ANALYSIS OF PERIPHERAL BLOOD MONONUCLEAR CELL DATA FOR FINDING A CELLULAR SIGNATURE OF GRAFT-VERSUS-HOST DISEASE

by

Yunfeng (Eric) Dai

BEng, Tianjin University of Technology and Education, China, 1990

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY
Fall 2008

APPROVAL

Name: Yunfeng (Eric) Dai

Degree: Master of Science

Title of thesis: Analysis of Peripheral Blood Mononuclear Cell Data for Finding a Cellular Signature of Graft-versus-Host Disease

Examining Committee: Dr. Richard Lockhart
Chair

Dr. X. Joan Hu,
Associate Professor, Statistics & Actuarial Science
Simon Fraser University
Senior Supervisor

Dr. Jinko Graham,
Associate Professor, Statistics & Actuarial Science
Simon Fraser University
Supervisor

Dr. Rick Routledge,
Professor, Statistics & Actuarial Science
Simon Fraser University
External Examiner

Date Approved: _____

Abstract

Graft-versus-host disease (GvHD) is the most important complication of bone marrow transplantation (BMT). It has high morbidity and mortality. Current diagnosis and treatment assessment rely primarily on ambiguous clinical symptoms. It is believed that the outcome of patients diagnosed with GvHD can be improved if they are treated before the development of full-scale clinical symptoms. In an attempt to provide early prediction of GvHD, we analyzed peripheral blood mononuclear cells (PBMC) data, together with GvHD outcome information, from a group of patients who underwent BMT. Various statistical methods were applied to overcome the difficulty arising from the high-dimensional longitudinal observations but with relatively small sample size. Analysis of "windowlized" data at a fixed time point did not find one subtype of PBMC or a subset of PBMCs that was important in the prediction of aGvHD. Longitudinal data analysis indicates that proportions of T_ cells over time may separate well the patients with aGvHD and without aGvHD. This is consistent with the main finding in Lee (2006). In addition, we considered a two-level hierarchical model to describe the proportions of PBMCs. It indicates that T_cells $CD3^{-}CD8\beta_{dim}CD8^{-}$ is marginally associated with and disease. We note that the trends over time of "T_helper" and "T_suppressor" are closely associated with aGvHD occurrence.

Acknowledgments

I am grateful to all those people who have kindly helped me during my study at Simon Fraser University. In particular, I would like to thank X. Joan Hu, Lihui Zhao, Shuli Ma, Ryan Brinkman, Egon Simons, my wife Xiangrong, and my thesis committee members: Dr. Jinko Graham and Dr. Rick Routledg.

I would also thank Dr. Brinkman's group at the BC Cancer Agency for allowing me to use the data from their leukemia study.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 Graft-versus-Host Disease

Graft-versus-host disease (GvHD) occurs after allogeneic hematopoietic stem cell transplantation (HSCT) or bone marrow transplantation (BMT). It is the most important complication of BMT. After transplantation, immune cells made by the transplanted marrow recognize the recipient's tissues as 'foreign', and attack them. This causes GvHD. GvHD has high morbidity and mortality. There are two types of GvHD: acute GvHD (aGvHD) and chronic GvHD (cGvHD) . They can be distinguished on the different time of onset. aGvHD usually occurs within 100 days post-transplant, and cGvHD occurs approximately four months after transplantation. This distinction is not arbitrary: aGvHD and cGvHD appear to involve different immune cell subsets, different cytokine profiles, and different types of target organ damage. The current diagnosis relies primarily on ambiguous clinical symptoms, such as fever, rash and diarrhoea. Patients' condition could be improved if they were treated in a preemptive fashion which requires an earlier diagnosis of GvHD. It is thus in demand to develop a method that can diagnose GvHD before it is well established.

1.1.2 A BCCA Study on Predicting GvHD

A project at the British Columbia Cancer Agency (BCCA), 'Prediction of Graft-versus-Host Disease', investigated the utility of Support Vector Machine (SVM)-based classifiers trained flow cytometry data to assist in GvHD diagnosis, and aimed at elucidating the molecular

profile of GvHD. The study goal was to define a cellular signature that could guide early diagnosis of GvHD.

The study collected information from a total of 31 patients who underwent bone marrow transplantation. Among the 24 patients who are included in a preliminary data analysis (Lee 2006), 21 patients developed aGcHD, cGvHD or both, and three patients did not have either aGvHD or cGvHd. The data of the other patients were not included in the preliminary study due to the following : (1) the patients died prior to 100 days post-transplant and it could not be determined if they would have subsequently developed GvHD or not ($n = 2$); (2) the patients developed de novo chronic GvHD, which might have confounded the analysis for aGvHD ($n = 3$); (3) the patients were lost to follow-up ($n = 1$); (4) the patients had insufficient clinical samples collected ($n = 1$). Since these seven patient did not have aGvHD, we consider them as non-aGvHD patients in our study.

The patterns of change in PBMCs over time may unveil the fact of developing GvHD. The main hypothesis of the study was that "onset of aGvHD or cGvHD may be predicted by identifying patterns of cellular markers in peripheral blood mononuclear cells via a flow cytometric high content screening (FC-HCS)".

1.1.3 Diagnosis of GvHD

Current diagnosis of GvHD is mainly based on clinical features and tissue biopsies. A noninvasive, unbiased laboratory test for GvHD diagnosis does not exist. Scientists are trying to develop an efficient and accurate method to predict the disease. The serum proteomic pattern, the gene microarray and the flow cytometry data are believed to be of potential usefulness.

Flow cytometry (FCM) is a technology that can separate and count different particles by combining their light reflecting and scattering properties. Flow cytometric high-content screening (FC-HCS) combines advances in robotic fluid handling, flow cytometric instrumentation and bioinformatic software so that relatively large numbers of flow cytometry samples can be processed and analyzed in a short period of time. Hematopoietic membrane and cytoplasmic markers that are stained by fluorescent dyes are classified as clusters of differentiation (CD). CDs generally represent cell-surface antigens. Different immune cell lineages and functions can be identified using different combination of the CD markers. Immune cells are believed to be associated with GvHD. Because GvHD is induced by the reaction of donor T cells to recipient histoincompatible antigens, the goal may be achieved

by studying relationship among onset of GvHD, the change patterns of peripheral blood mononuclear cells (PBMCs) , transplantation types, GvHD grades, and the demographic information on patients. aGvHD and cGvHD are different in terms of the time of onset, clinical manifestations and distinct pathobiological mechanisms. The cellular signatures of aGvHD and cGvHD may also be different.

Jessica Lee in her thesis (Lee, 2006) conducted a quality assurance (QA) test on the FCM data and set up a suitable temporal analysis pipeline to process the high-throughput FCM dataset. The QA test was to identify values motivated by experimental errors. By separately analyzing data with each PBMC, she compared samples taken from the aGvHD and the non-GvHD patients, and samples taken from the 7 aGvHD & cGvHD and the 9 aGvHD only patients. She used samples taken between 7 and 21 days post-transplant to find the classifier for the onset of aGvHD. The most persistent correlation to the onset of aGvHD was observed from the immune cells $CD3^+CD4^+CD8\beta^+$ and its subpopulation $CD3^+CD4^+CD8\beta^+CD8^+$. No uniform classifier for the cGvHD was found by the analysis across different PBMCs. The primary statistical method was functional linear discriminant analysis (FLDA), which is designed specifically for data sets generated by sparse sampling over a time period.

1.2 Description of the BCCA GvHD Study Data

1.2.1 Patient Information & Clinical Outcomes

The study enrolled 31 patients who underwent BMT between 20 Apr, 2001 and 23 Jan, 2003. The patient information and the clinical outcomes are listed in Table 1.1. Figure 1.1 shows when patients had GvHD, died or dropped out, and how long they stayed in the study. For patients 3, 7 and 20, their last measurements were recorded later than their death. We have asked the data provider for the reason, but have not obtained the reply.

1.2.2 Lab Test Outcomes

Concentration of Mononuclear Cell

A mononuclear cell is a leukocyte with a single non-segmented nucleus in the mature form. The mononuclear cell (MNC) concentration values (mm^3) were obtained separately using different samples taken from the same patient at multiple time points. Figure 1.2 shows

collection times of MNCs data and PBMCs during a period from −16 day to 200 day post BMT.

Proportion of Peripheral Blood Mononuclear Cells (PBMCs)

There are two major lab test results: proportion and concentration of PBMCs. Data on a total of 123 subsets of PBMCs were obtained by flow cytometry from the 31 BMT patients. At the time of flow cytometry analysis, cells are thaw and aliquot into 96-well plates. The plates are stained with antibody combinations. These stained blood samples can be used for flow cytometry analysis. The data obtained from FCM are combined with immune cell concentration data and transformed into a proportion data set and a concentration data set. The proportion dataset contained 123 subsets of immune cells; each is corresponding to the proportion of cells (proportion of either the total PBMCs or total CD3+ cells) in the gate. The concentration dataset was obtained by multiplying each proportion value with the MNC concentration of samples taken at the closest date. This project focuses on the PBMC proportion data since they are obtained directly from samples.

The subsets of immune cells from each of the ten aliquots are described in *Appendix A*.

1.2.3 Summary of the Measurements

On average, the study had 14 (±3) blood samples per patient collected on an approximate weekly basis. Samples were collected from 0 to 16 days before the transplantation and 49 to 400 days after the transplantation. The responses are multivariate longitudinal for each of the 31 patients. A total of 123 subtypes of PBMC were measured over a period of approximate 400 days. Figure 1.3 gives an intuitive view of the measurements of PBMCs. The collection times vary from patient to patient, and are unequally-spaced within a patient. Figure 1.4 shows the trajectories of one type PBMCs of all patients, which was identified in Lee (2006) as a cellular marker of GvHD.

