)
Other	214 (21)	189 (19)
South Asian	161 (16)	267 (26)
White	558 (56)	510 (50)
Liver disease (%)	64 (6)	81 (8)
COPD (%)	269 (27)	257 (25)
Asthma (%)	157 (16)	183 (18)
Breathlessness (%)	551 (61)	429 (43)
Chest pain (%)	39 (4)	65 (6)
Diabetes (no comp.) (%)	250 (25)	270 (27)
Diabetes (with comp.) (%)	106 (11)	48 (5)
Headache (%)	42 (5)	48 (5)
Hypertension (%)	628 (63)	519 (51)
Malaise (%)	183 (20)	171 (17)
New diarrhoea or vomiting (%)	55 (6)	70 (7)
Peptic ulcer (%)	26 (3)	21 (2)
Rheumatoid inflammation (%)	49 (5)	42 (4)
Sleep apnoea (%)	47 (5)	34 (3)
Sputum (%)	84 (9)	72 (7)
Thyroid (%)	102 (10)	82 (8)
Alkaline phosphatase (U/L) [IQR]	81.00 [63.00, 112.00]	82.00 [65.00, 107.00]
Anion gap (mmol/L) [IQR]	19.20 [17.20, 21.40]	19.20 [17.70, 20.90]
Bilirubin (umol/L) [IQR]	11.00 [8.00, 16.00]	10.00 [7.00, 14.00]
Blood glucose (mmol/L) [IQR]	7.14 [6.09, 9.09]	6.90 [5.85, 8.80]
Corrected calcium (mmol/L) [IQR]	2.32 [2.24, 2.42]	2.29 [2.21, 2.37]
*Eosinophils (10*9/L) [IQR]*	0.01 [0.00, 0.06]	0.02 [0.00, 0.08]
HCT (L/L) [IQR]	0.39 [0.34, 0.43]	0.39 [0.35, 0.42]
MCV (fL) [IQR]	88.80 [84.15, 93.20]	86.50 [82.20, 90.70]
*Monocytes (10*9/L) [IQR]*	0.51 [0.33, 0.72]	0.53 [0.35, 0.76]
*Neutrophils (10*9/L) [IQR]*	5.80 [4.05, 8.35]	5.20 [3.60, 7.97]
*Platelets (10*9/L) [IQR]*	214.00 [164.50, 283.00]	235.00 [180.00, 305.00]
Potassium (mmol/L) [IQR]	4.10 [3.80, 4.50]	4.10 [3.80, 4.50]
RDW (%) [IQR]	13.80 [12.80, 15.20]	13.60 [12.70, 14.80]
Sodium (mmol/L) [IQR]	138.00 [135.00, 141.00]	137.00 [135.00, 140.00]
*WBC (10*9/L) [IQR]*	7.50 [5.60, 10.30]	7.20 [5.30, 10.20]

[1]Comp., complications; COPD, chronic obstructive pulmonary disease; fL, femtolitre; HCT, haematocrit; IQR, interquartile range; L, Litre; MCV, mean corpuscular volume; min., minute; mmol, Millimoles; N, number; RDW, red cell distribution width; U, units; umol, micromole; WBC, White blood cell count

Table S. 2 Thresholds and factor labels

Variable[1]	*Breaks / Levels*	*Factor labels*
Age	0, 30, 40, 50, 60, 70, 80, 90, Inf	
Alanine Aminotransferase	0, 55, 100, 8000	↓↓,↕,↑↑
Albumin	0, 25, 35, 8000	↓↓,↕,↑↑
Alkaline Phosphatase	0, 130, 8000	↓↓,↑↑
Anion Gap	0, 6, 16, 8000	↓↓,↕,↑↑
Asthma		
Base Excess	-50, -2, 2.01, 8000	↓↓,↕,↑↑
Bilirubin	0, 21, 8000	↓↓,↑↑
Blood Glucose	0, 7.8, 8.5, 8000	↓↓,↕,↑↑
BMI	-Inf, 18.5, 25, 30, Inf	↓↓,↓,↑, ↑↑
Breathlessness		
Cancer (any malignancy)		
Cardiovascular disease		
Chest pain		
Chest x-ray	Appears clear, Local consolidation, Ground glass opacity / Bilateral infiltrates	Clear x-ray, Local consol. x-ray, GGO/bilat. x-ray
Clinical frailty scale	0, 3, 6, 10	↓↓,↕,↑↑
COPD		
Corrected calcium	0, 2.2, 2.6, 8000	↓↓,↕,↑↑
Cough		
CRP	0, 10, 100, 8000	↓↓,↕,↑↑
Delirium		
Dementia		
Diabetes (no comp.)		
Diabetes (with comp.)		
Diastolic BP	0, 90, 8000	↓↓,↑↑
Eosinophils	0, 0.405, 8000	↓↓,↑↑
EGFR	-Inf, 30, 59, 89, Inf	↓↓,↓,↑, ↑↑
Fever		
Glasgow coma score	3, 15, 8000	↓↓,↑↑
Haemoglobin	0, 115, 154, 8000	↓↓,↕,↑↑
HCO3	0, 22, 29, 8000	↓↓,↕,↑↑
HCT	0, 0.5, 8000	↓↓,↑↑
Headache		
Heart Rate	0, 80, 100, 8000	↓↓,↕,↑↑
Hydrogen ion conc.	0, 7.299999, 7.350001, 7.450001, 8000	↓↓,↓,↑, ↑↑
Hypertension		
Lactate	0, 2.21, 8000	↓↓,↑↑
Liver disease		
Lymphocytes	0, 1.5, 8000	↓↓,↑↑

Malaise		
MCV	0, 80, 96, 8000	↓↓,↕,↑↑
Monocytes	0, 0.2, 0.81, 8000	↓↓,↕,↑↑
Neutrophil:Lymphocyte ratio	0, 2.21, 4.83, 8000	↓↓,↕,↑↑
New diarrhoea or vomiting		
O2 saturation	0, 80, 89, 94, 8000	↓↓,↓,↑, ↑↑
pCO2	0, 4.67, 6.4, 8000	↓↓,↕,↑↑
Peptic ulcer		
Platelets	0, 150, 400, 8000	↓↓,↕,↑↑
pocFiO2	-Inf, 0.28, 0.5, Inf	↓↓,↕,↑↑
Potassium	0, 2.5, 5.3, 8000	↓↓,↕,↑↑
RDW	0, 11.5, 15.5, 8000	↓↓,↕,↑↑
Respiratory rate	0, 20, 8000	↓↓,↑↑
Rheumatoid inflammation		
Sex		
Sleep apnoea		
Sodium	0, 133, 145, 8000	↓↓,↕,↑↑
Sputum		
Systolic BP	0, 140, 8000	↓↓,↑↑
Temperature	0, 37.8, 8000	↓↓,↑↑
Thyroid		
Urea	0, 7.8, 8000	↓↓,↑↑
WBC	0, 3.9, 10.9, 8000	↓↓,↕,↑↑

[1]BP, blood pressure; Comp., complications; conc., concentration; CRP, C-reactive protein; pCO2, partial pressure of carbon dioxide; pocFiO2, fraction of in-spired O2; RDW, red cell distribution width; WBC, White blood cell count