The proportions are not always based on all PBMCs. There are intermittently missing data due to drop-outs. Some patients did not have data with certain types of PMBCs. In the clinical study, these intermittent missing values are caused by failing to follow-up, and are usually considered as missing at random. In the cases with data missing at random, data can be analyzed by any method which can accommodate unbalanced data. It appears that the drop-outs do not influence much the analysis of aGvHD, but confine the study of

cGvHD.

The histograms of the proportion data suggest that the responses be not normally distributed. Transformation is needed before an analysis that requires normality assumption.

1.3 Outline of Statistical Analyses

Table 1.2 summarizes the available study data: Items $1, 2, 3$ are lab test outcomes, items $4, 5, 6, 7$ are clinical outcomes, and the remainders are demographic data. To achieve the study goal of defining a cellular signature for early diagnosis of GvHD, we conduct the following statistical analyses.

1.3.1 Analysis of the Clinical Outcomes

This is to explore the clinical outcomes in association with patients' demographic information. Survival and generalized linear regression methods are applied. The analysis is presented in Chapter 2.

1.3.2 Exploring the Lab Test Outcomes

We visualized the data in Chapter 3 to show change patterns of PBMC proportions and relationship between patient groups or subtypes of PBMCs.

In order to apply some exploratory data analysis methods, such as principal components analysis, and clustering, we first rescaled the collection time as if the samples were collected at fixed time points. In this study, the width of window is 7 days (one week) and the midpoints are $-14, -7, 0, 7, 14, 21, \ldots$. It roughly consists with the interval of collection. Registration is used to synchronize different patient response times. The subtypes of PBMCs whose proportions are determined by the proportions of other subtypes are excluded in the analysis. *Appendix A* gives the list of the PBMCs types that are redundant. Chapter 3 presents a detailed description of the data manipulation.

For each fixed time point, we conduct cluster analyses to search for patterns in PBMCs which are associated with the clinical outcomes. If a pattern agrees with clinical outcomes, it may give a GvHD prediction. The investigation also includes finding most important components by using principal component analysis methods. In our study, there are 123 types of PBMCs. Some of them are redundant, and can be eliminated without losing any

information. Some redundant cell types may be picked out by statistical methods. Two approaches are considered to select important/non-redundant cell subtypes.

1. To view proportions of PBMCs, denoted by $X_1, X_2, \ldots, X_K$, as predictors and the time to the onset of GvHD or status of disease as responses, and select a subset of the K cells by using the stepwise selection method in regression.

2. To view $X_1, X_2, \ldots, X_K$ as responses, and to conduct a stepwise discriminant analysis to select a subset of the responses.

We report these analyses in Chapter 4.

1.3.3 Longitudinal Data Analysis

Consider a subset of PBMC proportions as responses collected over time, and the collection times, the status of disease and cell types as predictors. MANOVA and multilevel mixed effects models can be applied to find desirable pattens in PBMC.

The "windowlized" data described in Section 1.3.2 are analyzed by MANOVA. We consider the status of disease and the cell type as factors, PBMC proportions obtained over time as response vectors. A two-way MANOVA are conducted to examine association of PBMC with the disease. If the status of disease or interaction between cell subtypes and the status are significant, the corresponding PBMCs may be the cellular markers for GvHD prediction.

The multilevel mixed effects models are considered in the analysis of raw (the original) data. We analyze the data with different models, and investigate which model is best in description of the data. Based on the best model, we investigate the relationship and change patterns between the PBMC responses and PBMC subtypes, collection times, and status of disease.

Longitudinal data analyses are presented in Chapter 5.

We provide a brief summary of the data analyses mentioned above in Chapter 6, along with comments and suggestions for further investigation.

Table 1.1: Patient Information and Clinical Outcomes

patient ID	BMT date	donor-patient relationship	aGvHD date	aGvHD grade	cGvHD date	drop-out date
1	8/16/01	MUD	9/11/01	3		2/19/02
2	8/27/01	SIB		0		
3	8/24/01	MUD	9/16/01	4		10/24/01
4	9/13/01	SIB		0		6/28/02
5	10/12/01	SIB	12/10/01	3		
6	9/20/01	SIB	10/9/01	3		
7	5/25/01	SIB	7/3/01	3		8/22/01
8∗	10/26/01	SIB		0	2/25/02	
9	10/25/01	SIB	12/7/01	3	5/24/02	
10	10/18/01	MUD	10/29/01	1		
11	11/9/01	SIB	1/16/02	1	8/9/02	
12	11/8/01	SIB	11/30/01	3		
13	11/1/01	SIB	12/19/01	3		
14	11/9/01	MUD	12/7/01	2		
15	6/14/01	SIB	7/3/01	2	9/20/01	
16	12/8/01	MUD	12/18/01	2		2/20/02
17	1/31/02	SIB		0		
18∗	1/17/02	SIB		0		3/12/02
19	5/17/02	SIB	8/2/02	2	8/6/03	
20∗	6/18/01	SIB		0		8/2/01
21	7/5/01	MUD	8/28/01	3	4/25/02	
22	1/31/02	SIB	3/4/02	3	9/11/02	
23	7/19/01	SIB	8/10/01	3		10/27/01
24	7/25/01	SIB	8/31/01	3		
25	5/18/01	SIB	7/1/01	1		8/16/01
26∗	5/18/01	SIB		0	9/12/01	
27	4/20/01	SIB	5/21/01	2		
28	5/9/01	MUD	6/29/01	1	11/2/01	
29	11/15/02	SIB		0		2/20/03
30∗	1/23/03	SIB		0	5/7/03	
31∗	1/17/03	SIB		0		5/4/03

patient ID with a '*' indicates the patient was not used in Lee's study

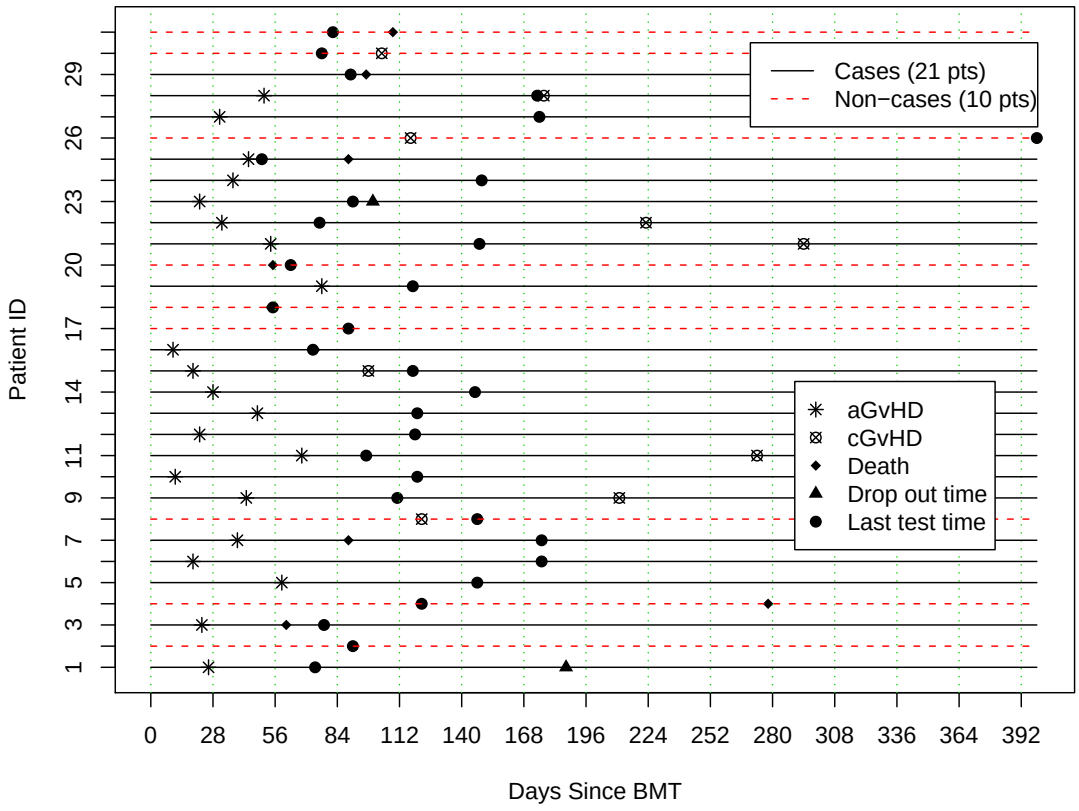

Figure 1.1: Events over Time

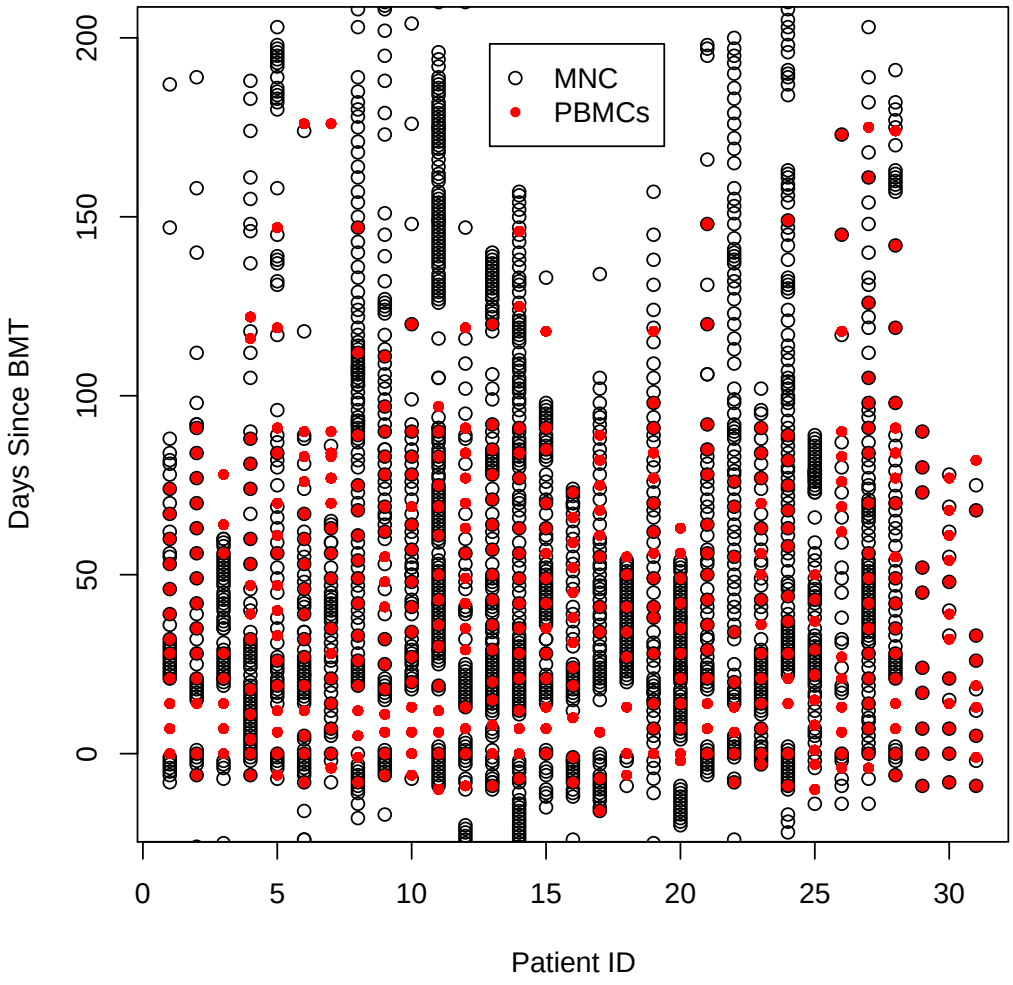

Figure 1.2: Collection Times of MNC and PBMC Data for Patients

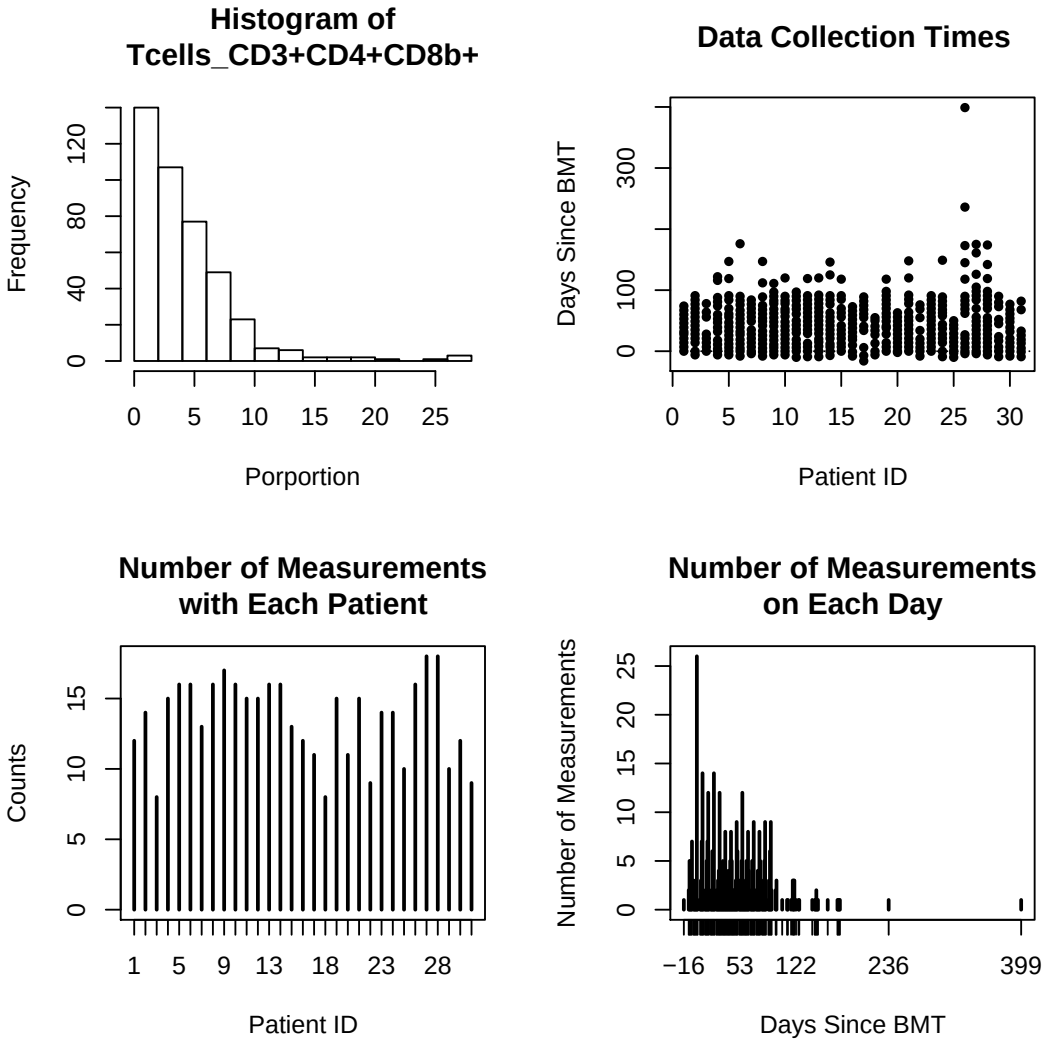

Figure 1.3: Summary of PBMC Measurements

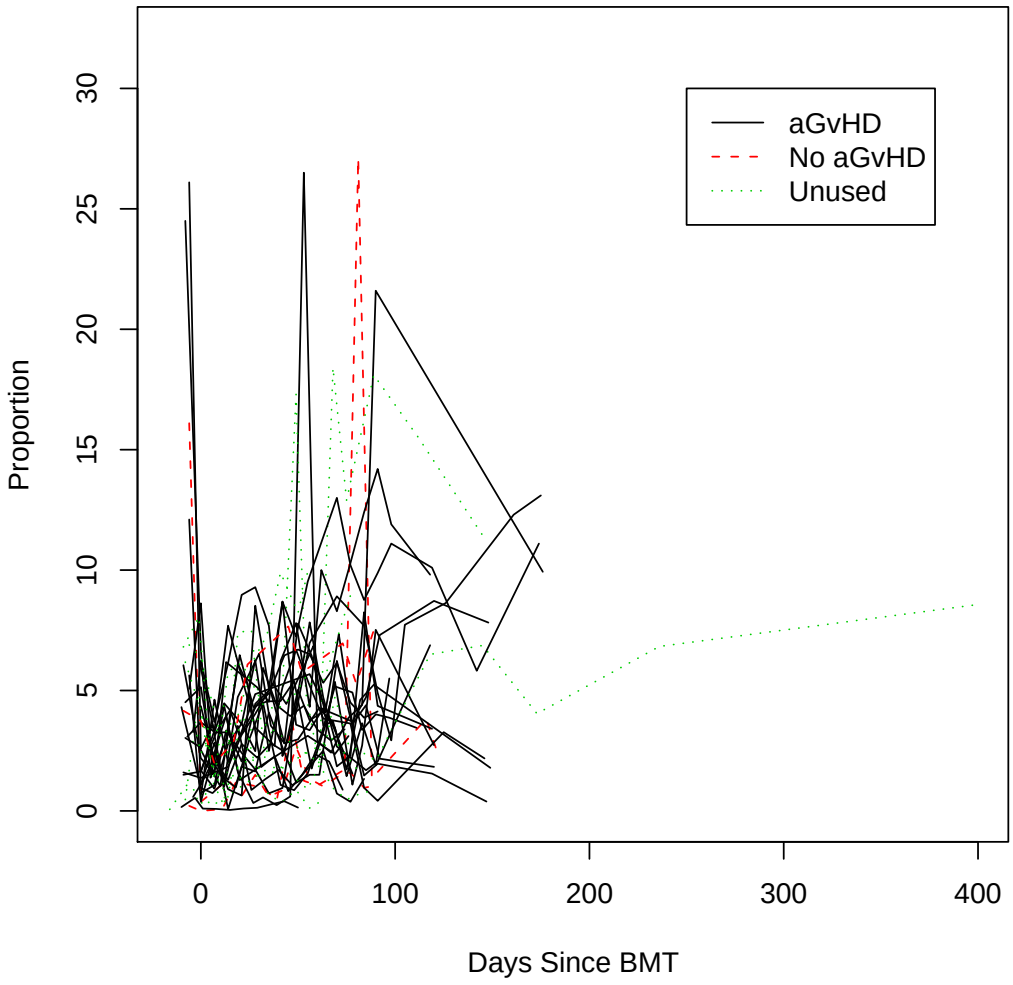

Figure 1.4: Trajectories of T_cell CD3+CD4+CDb8+

Table 1.2: Data Used in Analysis

Item #	Name	Value	Comment
1	PBMC types	$1 \sim 123$	123 subsets of PBMCs.
2	Proportion and Concentration of 123 subset of PBMCs	$0 \sim 100$	Not independent, with missing and drop-out.
3	Measuring Time	$-16 \sim 399$	Unequal space, different from patient to patient.
4	aGvHD Post-transplant	$11 \sim 68$	# of days from BMT to diagnosis of aGvHD.
5	Max aGvHD Grade	$0 \sim 4$	0 means no GvHD.
6	cGvHD Post-transplant	$98 \sim 446$	# of days from BMT to diagnosis of cGvHD.
7	Donor-patient Relationship	MUD/SIB	
8	Age		Age at transplantation.
9	Gender		Gender of patient.