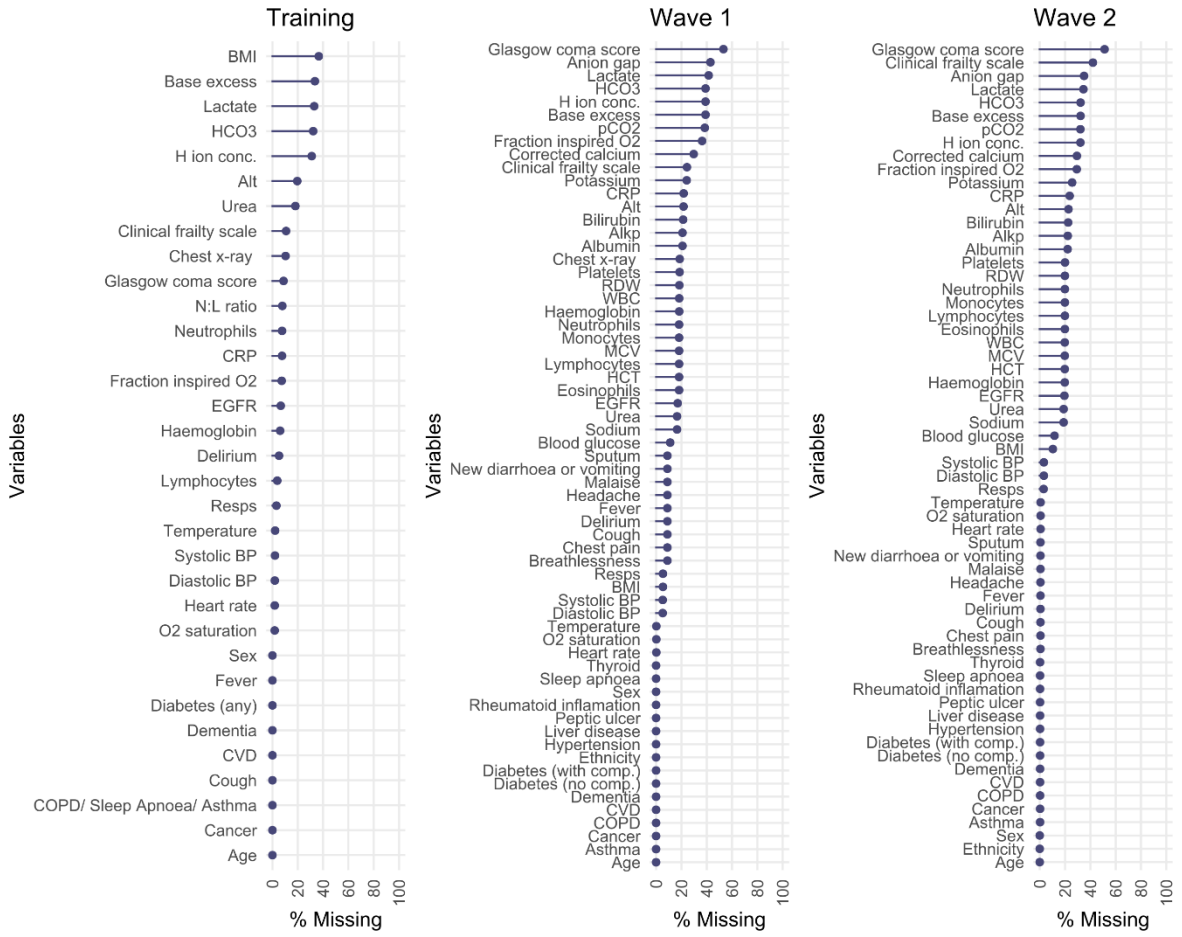


Figure S. 1 Percentage of missing observations. The percentage of missing observations for each variable in the training, wave 1 and wave 2 cohorts.

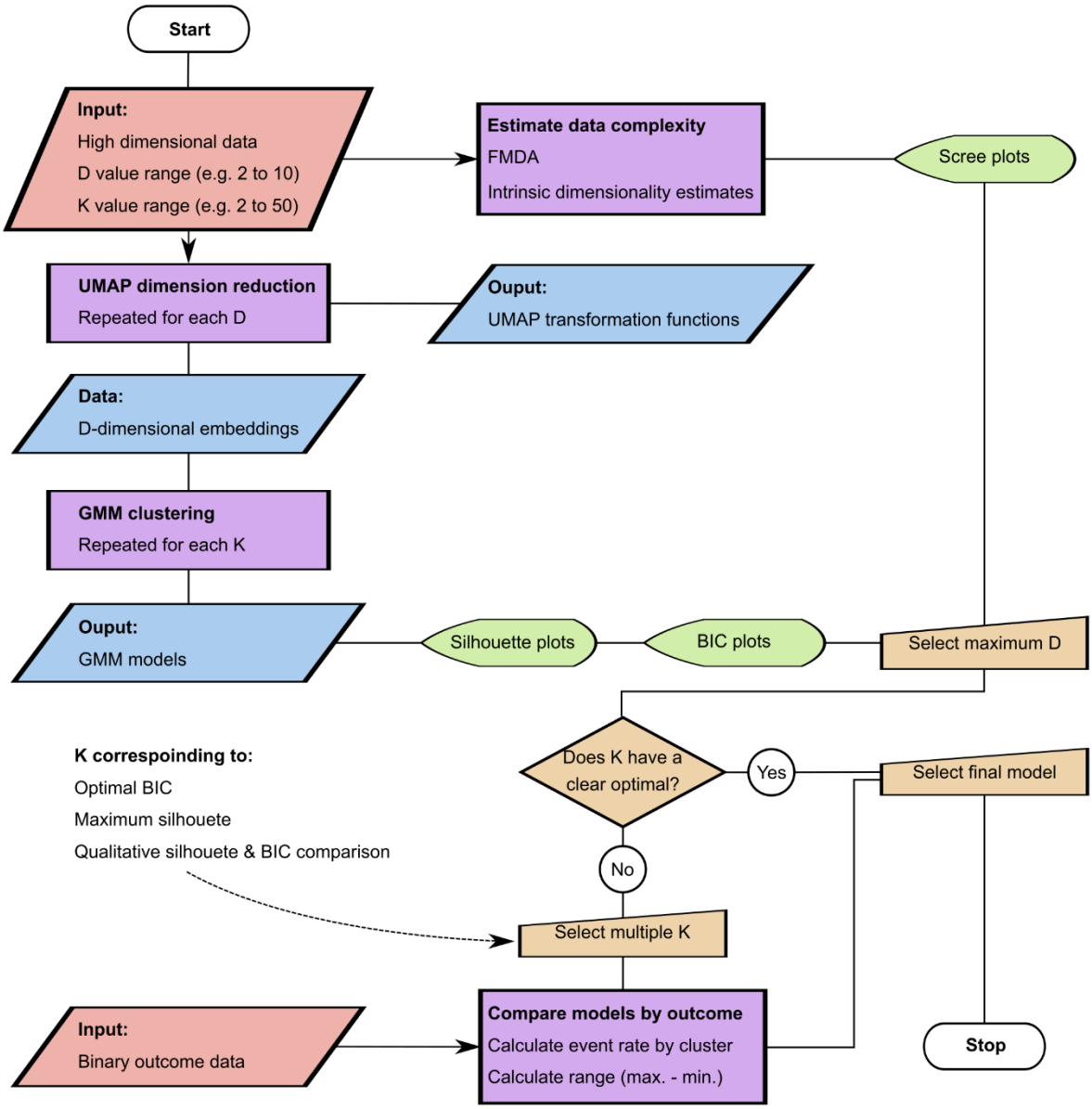


Figure S. 2 Workflow for clustering model development. A clustering model was developed by training a GMM with UMAP dimension reduction as a pre-processing step. The number of dimensions, D, to embed the data and the number of mixture components, K, for GMM fitting were determined as follows. Data complexity as measured by intrinsic dimensionality estimation and FAMD was used to select a range for D. UMAP was applied iteratively with each value of D, and models were fitted to each embedding. The maximum D was selected which did not substantially reduce silhouette width or result in high variability in BIC between models. Possible values of K were selected based on BIC and silhouette width. If both methods indicated a similar K, model development was complete. If multiple values for K were selected, the diversity of mortality rates was compared between models. Mortality rate at day 28 after hospital admission was calculated for K clusters. The model with the largest range (maximum-minimum) was retained.