Chapter 2

Analysis of Clinical Outcomes

2.1 Clinical Outcomes

The clinical outcomes include the time to aGvHD or cGvHD and the associated grades. Bone marrow transplantation date, the type of BMT (relationship between donor and patient), the type of leukemia, the drop out time, the death time and the last visit time are available for each patient. Table 2.1 provides a description of nine variables of interest. There are 20 categories of types of leukemia in the 31 patients. We exclude the variable of leukemia type from the analysis. We investigate how the disease or death is associated with the patients' characteristics.

In particular, survival analyses are conducted regarding four types of events: onset of GvHD (including acute GvHD and chronic GvHD), onset of aGvHD, onset of cGvHD, and death.

The time is calculated from bone marrow transplantation (BMT) to the occurrence of the event. Survival analysis and logistic analysis focus on different aspects of events. Survival regression analysis focuses on the procedure. I also conducted a logistic regression analysis to study the clinical outcomes. Logistic regression analysis focuses on the consequence. It is of interest to take the grades of GvHD into account, so is exploring the transition probabilities between the different states.

2.2 Analysis of Times to Events

To examine the characteristics of the overall survival, both nonparametric and parametric methods are used in the analysis. Kaplan-Meier estimates (Figure 2.1) show patients developed aGvHD before 100 days post transplantation and cGvHD after 100 days post transplantation. Most deaths occurred before 100 days. The first 100 days seems critical for BMT patients.

We chose three common parametric models for the survival function, the Weibull, Log-normal and Log-logistic models. The associated probability density functions and survivors are provided by Table 2.2. From the quantile-quantile plots (Figure 2.2), parametric survival curves (Figure 2.3) and estimates of parameters (Table 2.3), we see the Log-logistic model is the best parametric model to describe times to GvHD, aGvHD, or cGvHD. The estimates based on the model are in a good agreement with the Kaplan-Meier estimates.

2.3 Relationship between Events and Covariates

We are interested in modeling and determining the relationship between events and covariates. The BMT type and transplantation date are the covariates in the following regression analyses.

2.3.1 Regression Analysis Based on Cox Proportional Hazards Model

I considered the Cox proportional hazards regression model (Cox PH model). The Cox PH model assumes that the hazard conditional on his covariates X_i for a subject is in the form: $h_0(t)e^{\beta' X_i}$, where β is a vector of coefficients, $h_0(t)$ is an arbitrary and unspecified baseline hazard function, and X_i is the vector of covariates for the i^{th} individual.

The initial model included two covariates, BMT type and BMT time. The estimated coefficients of Cox PH model are provided in Table 2.4. It appears that the time to GvHD or to aGvHD depends on BMT type. The relative risks of having GvHD or aGvHD for patients with non-sibling or sibling donors are 3.17 or 2.87 fold higher, respectively. The probability of having GvHD or aGvHD for patients who received the bone marrow from their siblings is much lower. However, in a long time period, the two groups have no significant difference in death or having chronic GvHD.

Cox PH model assumes that the baseline hazard function is the same for all individuals.

It implies that

$$log(-logS(t|x)) = log(-log(S_0(t)) + \beta' x.$$

We plot estimates of $log(-logS(t|\mathbf{x}))$ against t in Figure 2.4. The lines of the two groups with different donors look parallel. This indicates that the Cox PH assumption is appropriate with the data.

2.3.2 Logistic Regression Analysis

Our interest then focuses on whether the occurrence of an event of interest depends on covariates. Consider response Y as a binomial variable with value 1 if a certain event occurred in the study and 0 otherwise.

Table 2.5 provides the summary of the logistic regression analysis. The likelihood ratio tests (LRT) of null model vs model #3 and #4 are not significant, p-value = 0.935 and p-value = 0.691 respectively. Adding covariates BMT type and BMT_time would not help to predict occurrence of cGvHD and death. There is mild evidence that the probability of having aGvHD will decrease if a patient postpone the operation. The odds of having aGvHD will decrease by 0.5%. The logistic regression model does not address the censoring. It may underestimate the influence of the covariates. Results from survival analysis could be more reliable.

2.4 Discussion

The aGvHD was classified into five categories with grades from 0 to 4 according to the severities, where 0 indicates no aGvHD. The onset of aGvHD is a marked event. One way to conduct a marked event analysis is to consider the grade as a covariate. Further investigation is required. The result from the Cox PH model shows aGvHD grades are not significant with p-values close to 1.

The status of a patient can transit from healthy to aGvHD, cGvHD or death. aGvHD also can change to cGvHD or death, and cGvHD can transit to death. What are the probabilities of transition from one stage to another stage? Dose transition depend on time or other factors? Multistate analysis may provide answers to these questions.

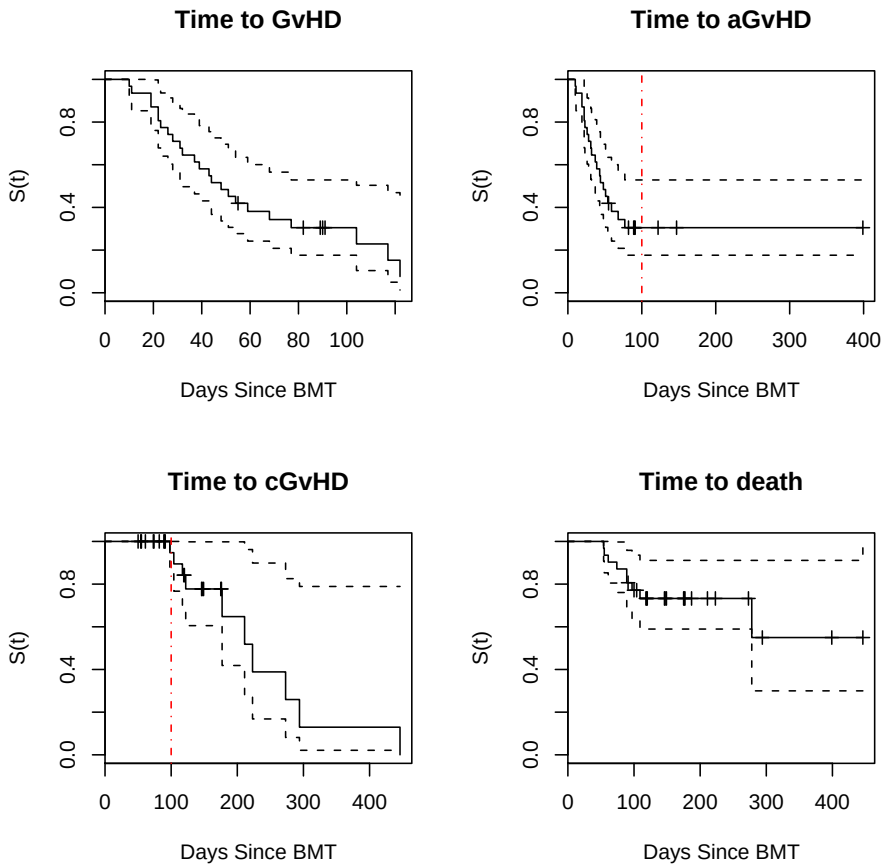

Figure 2.1: Kaplan-Meier Estimates: Solid Lines are Estimates of Survivor Functions, Dashed Lines are 95% Confidence Limits.

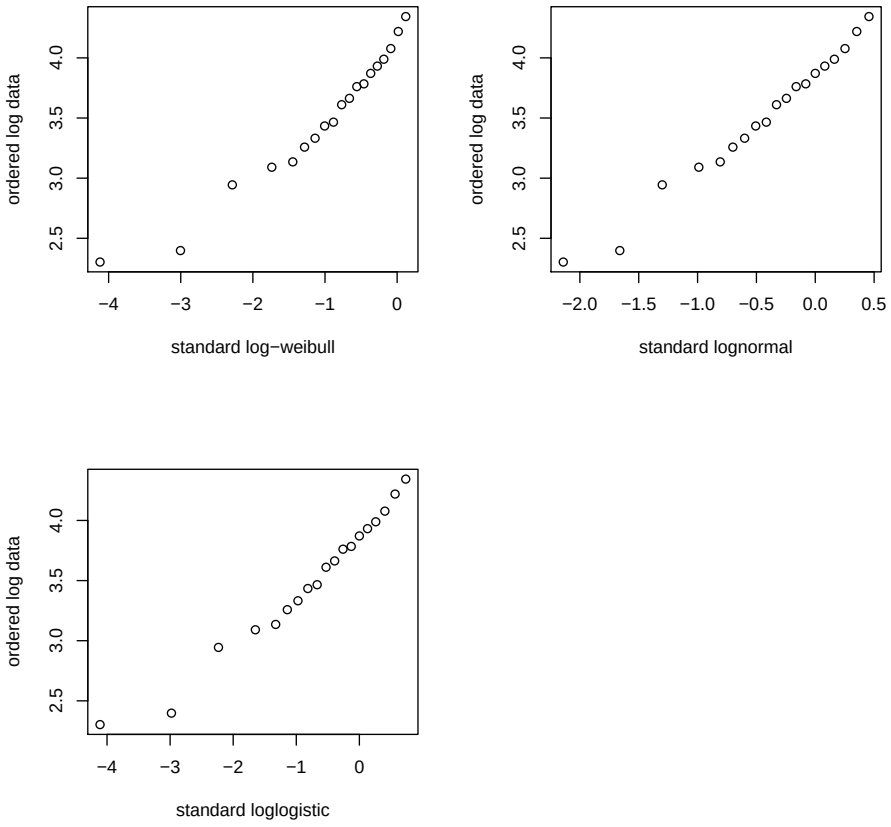

Figure 2.2: Quantile-quantile Plots

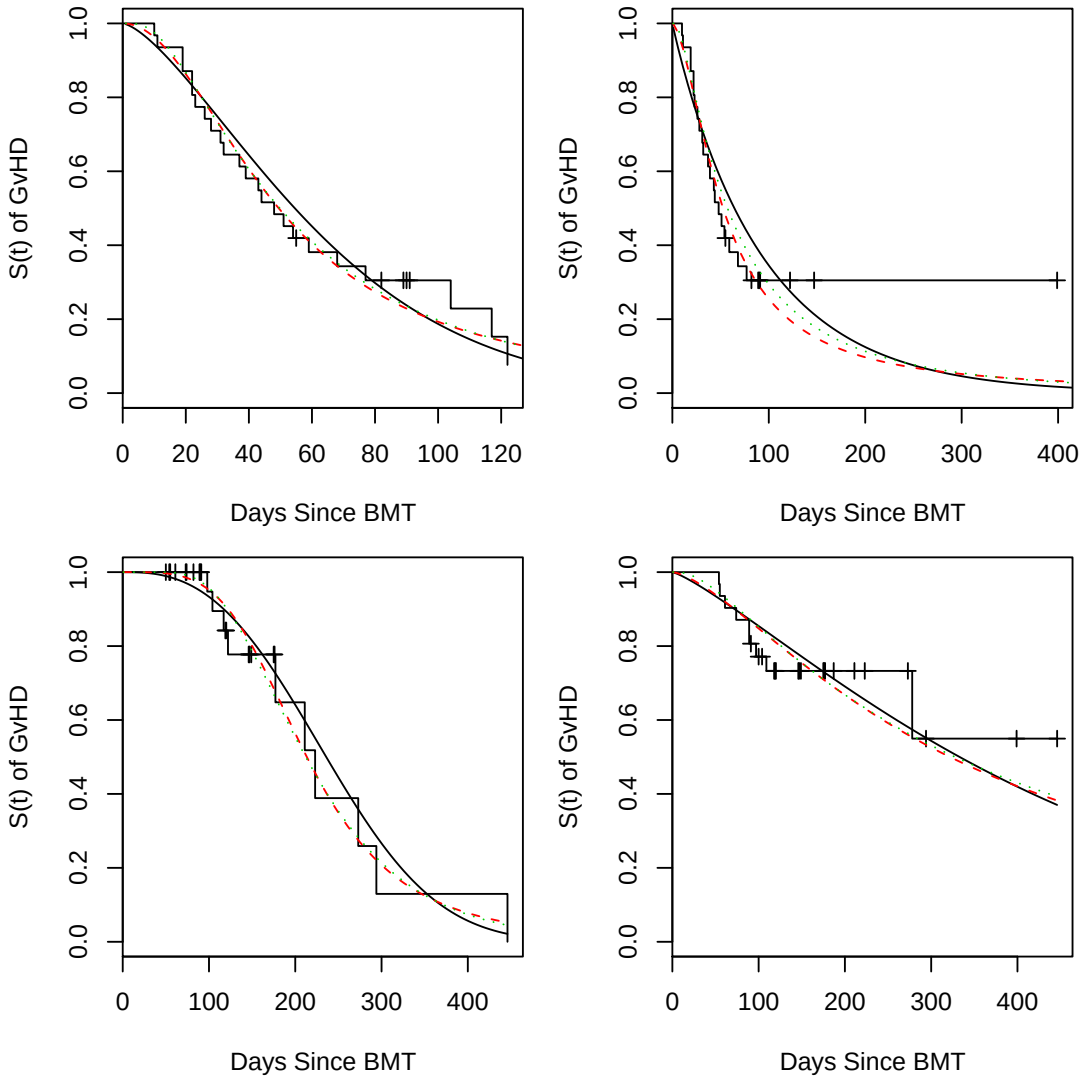

Figure 2.3: K-M Estimates vs Parametric Estimates: Step Lines are K-M Estimates, Solid Curves are the Weibull Estimates, Dashed Lines are the Log-logistic Estimates, Dotted Lines are the Log-normal Estimates.

Table 2.1: Variables in Clinical Outcome Analysis

Seq No.	name	value	comment
1	BMT_time	Apr 20, 2001 to Jan 23, 2003	There are two groups separated by median
2	diagnosis		20 types
3	relationship	SIB or MUD	
4	agvhd	$10-77$ days post BMT	$median = 32$ days post BMT
5	agrade	$0, 1, 2, 3, 4$	0 refers to no GvHD
6	cgvhd	$89-446$ days	$median = 194$ days
7	death	$50-122$ days	$median = 73$ days
8	last_visit	$50-399$ days	$median = 118$ days
9	dropout		only two dropouts

Table 2.2: PDFs and Survival Functions

Distribution	**pdf**	**S(t)**	**Note**
Weibull	$\lambda\alpha(\lambda t)^{\alpha-1}exp(-(\lambda t)^{\alpha})$	$exp(-(\lambda t)^{\alpha})$	shape:α, scale:λ
Log-normal	$\frac{1}{\sigma t\sqrt{2\Pi}}exp(\frac{-(log(t)-\mu)^2}{2\sigma^2})$	$1-\Phi(\frac{log(t)-\mu}{\sigma})$	$\Phi(.)$ denotes std. normal.
Log-logistic	$\lambda\alpha(\lambda t)^{\alpha-1}(1+(\lambda t)^{\alpha})^{-2}$	$\frac{1}{1+(\lambda t)^{\alpha}}$	$\frac{S(t)}{1-S(t)} = (\lambda t)^{-\alpha}$

Table 2.3: Estimates of Parameters

Life Time	Model	Parameter	Estimate)	Std. Error
Time to GvHD	Weibull	α	1.458	0.242
		λ	0.014	0.002
	Log-logistic	α	2.032	0.341
		λ	0.020	0.003
	log-normal	μ	3.909	0.153
		σ	0.819	0.757
Time to aGvHD	Weibull	α	0.970	0.160
		λ	0.011	0.002
	Log-logistic	α	1.672	0.310
		λ	0.019	0.004
	log-normal	μ	4.045	0.200
		σ	1.035	4.847
Time to cGvHD	Weibull	α	2.689	0.500
		λ	3.699×10^{-3}	4.352×10^{-4}
	Log-logistic	α	3.934	0.895
		λ	4.686×10^{-3}	6.092×10^{-4}
	log-normal	μ	5.357	0.126
		σ	0.438	0.240
Time to death	Weibull	α	1.232	0.340
		λ	2.230×10^{-3}	7.863×10^{-4}
	Log-logistic	α	1.469	0.409
		λ	3.108×10^{-3}	1.069×10^{-3}
	log-normal	μ	5.792	0.356
		σ	1.134	2.106

Table 2.4: Coefficients Estimates with Cox PH Model

Life Time	Coefficient	Estimate	Std. Error	p-value
Time to GvHD	BMT type(SIB)	-1.155	0.515	0.025
	BMT_time	-1.55×10^{-3}	1.45×10^{-3}	0.280
Time to aGvHD	BMT type(SIB)	-1.053	0.511	0.039
	BMT_time	-2.57×10^{-3}	1.69×10^{-3}	0.130
Time to cGvHD	BMT type(SIB)	0.774	1.048	0.460
	BMT_time	-2.62×10^{-3}	3.69×10^{-3}	0.476
Time to death	BMT type(SIB)	-0.152	0.840	0.856
	BMT_time	1.43×10^{-3}	1.83×10^{-3}	0.437