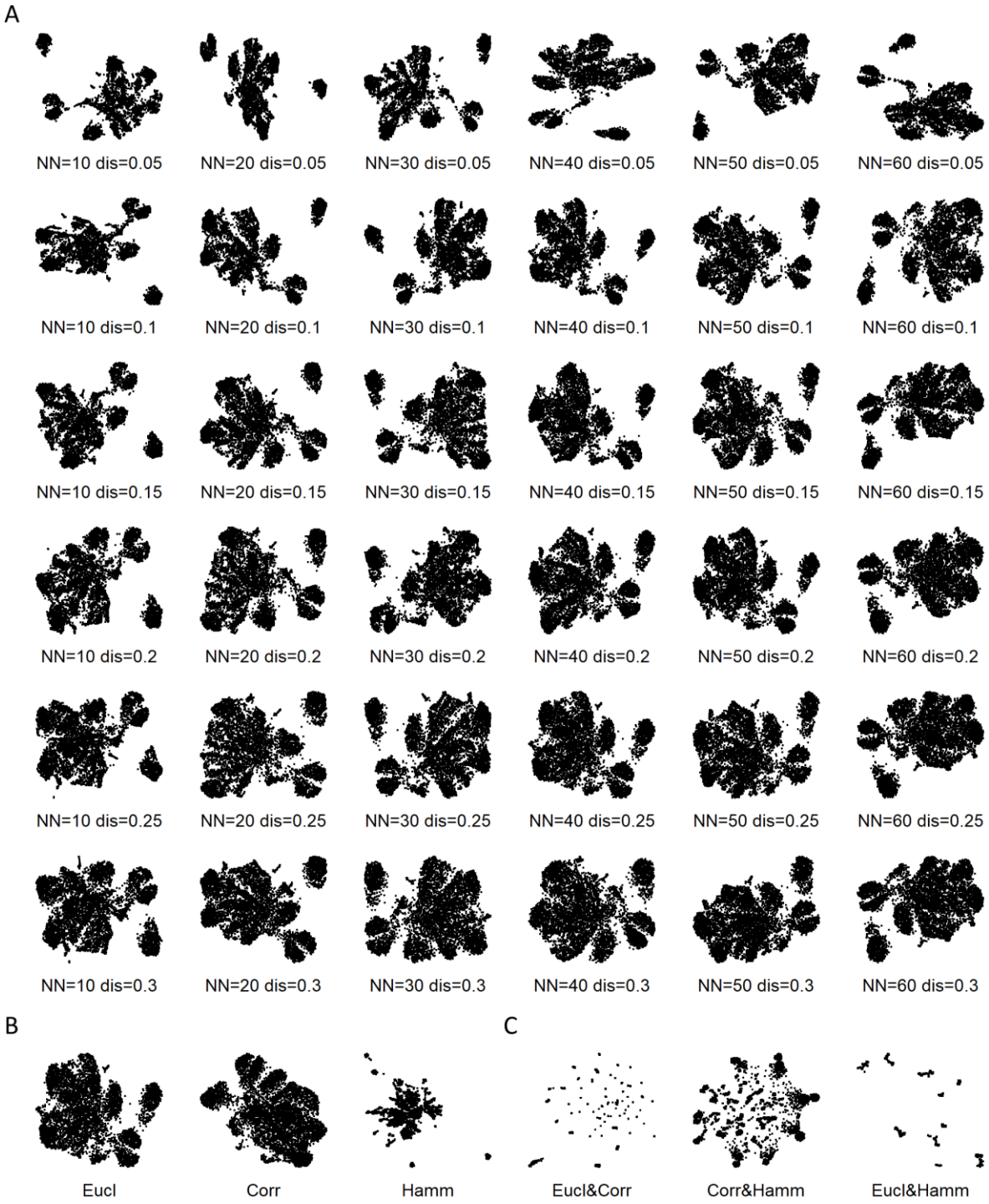


Figure S. 3 Selection of UMAP hyper-parameters. (A) UMAP hyper-parameters, nearest neighbours (NN) = 40 and minimum distance (dis) = 0.25 were selected by visualising a 2-D embedding of the Training cohort using a Euclidean distance metric (Eucl). **(B)** Additional distance metrics, Hamming (Hamm) and Pearson correlation (Corr) were also compared, the examples shown used NN=45 and dis=0.25. **(C)** Combinations of metrics, Eucl/Corr for continuous data with Corr/Hamm were also tested with NN=45 and dis=0.25. However, this would prevent the use of a UMAP transformation with new observations so was deemed unsuitable for model development purposes.

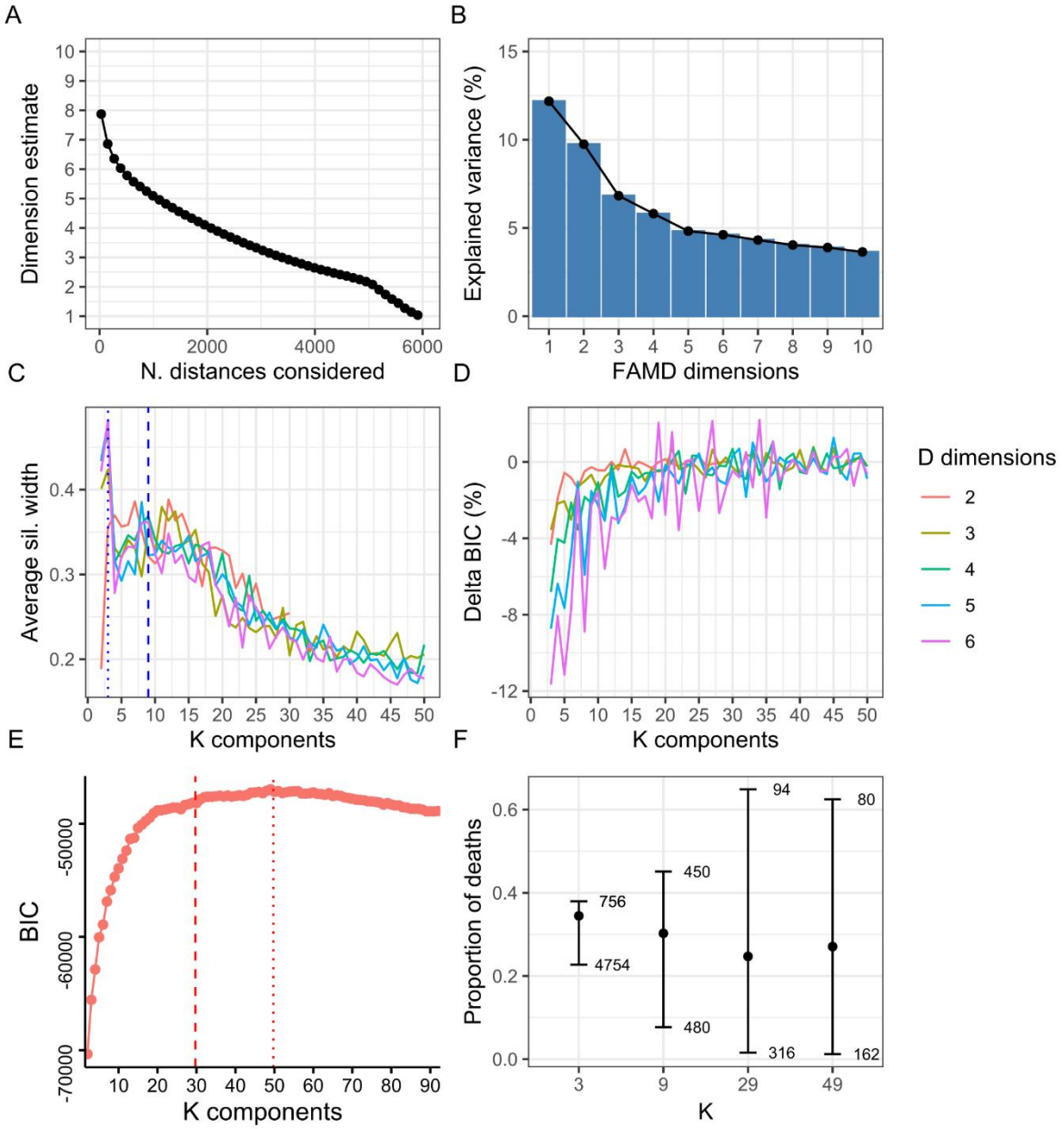


Figure S. 4 Selection of the number of UMAP dimensions and GMM components based on the training cohort. (A) Intrinsic dimension estimation by number (N.) distances measured, estimated globally. **(B)** FAMD scree plot of the percentage of explained variance by eigenvector. **(C)** Average silhouette (sil.) width by number of components in the GMM (K), grouped by the number of UMAP dimensions used in model fitting (D). Dotted and dashed lines indicate K selected by maximum sil. width (K=3) and manually by qualitative assessment (K=9). **(D)** Change in BIC (delta BIC) with increasing K, grouped by D. **(E)** BIC by K for a 4-D embedding. Dotted and dashed lines indicate K selected by optimal BIC (K=49) and manually by qualitative assessment (K=29). **(F)** Maximum, median and minimum proportion of deaths by cluster for selected values of K on a 4-D embedding. Annotated by number of patients in the cluster.

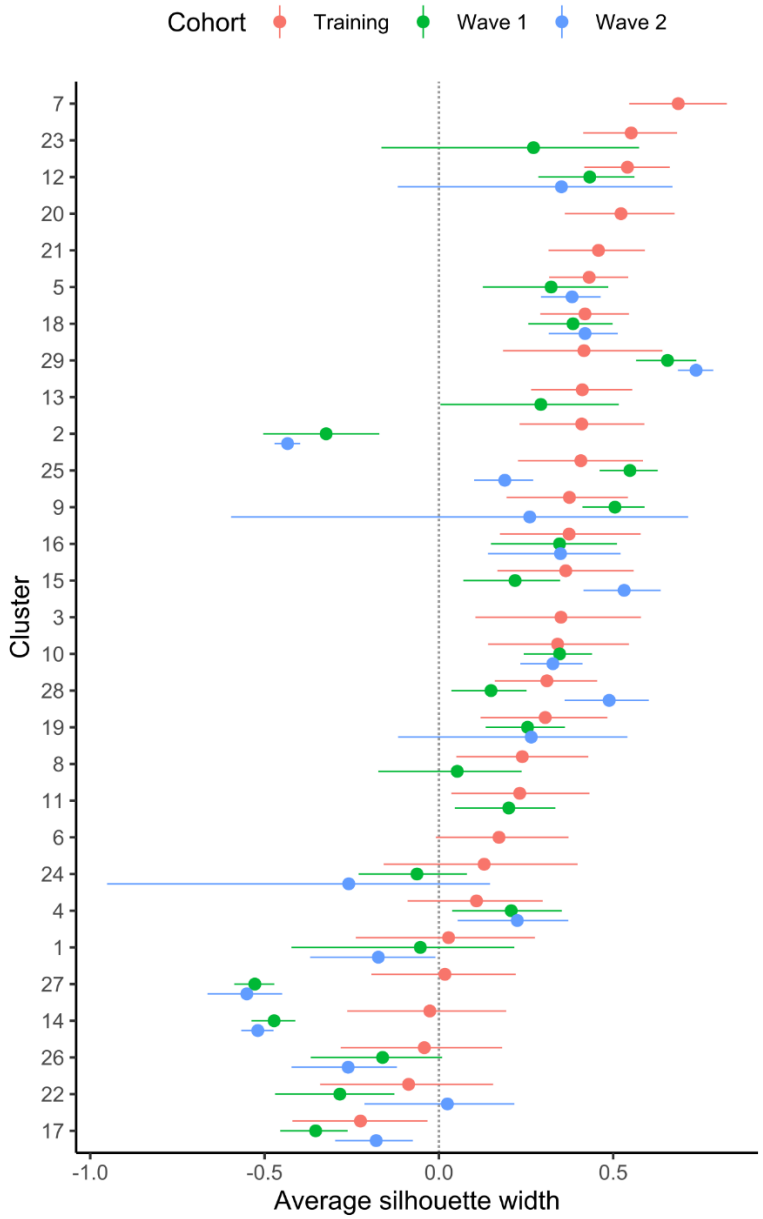


Figure S. 5 Cluster separation by silhouette width. Average cluster silhouette width, a measure of cluster separation, defined for each cluster. This measures how similar observations are within a cluster compared to the most similar cluster, with higher values suggesting better clustering configuration. Single imputation testing was used for the training cohort whilst validation cohorts were tested across 5 imputations and a single estimate was pooled (mean, 95% CI).