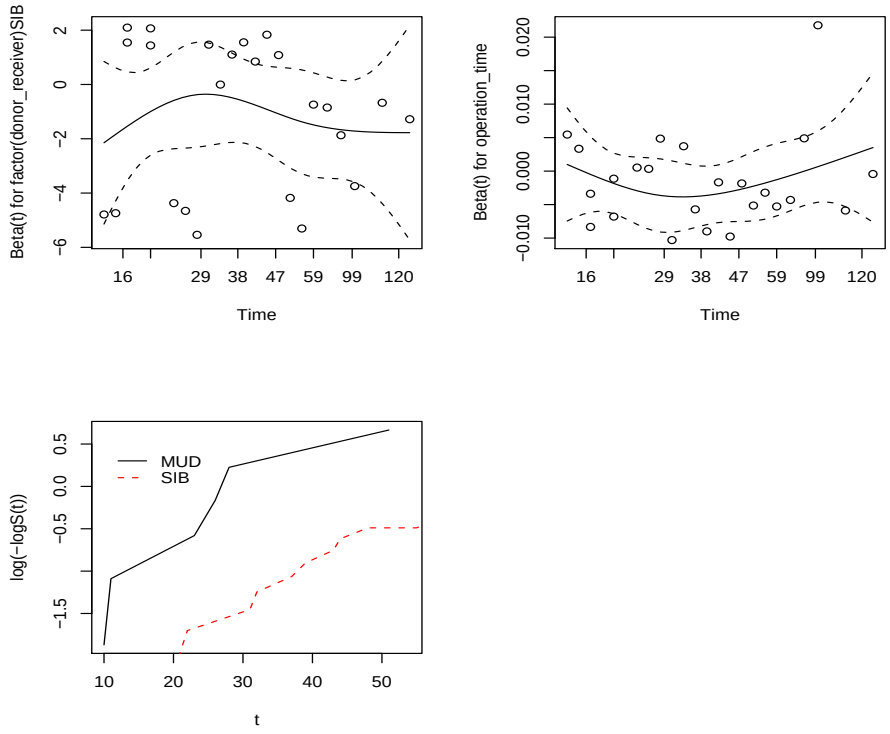

Figure 2.4: Diagnostic Plots of Constant Coefficients and Hazards Ratio

Table 2.5: Summary of Logistic Regression Analysis

Model 1:	GvHD ∼ BMTTYPE + BMT_time		LRT=6.453	P-value=0.040
Coefficients	Estimate	Std. Error	Z Value	p-value
Intercept	$1.949e+01$	$2.440e+03$	0.008	0.994
BMT type(SIB)	$-1.734e+01$	$2.440e+03$	−0.007	0.994
BMT_time	$-3.707e-03$	$2.531e-03$	−1.465	0.143
Model 2:	aGvHD ∼ BMTTYPE + BMT_time		LRT=10.419	P-value=0.005
Coefficients	Estimate	Std. Error	Z Value	p-value
Intercept	$1.985e+01$	$2.419e+03$	0.008	0.994
BMT type(SIB)	$-1.787e+01$	$2.419e+03$	−0.007	0.994
BMT_time	$-5.158e-03$	$2.962e-03$	−1.742	0.082
Model 3:	cGvHD ∼ BMTTYPE + BMT_time		LRT=0.135	P-value=0.935
Model 4:	Death ∼ BMTTYPE + BMT_time		LRT=0.739	P-value=0.691

Chapter 3

Exploring Lab Test Data

Proportions and concentrations of 123 peripheral blood mononuclear cells (PBMCs) from 31 patients were collected at multiple time points over the study course. These proportions and concentrations comprise the lab test data. The data studied in this thesis were taken from 16 days prior-transplantation to 399 days post-transplantation. The concentration values were obtained by multiplying each proportion value with the mononuclear cell (MNC) concentration of samples taken at the closest date. This would introduce additional systematic error and random error to concentrations of PBMCs. These additional errors make the concentration data less reliable, compared to the proportion data. We focus on the proportion values of the PBMCs.

The hypothesis of my study is that some PBMCs are associated with GvHD. We try to show patterns in PBMC proportions. I focus on the data obtained from date of transplantation to 100 days post-transplantation, which is the period when aGvHD may occur.

3.1 Proportions of Peripheral Blood Mononuclear Cells

The cluster of differentiation (often abbreviated as CD) is a protocol used for the identification and investigation of cell surface molecules (proteins) present on leukocytes. The CD system is commonly used as cell markers. A '+' or a '-' indicates whether a certain cell fraction expresses or lacks a CD protein. Cells can be defined based on what molecules are present on their surface. In flow cytometry, cells are typically stained by fluorochromes that are used to detect the presence of cell surface proteins. The PBMC samples were divided into ten aliquots in 96 well plates. Aliquots were stained with antibodies and were

named by targeted different immune cells. They are 1Activation, 2Activation, 3Activation, resting/activate (rest/act) T helper, rest/act T suppressor, T cell, myeloid, B cell, NK cell, and T cell receptor (TCR). There are 123 subsets of PBMCs in total. Among the 123 PBMCs, 3Activation_CD3-CD8lowCD122hi, Bcells_CD20+, TCR_CD3-TCRabgd-CD5+, and TCR_CD3+TCRab+CD5+ were only sampled from one patient. Patients 1, 2, 3, 4, 5 and 18 did not have observations for aliquots resting/activate (rest/act) T helper and rest/act T suppressor.

The original flow data from a flow cytometer are the counts of tested cells with light property parameters attached. The gating procedures classify cells as subtypes. Figure 3.1 shows how the proportion of a cell is obtained. The whole area is divided into four blocks by manual gating. The number at the corner of each block is the proportion of particles in that block. If FSC-A represents CD3, and SSC-A represents CD4, then the proportions of CD3-CD4-, CD3+CD4-, CD3+CD4+, and CD3+CD4- are 65.20, 0.01, 0.36 and 34.43 respectively. Since clusters' boundaries are usually rather vague, these PBMC proportions usually contain measurement errors. The measurement error highly depends on the technician who gates the data.

In clinical studies, it is common that observations are missing. Little and Rubin (1987) classify data missing as missing completely at random, missing at random and missing not at random. The first two types of missing data are referred to as ignorable. Most missing data in the clinical studies belong to this category. Some statistical techniques cannot handle missing values. In this situation, we need to insert artificial data to complete the data. There are many kinds of imputation. Here, we use the predicted values from non-parametric regression to replace missing values.

3.2 Data Cleaning and Manipulation

As mentioned in Chapter 1, the measurements were sparse and uneven. The collection was scheduled weekly for each patient. The actual patient visiting times shifted forward or backward. Some patients' data were not available at certain time points. Even with a blood sample taken, not all PBMC proportions could be obtained. Table 3.1 provide a summary of measurements in the interval $[-16,\ 399]$ days since BMT. If the last column is "YES", that means the same subset of PBMCs was not available over time for the patient. If the last column is "NO", then the corresponding patient has different missing blood cells over

time. The number in parenthesis in column 2 is the numbers of observations for a patient in the interval [0, 100]. A subset of PBMC samples might be missing. Figure 3.2 is the bar chart of a subset PBMC proportions of patient 1 on day 7. It shows that proportions of two PBMCs are missing and the sum of proportions of CD3+ and CD3- approximately equals 1 on day 7. It depends on the analysis methods. *Appendix A* gives the details about the 123 PBMC subtypes.

We first approximate the observations as collected on a common time basis. Since measurements were planed once a week for each patient, the proportions of PBMCs are rearranged in a grid with width equal to one week. The midpoints of the time windows are $(-13, -7, 0, 7, 14, 21, 28, 35, 42, 49, 56, 63, 70, 77, 84, 91)$ days to BMT. Each measure time is adjusted to the closest midpoint. If there were more than one value within a time window, the average of these values was used. The measurement taken before onset of aGvHD was considered as collected at a negative time. Because we want to predict the onset of aGvHD, another reasonable time scale is to take as the time 0 the date when a patient is diagnosed with aGvHD. Here I only look at data obtained on days -14, -7, 0, $+7$ in the 2nd scale The proportions of PBMCs are between 0 and 100, and the histograms have long right tails. It is necessary to normalize the data for further study. We consider Box-Cox transformation and the logistic transformation in the analysis.

3.3 Visualizing Longitudinal Data

What are the cross-sectional and longitudinal patterns of the PBMC proportions? What are the most important cells in describing variation of the lab test data? Is there a cell or a subset of cells defining clusters of patients such that the clinical outcomes of patients in each cluster belong to the same category?

Bearing those question on mind, we plot the PBMC proportions over time and draw loess curves of group means. loess is an extension of kernel smoothing to deal with unequally spaced data. It is less sensitive to outliers, because its weight function is no longer a function of fixed window width. The window width is negatively correlated with the density of data points near the estimated point. loess-smoothed curves are usually used to highlight aggregate patterns of longitudinal data. Figure 3.3 shows scatter plots for four selected cell types. Solid lines are the loess regression lines for the aGvHD group and dashed lines are the loess lines for the non-aGvHD group. For most PBMCs, the two smoothing lines are

intersect. Curves that do not cross over which be potential cellular indicators. We present those PBMCs in Table 3.3.

We fit the PBMC proportions with a non-parametric regression model, and obtain the residuals (observations minus expected values by loess regression) for each cell type. Figure 3.4 shows the residuals of two groups with repeated values from 10 patients connected for cell T_cells CD3+CD4+CD8b+, which is found by Lee (2006) as a potential cellular indicator. These patients had median residual value at the 0th, 25th, 50th, 75th, or 100th percentile. Note that the data for each patient tend to go through different levels of the cell proportion. Variation of the non-aGvHD group is greater than that of the aGvHD group. Variation increases as time goes by within a patient. Residual graphics of some other cells do agree with it, but some PBMCs' variations decrease along the time.

The pairwise residual plots of T_cells CD3+CD4+CD8b+ proportions with correlations are given in Figure 3.5. The residuals are obtained by fitting a linear regression model on the proportions of T_cells CD3+CD4+CD8b+ against time. Here we use the "windowlized" data from day 0 to day 42. Then we draw pairwise plots of the residuals for different fixed time points and calculate their correlations. In the graph, the small plots under the diagonal are pairwise residual plots. The plots on the diagonal are the histograms of residuals. "lmres" means residuals from linear regression models, and the numbers following it are the times at which the residuals are calculated. The correlations between two times are given in the squares above the diagonal. The graph shows that the correlations not only depend on distance between observations, but also depend on when observations were obtained. Particularly, the later two observations are obtained, the greater correlation is between them. This observation guides us to use an appropriate correlation structure in the later longitudinal data analysis.