A.10 References

References are integrated into the thesis **Bibiliography**.

Appendix B Variable selection

B.1 Importance of variable selection in model development

Numerous variables may be measured in a prognostic study, yet not all will be relevant to the outcome of interest. Determining which variables to include in multivariate analyses is an important decision that can profoundly change the final interpretation of results.

Historically, data would primarily have been collected to validate a pre-existing hypothesis. This has changed since the widespread adoption of high-dimensional "omics" technologies and the abundance of data collected for other purposes (e.g., routine medical observations).

It is now common in biomedical and clinical studies to examine very large numbers of variables (598). As a result, scientific discoveries and novel hypotheses can now be generated directly from the data using data analytics.

However, this abundance is not without its drawbacks. The size of a data set can be measured in two dimensions, the number of observations, and the number of variables. The larger the number of variables relative to the number of observations, the worse the performance of statistical tools such as Cox regression (599-601), due to the so-called "curse of dimensionality"(602). This can lead models to overfit to a particular data set, meaning inferences made from the model may not generalise to new unseen observations.

Even if the ratio of observations to variables is acceptable, the presence of irrelevant or redundant attributes can also harm model development. As models may treat random fluctuations and noise present in the data as learned concepts(603). Collinearity may also exist between variables if they convey similar information.

Variable selection, also known as feature selection, is the process of selecting relevant variables that accurately describe a given problem, whilst discarding irrelevant or redundant features with a minimum reduction in performance(341). This enables simpler, more parsimonious, models to be constructed. Thereby overcoming the issues introduced by high-dimensional data and making the resulting models more interpretable.

B.2 Variable selection methods for survival data

This section will introduce some variable selection strategies compatible with censored survival and competing risks data. Focusing on methodologies utilised in biomarker discovery in allo-HSCT research (349-351) and applied to OLINK proteomics data (352, 353).

B.2.1 Prior knowledge

The most important strategy is to use of clinical knowledge and previous literature to guide variable selection(340). If a variable has an established association with the probability of an event, such as age in a study of all-cause mortality, it is often sensible for the variable to be included in model development.

This approach involves the principal investigator determining a list of variables, based on subject-specific knowledge of the problem being modelled, and on the availability of the data. This selection should not be influenced by associations measured in the data set for model development.

The number of variables to include may be guided by calculating the events-per-variable (EPV) ratio. This is the ratio between the number of observations (events in the case of survival analysis) and the number of variables. Whilst subject to debate, it is often cited that EPV ratios above 10 are desirable to limit overfitting to the observed data(345). Whereas recent simulation studies have argued this

could be relaxed to an EPV of 5-9 without substantially increasing the risk of bias(361). However, the EPV is only a rough guide to aid decision making.

By discussing the known impact of variables, the availability of the data, and the EPV ratio, a subset of variables for model development can be determined.

B.2.2 Filtering based on univariate effects

A common strategy is to filter variables based on a univariate significance criterion. For example, variables with univariate hazard ratios of an arbitrary significance (e.g., P < 0.1) may be included as covariates in a multivariate survival model(604).

This is intuitive, scales well to very large data sets, and reduces the risk of overfitting(345). However, variables with weak individual effects may be excluded which have a larger effect when combined(340). Additionally, this introduces unaccounted multiple testing that can complicate the interpretation of significance from the multivariate model.

B.2.3 Information criteria

Significance criteria are useful to assess if individual variables should be included, whereas information criteria enable the selection of a model from a set of plausible models.

As the number of variables increases, the apparent fit of a regression model will also increase. Information criteria were developed to penalize the apparent model fit based on the complexity of the model (the number of variables). In this context, the model fit is measured by the likelihood function, for an intuitive explanation of the concept of likelihood and its use in model fitting I would point the reader to Alexander Etz 2018 (605).

The Akaike information criterion (AIC) (346, 606) deals with the trade-off between simplicity and having an adequate model fit. This compares the number of parameters estimated in the model k (equal to the number of variables) with the maximum value of the log-likelihood function $\hat{L}$, using the formula

Equation 14

$$AIC = -2\log(\hat{L}) + 2k.$$

AIC estimates the relative loss of information in a model relative to other candidate models (340). The lower the AIC value, the better the perceived quality. This provides a way to compare various candidate models for a specific set of data.

Alternatively, the Bayesian information criterion (BIC) can be used (277). Whilst BIC was developed from a different statistical perspective, where it is assumed that a true model exists amongst the set of models being compared. In practice, the main distinction is that the penalty term is determined by the number of observations n (the number of events for survival models), using the formula

Equation 15

$$BIC = -2\log(\hat{L}) + \log(n) \cdot k.$$

BIC typically results in a less complex model being selected compared with AIC. As given a suitable number of events n, the penalty applied by BIC for complexity is larger than that of AIC.

By comparing the impact of different combinations of variables, information criteria can be used for variable selection.

This can be in the form of “best subset” selection, in which all possible combinations of variables are compared (345). However, there are 2^k possible models with k different variables. Less exhaustive approaches such as forward selection are typically employed. In which variables are added sequentially, or backward selection which reverses the process. The set of variables that result in the model with the lowest AIC or BIC are then selected.

Due to its ease of implementation stepwise selection with AIC or BIC is amongst the most widely used variable selection strategies(340). It provides an automated, and reproducible way to select a set of variables, the relative importance of which can be assessed by the magnitude of the coefficients.

However, there are several disadvantages. Unless applied to very large sample sizes, there can be instability in the variable selection results. Irrelevant features may be retained if there is a large number of candidate variables, and the presence of collinearity can influence selection(340). There is also a relationship between AIC/BIC and significance testing, so it could introduce unaccounted multiple testing (345).

B.2.4 Regularisation

The information criteria discussed apply a penalty term to the apparent model fit. In contrast, shrinkage, also known as regularisation methods, apply a penalty term to the size of individual regression coefficients. This shrinks the size of uninformative coefficients, so they have less weight in the model.

One such method which is useful for variable selection in high-dimensional settings is the least angle selection and shrinkage operator (LASSO)(343). LASSO introduces sparsity into a model meaning some of the regression coefficients are equal to zero. LASSO starts by fitting a model including all candidate variables. LASSO imposes an upper constraint on the sum of squared regression coefficients, this results in shrinkage. The non-zero coefficients after shrinkage correspond to the selected variables. The relative importance of the non-zero coefficients can be assessed by the size of the resulting regression coefficients.

LASSO is an effective variable selection method that can be applied when the number of variables is far in more than the number of observations(345). It is fast, easily automated, and has been extended to both Cox and Fine-Gray regression (342, 354). LASSO is also relatively robust to the presence of correlated variables.

However, LASSO is sensitive to the scale of covariates, so data often needs to be standardised. This may not be optimal if say the data contains a mix of binary and continuous variables. By definition, the coefficients of a LASSO model are biased by the penalty term. When interpretable regression coefficients are of interest, researchers may choose to refit the model post-selection. However, this introduces complications related to post-selection inference(607).

B.2.5 Machine learning methods

Unlike the variable selection algorithms described thus far (which all rely on an underlying regression model), machine learning (ML) methods can account for complex interactions between candidate variables, and non-linear effects on the outcome of interest.

ML methods use function approximation algorithms, such as random forests(608), gradient boosting (609), and neural networks(610), to fit a model which maps the association of variables onto the rate of an event. These are known as supervised classification algorithms as they associate patterns in the data with a set of known outcomes.

A wide array of effective ML-based variable selection methods have been developed for general regression or classification problems(341, 603). However, only a few such methods have been extended for the particular case of survival data, with the complexity introduced by censoring (611, 612). In particular Random Survival Forests (RSF)(356), and boosted Cox regression (613, 614), have shown to be effective tools for high-dimensional variable selection with survival data (348, 612, 615).

RSF and boosted Cox regression are also suitable for variable selection in the presence of competing risks using Gray's subdistribution hazards method (356, 616). A common limitation of machine learning methods is they may be more subject to overfitting, given a limited number of observations.

B.2.5.1 Random Survival Forests

RSF is a form of the random forest algorithm extended for use with censored survival data (347, 608). The algorithm uses an ensemble of decision trees, each of which is generated from a different sample of the candidate variables and observations.

At each node of a decision tree, observations are divided into groups based on a different variable. The order of variables and split points are both optimised to maximise the difference in survival between the resulting groups. For survival data, the log-rank test statistic is maximised (320), for competing risks data Gray's test is maximised (213, 356).

The initial random sampling of variables and observations introduces stochastic differences between each optimised decision tree (hence the term "random"). Individually, each decision tree is a weak learning algorithm with low accuracy (617, 618). However, by combining these decision trees into an ensemble classifier, the resulting model is better than the sum of its parts(618, 619), robust to overfitting, and capable of highly accurate predictions.

Two metrics can be used to assesses the relative importance of variables when applying the RSF algorithm, variable importance (VIMP) and minimal depth. Variable importance (VIMP) is the change in model prediction error when a variable is "noised-up" (356, 608). The larger the VIMP, the more substantial the degradation in accuracy, indicating the variable is important to predicting the event (347). In the case of competing risks, VIMP results are specific to an event.

Whereas minimal depth, which assess the depth in the decision trees a variable was first used to split the data. Variables higher in the decision tree are more predictive, so the smaller the minimal depth the more important the variable (356). Minimal depth is not event specific; however, a threshold can easily be estimated from the model, thus providing a cut-off for variable selection.

Variables are selected if they have a positive event specific VIMP and are above the minimal depth threshold estimated from the model.

RSF provides highly accurate predictions for survival and competing risks data (620). It is capable of detecting complex interactions between variables, providing an effective method for variable selection with high dimensional data. RSF is effective given a relatively small sample size(348). However, variable selection results may be biased towards variables with many potential split points(621).

B.2.5.2 Boosted Cox regression

Boosting is an iterative model-fitting routine used extensively in machine learning applications(609). Boosting is one of the most efficient machine learning methods for regression and classification (622). This approach relies on an ensemble of models, known as weak learners, each of which is learned sequentially. Mistakes of the previously fitted model are used to update the parameter estimates of the next model, iteratively improving accuracy.

Boosting has since been adapted to fit statistical models including the Cox regression model (613, 614, 623) and the Fine-Gray model (613). This approach provides an effective method for high-dimensional variable selection with survival data. As it is based on the Cox model, it also produces interpretable effect size estimates.

The only validated implementation of boosting using Gray's method, CoxBoost, was archived from the R repository on 2020-11-11 (624). As result, the software was unavailable for use in the investigation of competing risks in this thesis.

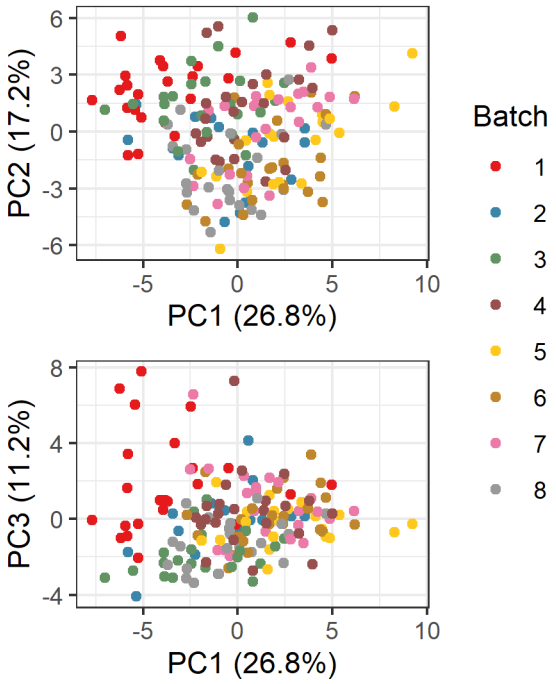


Figure C.1 Principal component analysis (PCA) of CyTOF data. The average marker expression per graft across all markers (graft centroids) visualised by first three principal components, with percentage of explained variance. Grafts are coloured by the batch in which they were subject to CyTOF data extraction.

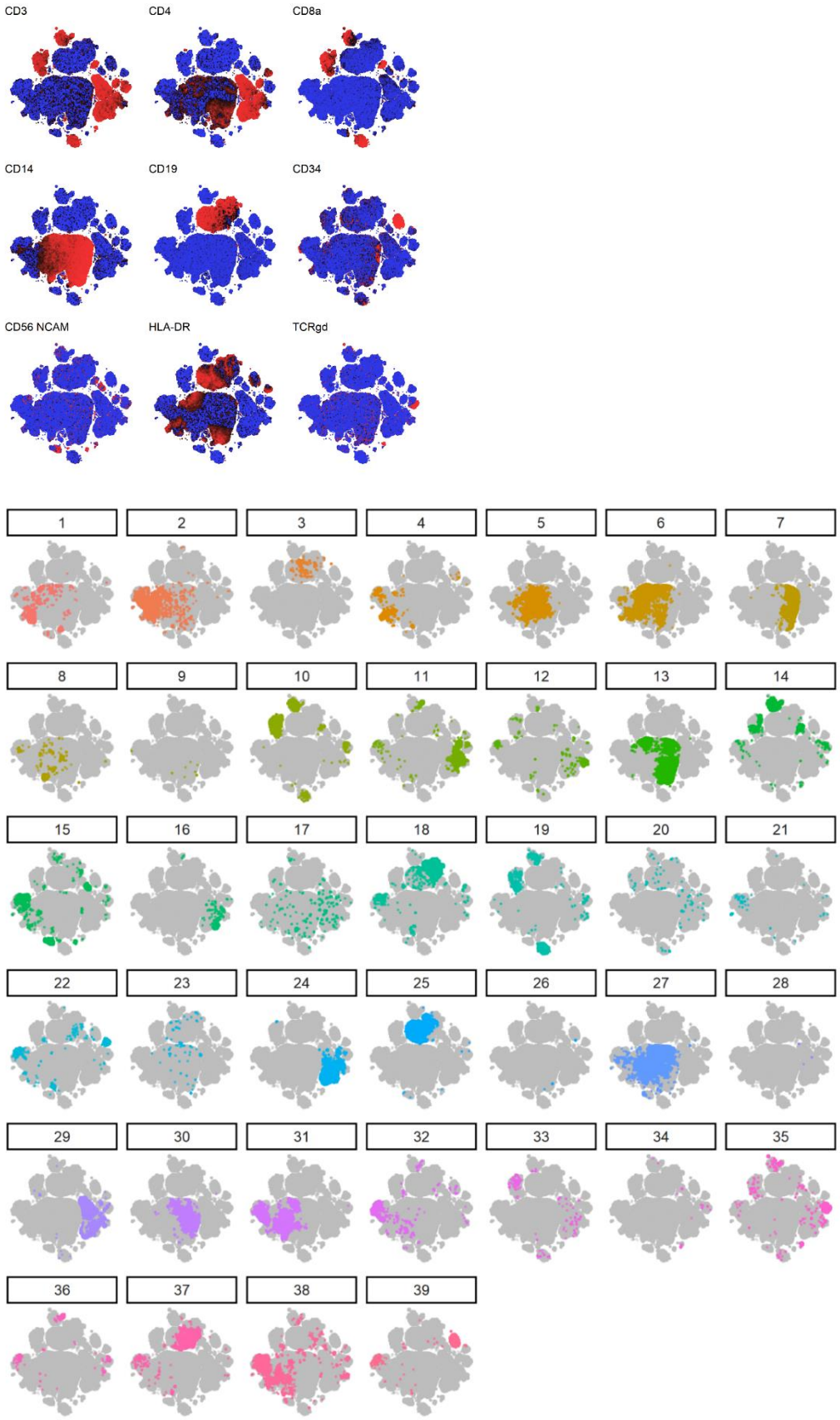


Figure C.2 t-SNE visualisation of cells by marker expression intensity and cluster.