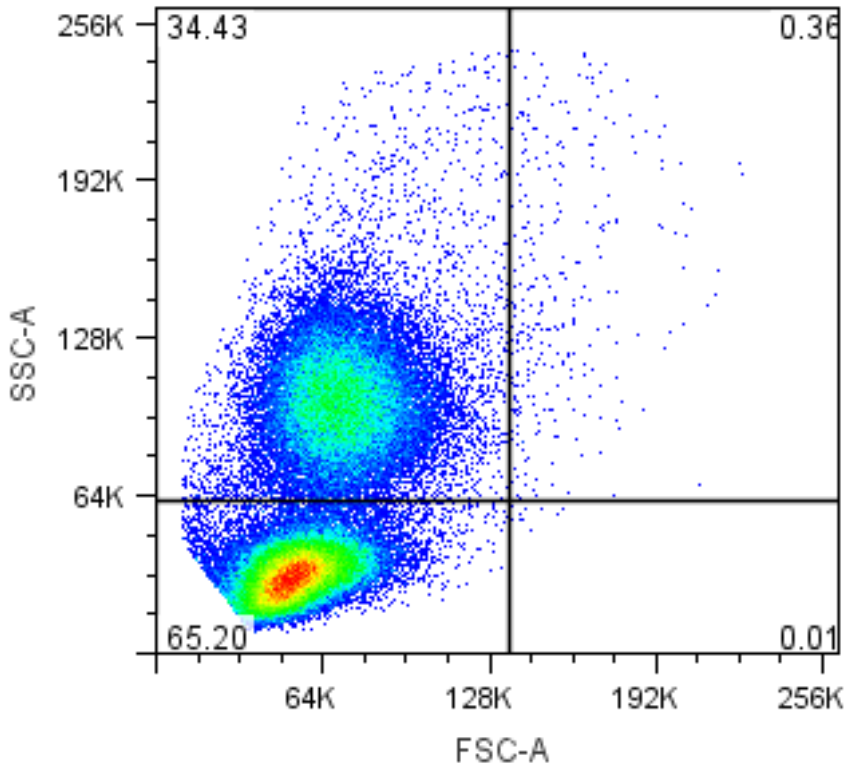

Figure 3.1: An Example of Gating

Table 3.1: Summary of Measurement

PID	# of Observations	Visiting Time (day)	Same Missing cell(s)
1	12(12)	0,7,14,21,28,32,39,46,53,60,67,74	YES
2	14(13)	-6,0,14,21,28,35,42,49,56,63,70,77,84,91	YES
3	8(8)	0,7,14,21,28,56,64,78	YES
4	15(12)	-6,0,4,11,18,32,39,47,53,60,74,81,88,116,122	YES
5	16(13)	-6,0,6,12,19,26,33,40,47,56,61,70,84,91,119,147	YES
6	16(14)	-8,0,5,12,19,27,32,39,46,53,60,67,76,83,90,176	NO
7	16(14)	-4,0,7,14,21,28,35,49,56,63,70,77,83,84,90,176	NO
8	16(12)	-8,-1,5,12,19,26,33,42,49,54,61,68,75,89,112,147	YES
9	17(15)	-6,0,6,11,18,25,32,41,48,55,62,69,78,83,90,97,111	YES
10	16(14)	-6,0,6,13,20,27,34,41,48,57,64,69,78,83,90,120	YES
11	15(14)	-10,0,6,12,19,30,36,43,50,61,69,75,83,90,97	YES
12	15(13)	-9,0,7,13,29,35,42,49,56,63,70,77,84,91,119	YES
13	16(14)	-9,0,8,15,20,29,36,43,50,57,64,71,78,85,92,120	YES
14	16(13)	-7,0,7,12,21,28,35,42,49,56,63,77,84,91,125,146	YES
15	14(13)	0,7,13,21,28,35,42,49,56,63,70,85,91,118	NO
16	12(10)	-8,-1,10,19,24,31,38,45,52,59,66,73	YES
17	11(9)	-16,-7,6,34,41,55,61,68,75,82,89	YES
18	8(7)	-6,0,13,27,34,41,50,55	YES
19	15(14)	0,7,14,28,38,41,49,56,62,70,77,84,91,98,118	YES
20	11(10)	-2,0,7,14,21,28,35,42,49,56,63	YES
21	15(13)	0,7,14,21,29,36,43,50,56,64,78,85,92,120,148	YES
22	9(8)	-8,0,6,13,20,34,55,69,76	YES
23	15(14)	-3,0,7,14,21,28,36,43,50,56,63,70,77,84,91	NO
24	14(12)	-9,0,14,21,28,37,44,58,63,68,75,82,89,149	YES
25	10(8)	-10,-3,1,8,15,22,29,37,43,50	YES
26	17(10)	-4,0,6,13,21,27,62,69,76,83,90,118,145,173,235,236,399	NO
27	18(13)	-4,0,7,14,21,28,35,42,49,56,70,84,91,98,105,126,161,175	NO
28	18(14)	-6,0,7,14,21,28,35,42,49,55,70,77,84,91,98,119,142,174	YES
29	10(9)	-9,0,7,17,24,45,52,73,80,90	YES
30	12(11)	-8,0,7,14,21,32,39,48,54,61,68,77	YES
31	9(7)	-9,-1,5,13,19,26,33,68,82	YES

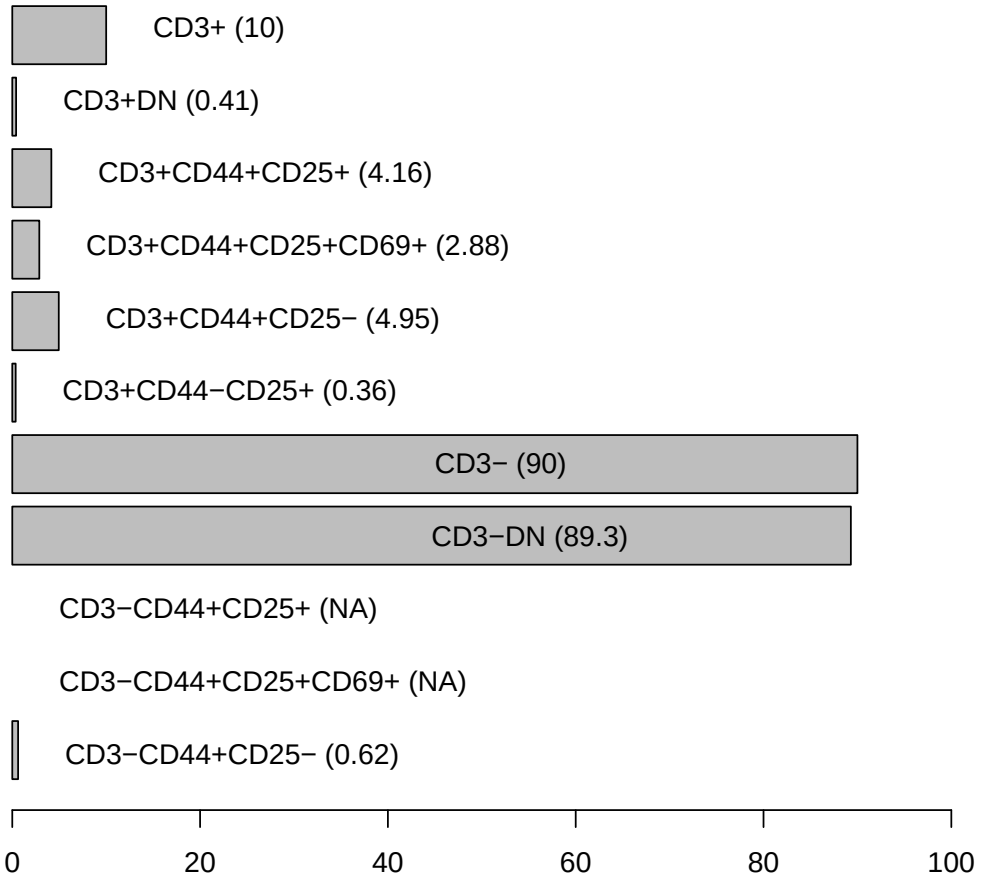

Figure 3.2: Proportions of a Subset of PBMCs for Patient No.1 on day 7 (Numbers in the Parentheses are Proportions, NA Means "not available")