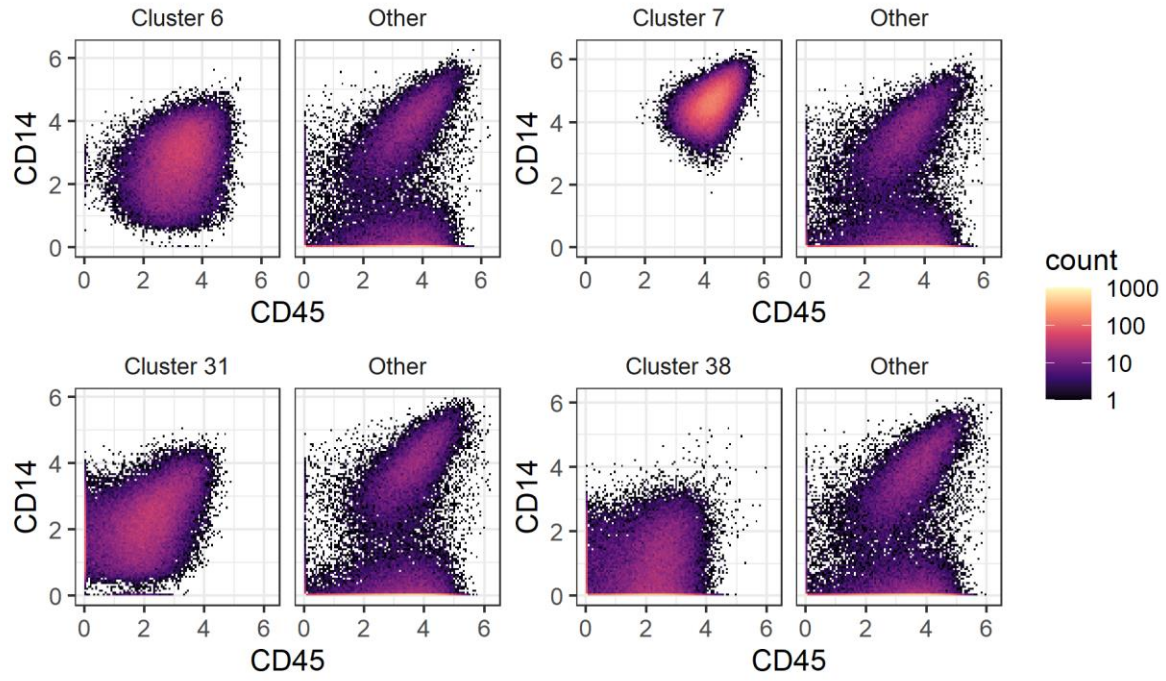


Figure C.3 Marker intensity in cells from Clusters 6, 7, 31, 38 – monocytes. Bi-axial plots of marker intensity by cell count across cell clusters compared to 500,000 cells sampled at random (Other).

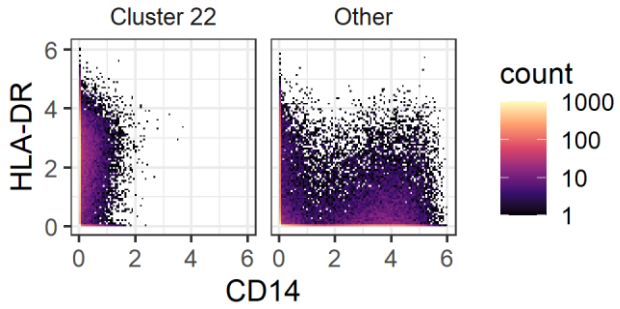


Figure C.4 Marker intensity in cells from Cluster 22 – DCs. Bi-axial plots of marker intensity by cell count across cells in Cluster 22 compared to 500,000 cells sampled at random (Other).

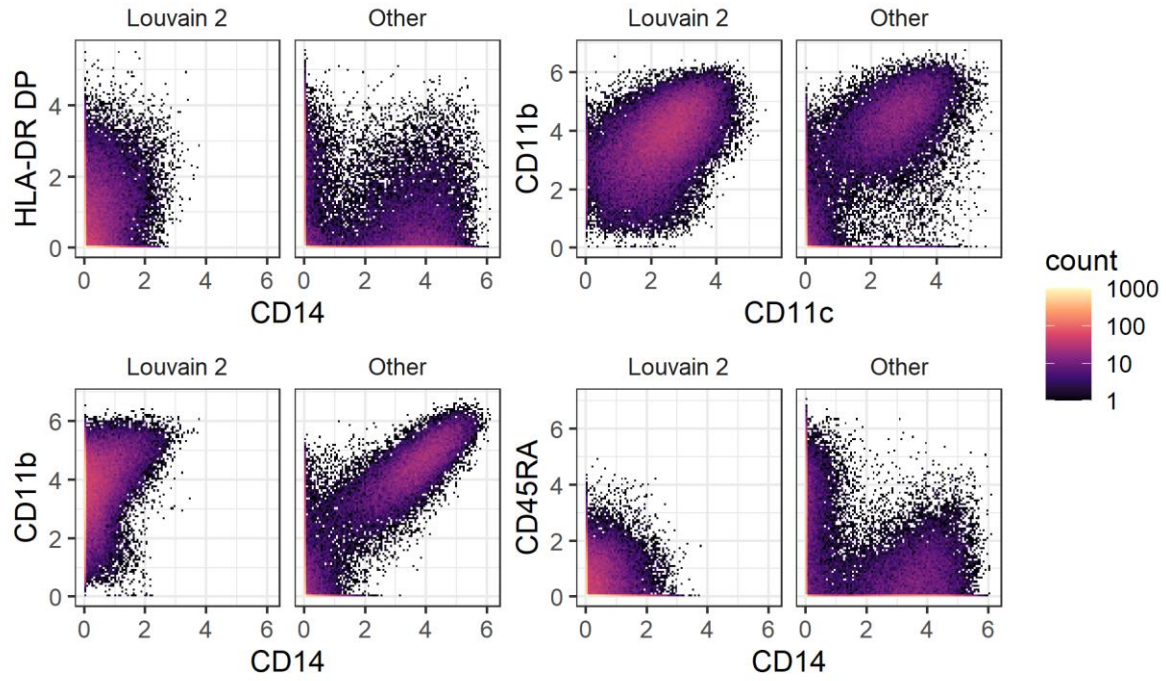


Figure C.5 Marker intensity in cells from Cluster 2 – Mixed M / DC. Bi-axial plots of marker intensity by cell count across cells in Cluster 2 compared to 500,000 cells sampled at random (Other).

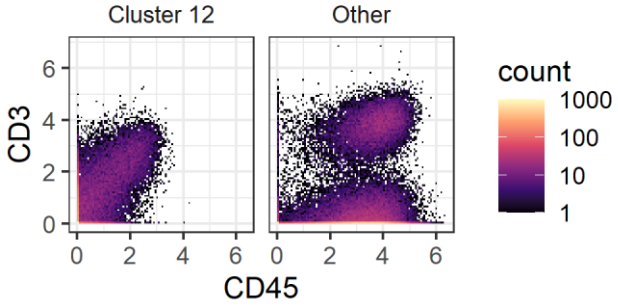


Figure C.6 Marker intensity in cells from Cluster 12 – T cells. Bi-axial plots of marker intensity by cell count across cells in Cluster 12 compared to 500,000 cells sampled at random (Other).

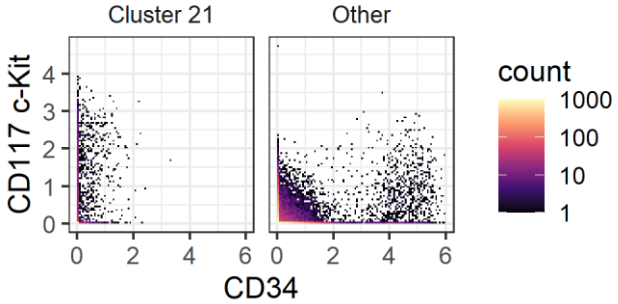


Figure C.7 Marker intensity in cells from Cluster 21 – CLP. Bi-axial plots of marker intensity by cell count across cells in Cluster 12 compared to 500,000 cells sampled at random (Other).

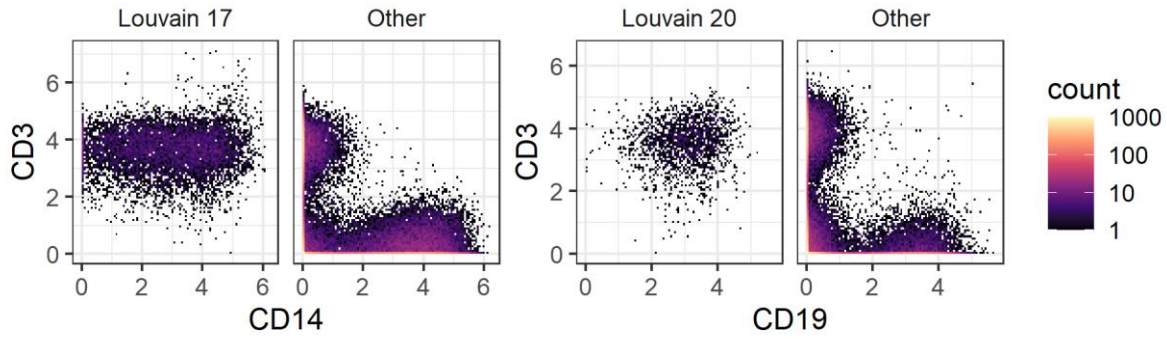


Figure C.8 Marker intensity in cells from Clusters 17 & 20 – doublets. Bi-axial plots of marker intensity by cell count across cell clusters compared to 500,000 cells sampled at random (Other).

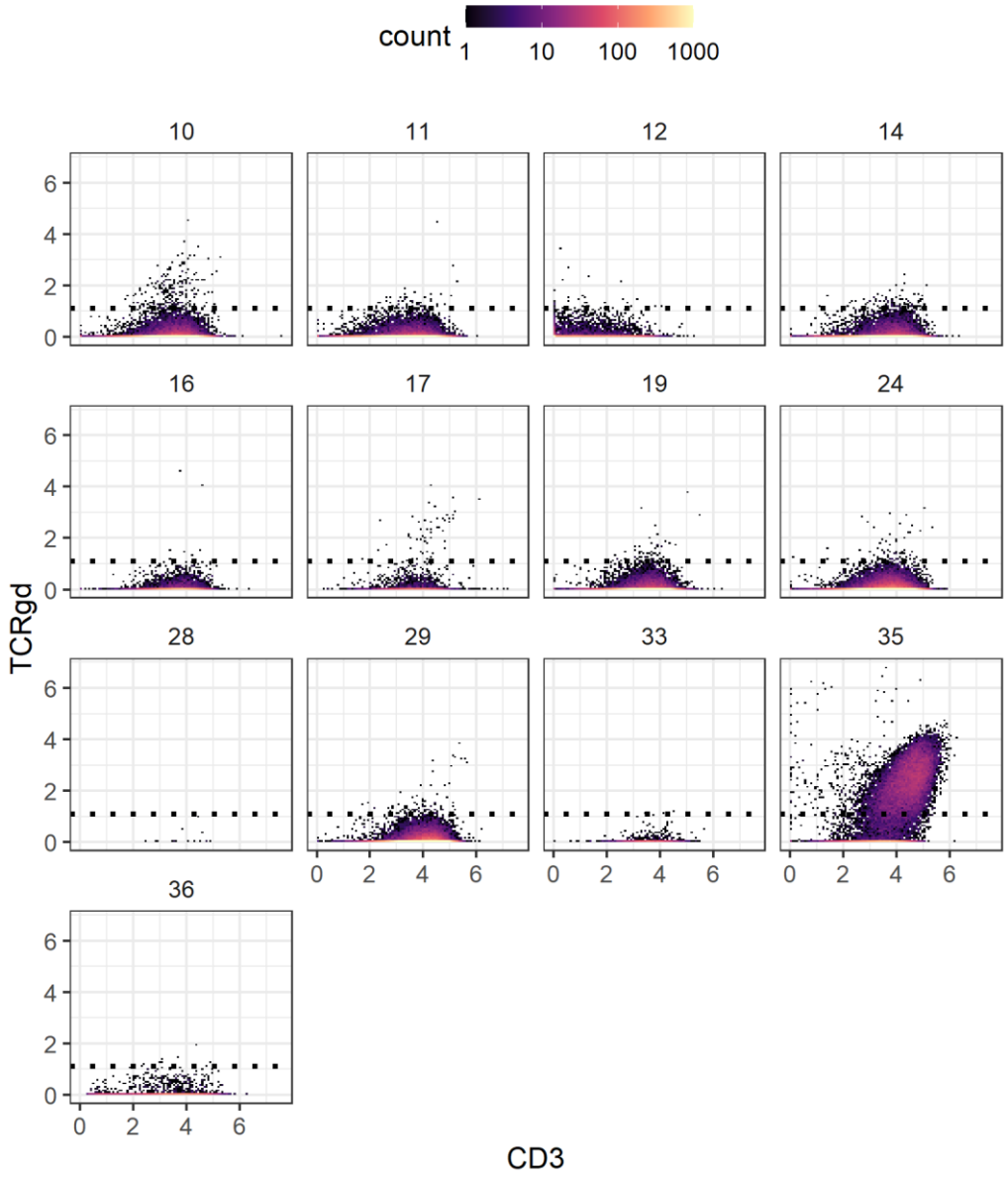


Figure C.9 Marker intensity in T cell clusters - γδ TCR & CD3 expression. Bi-axial plots showing marker intensity by cell count across clusters identified as T cells. Dashed lines indicate a qualitative threshold to aid visual interpretation.

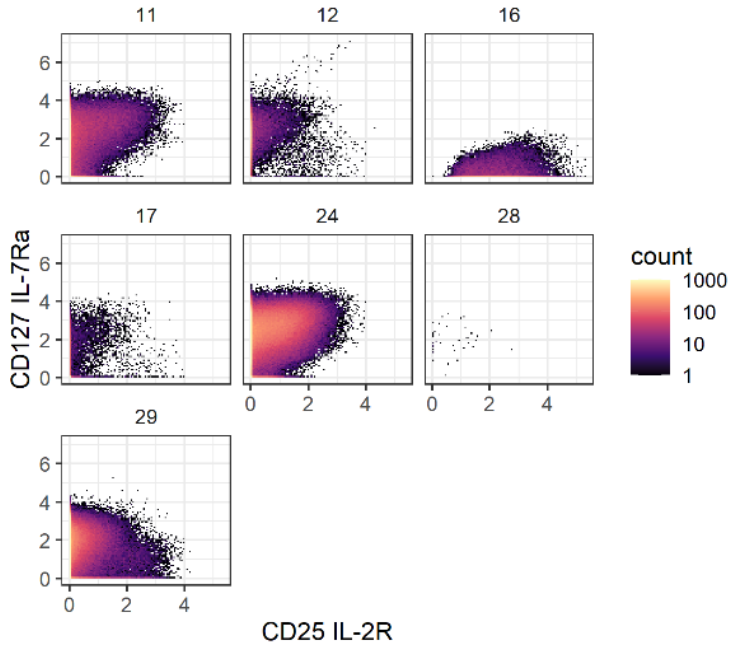


Figure C.10 Marker intensity in CD4+ T cell clusters - CD127 & CD25 expression. Bi-axial plots showing marker intensity by cell count across clusters identified as T cells.

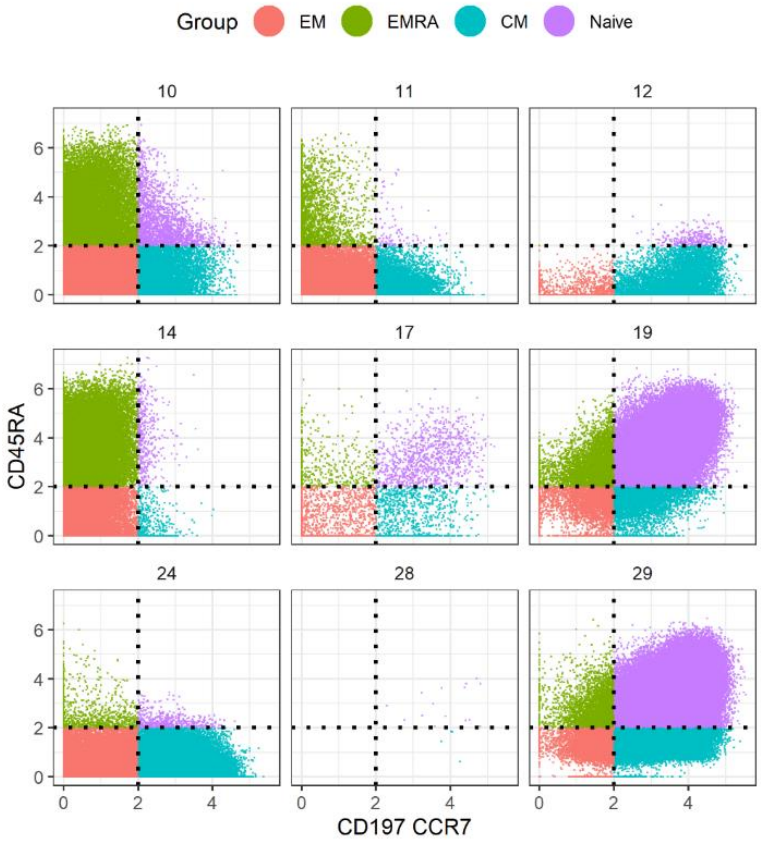


Figure C.11 Marker intensity in CD8+ and non-regulatory CD4+ T cell clusters – CCR7/CD197 & CD45RA expression. Bi-axial plots showing marker intensity of cells in T cell clusters. Dashed lines indicate the threshold selected to determine memory status. T cells characterised as naïve, effector memory (EM), effector memory cells re-expressing CD45RA (EMRA), and central memory (CM).

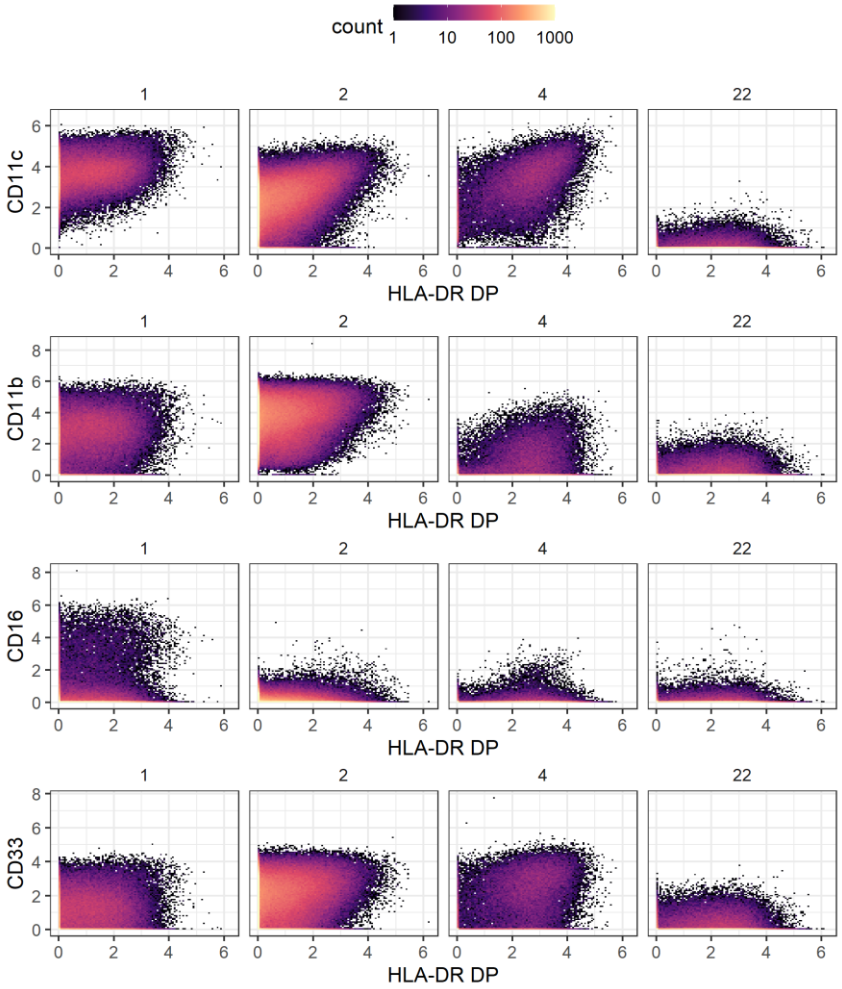


Figure C.12 Marker intensity in DC clusters. Bi-axial plots showing marker intensity by cell count across clusters identified as DCs.

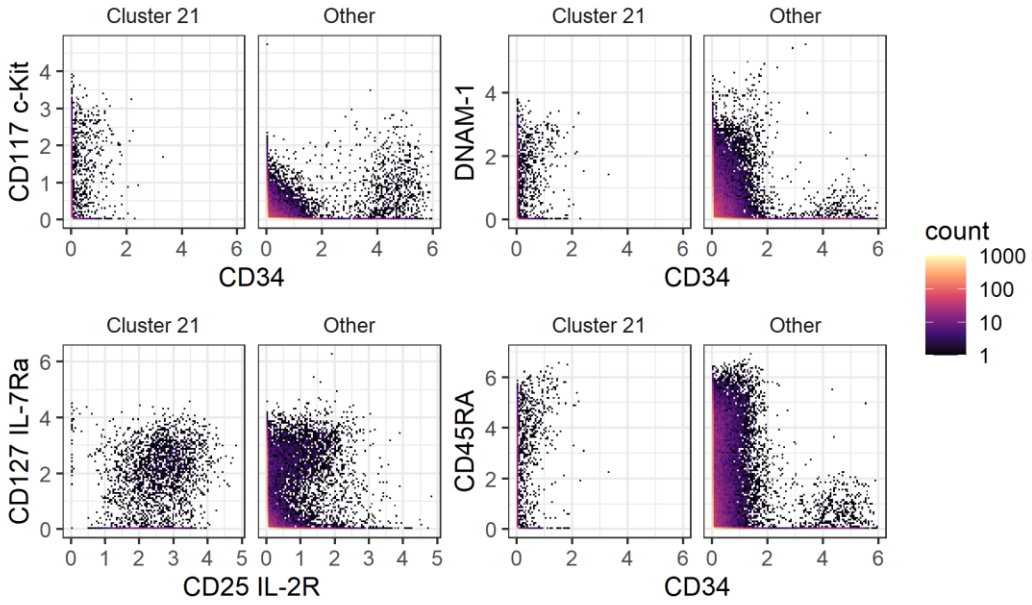


Figure C.13 Marker intensity in CLP cluster. Bi-axial plots showing marker intensity by cell count across Cluster 21 identified as CLPs.

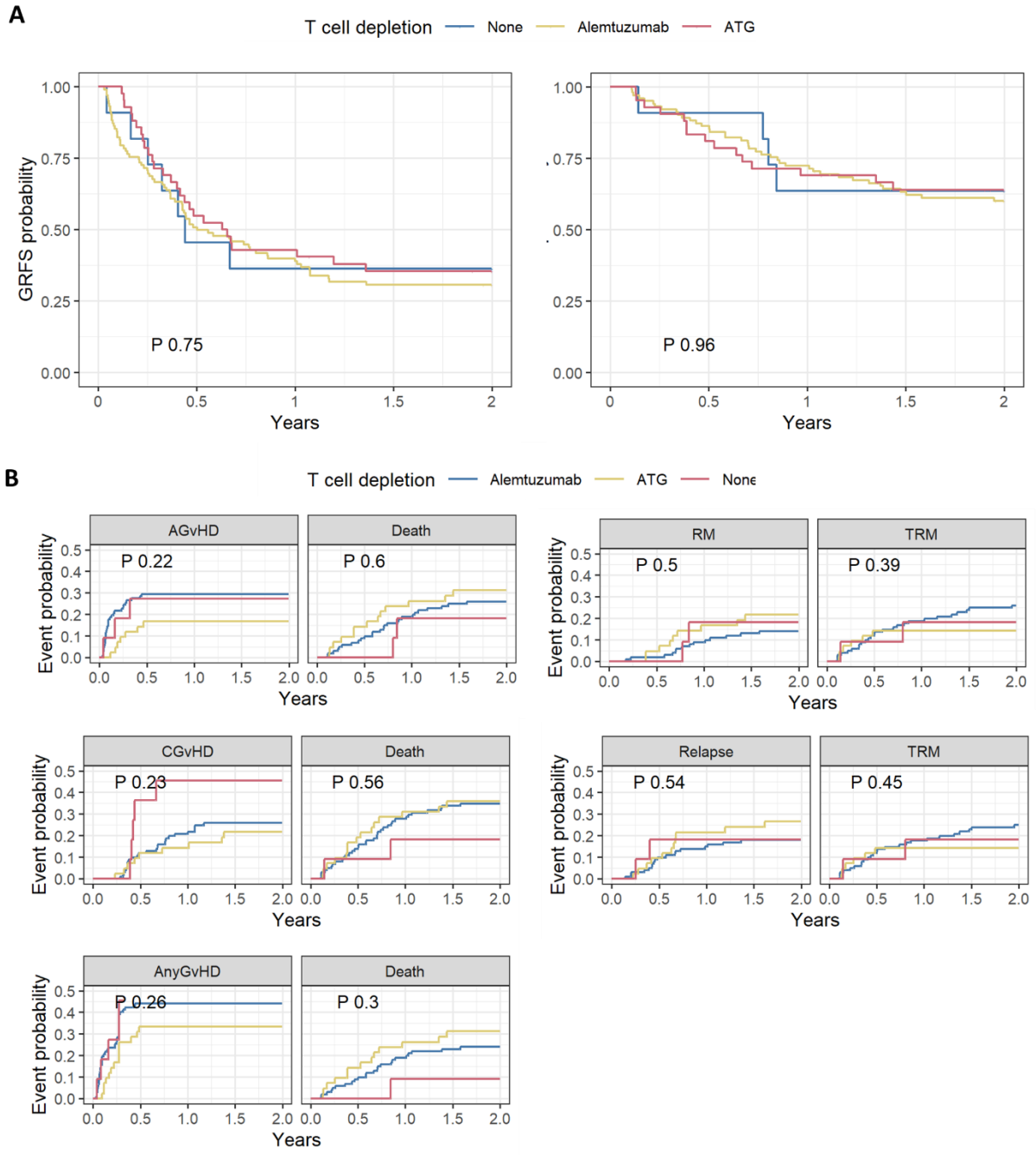


Figure C.14 Kaplan-Meier and cumulative incidence by in vivo T cell depletion. (A) Kaplan-Meier estimates of GRFS and OS. (B) Cumulative incidence estimates of AGvHD, CGvHD, GvHD of any kinda (AnyGvHD), relapse, RM, and TRM.

Table C.1 Procrustes correlation by choice of reference part for additive log-ratio transformation (ALR). Correlation between the transformed data and the exact log-ratio geometry, with a value of 1 corresponding to isometry.

Reference	Procrustes
HSC	0.90
B	0.92
DC	0.94
M	0.90
Lymph. Progen.	0.91
Mixed M/DC	0.90
NK	0.91
T	0.93

Table C.2 Events per variable (EPV) ratios for multivariate models in graft composition study. EPV ratios by clinical endpoint being modelled and type of variables included.

Clinical endpoint	Variable type	Vars	Events	EPV
OS	PBSC graft group	14	59	4.21
OS	Graft composition	19	59	3.11
GRFS	PBSC graft group	13	104	8.00
GRFS	Graft composition	18	104	5.78
AGvHD	PBSC graft group	13	40	3.08
AGvHD	Graft composition	18	40	2.22
CGvHD	PBSC graft group	NA	40	NA
CGvHD	Graft composition	18	40	2.22
Relapse	PBSC graft group	14	31	2.21
Relapse	Graft composition	19	31	1.63
TRM	PBSC graft group	14	34	2.43
TRM	Graft composition	19	34	1.79