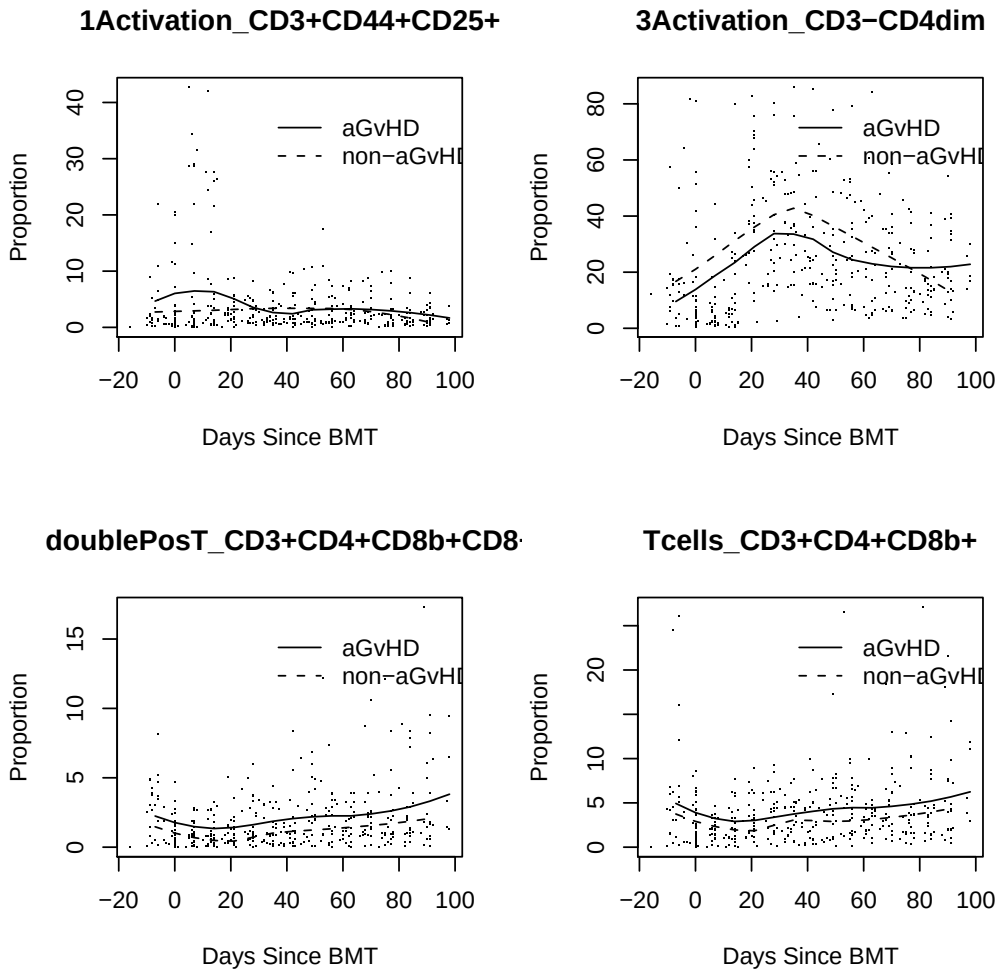

Figure 3.3: Proportions of Four PBMCs against Time since BMT with Loess-smoothed Curves.

Table 3.2: PBMCs Whose Mean Curves of Different Groups Show a Parallel Pattern

Aliquot	Cell Name	Aliquot	Cell Name
3Activation	CD3+CD8br	restactTs	CD45RACD3-CD8
doublePosT	CD3+CD4+CD8b+CD8+		CD45RACD3+CD8-
NK	CD2+CD16+CD56+CD3-		CD45RACD3+ofParent
restactTh	CD45RACD3+CD4+		CD8+CD3-ofParent
	CD45RACD3+CD4low	T cells	CD3+CD8+CD8b-
	CD45RACD3+ofParent		CD3+CD8+CD8b+
	CD45ROCD3+CD4low		
	CD3+CD4-ofParent		

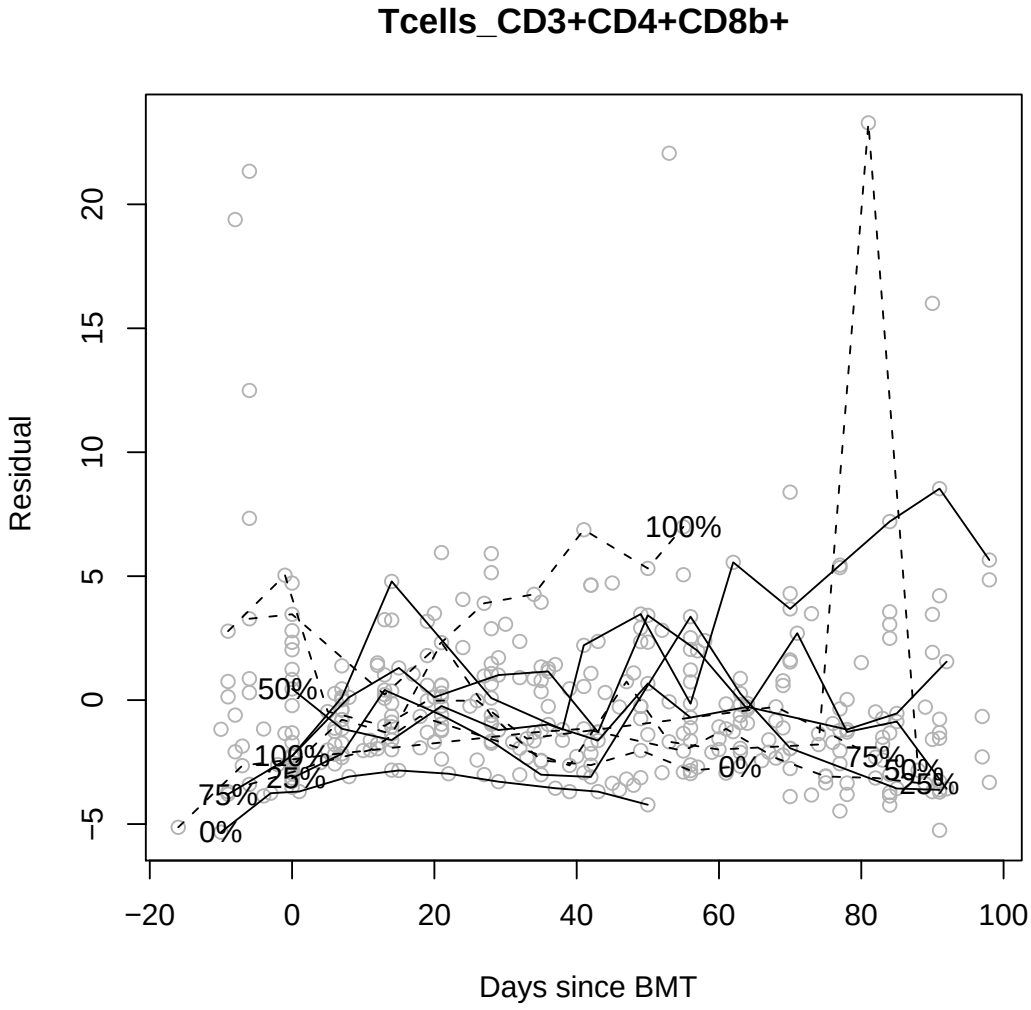

Figure 3.4: Non-parametric Regression Residuals against Time since BMT. Solid Lines: from the aGvHD Group; Dashed Lines: from the non-aGvHD Group.

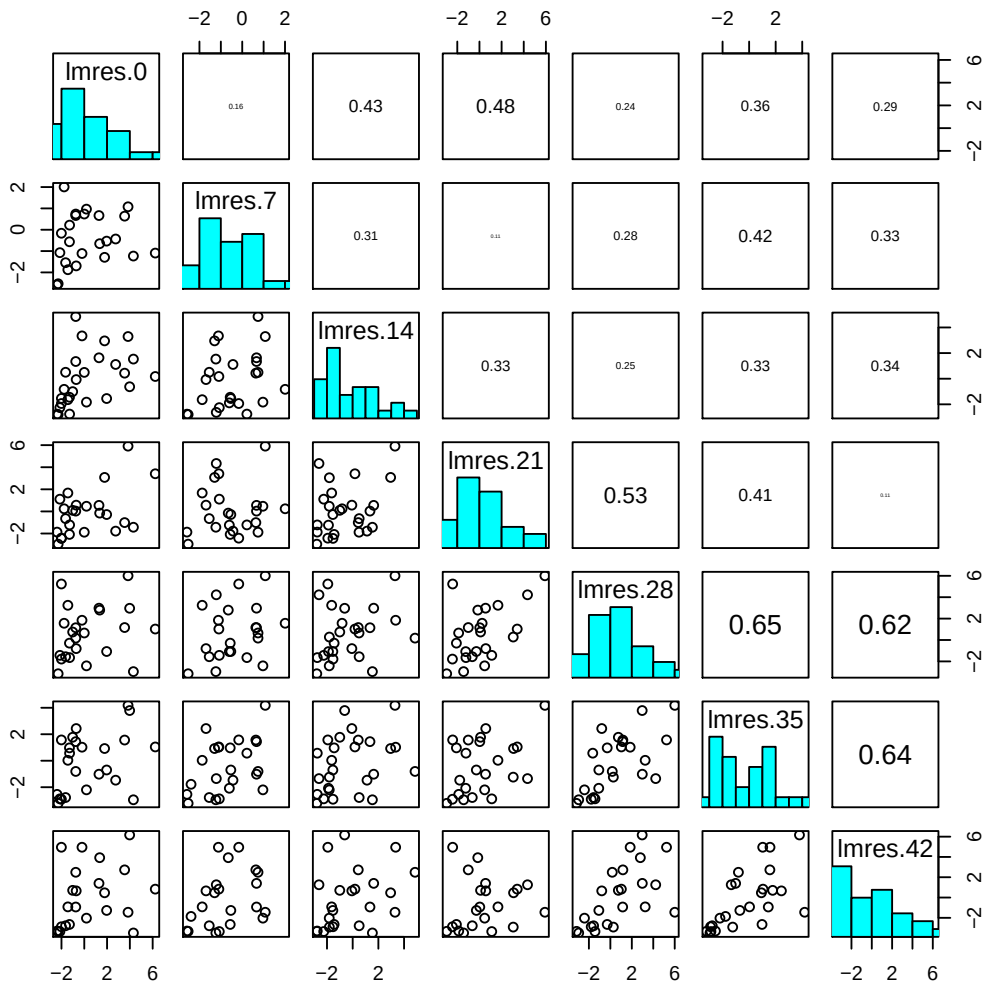

Figure 3.5: Scatter Plot Matrix of Residuals at Different Time for T_cells CD3+CD4+CD8b+

Chapter 4

Analysis of PBMC Data at a Fixed Time Point

This chapter focuses on the "windowlized" as described in Chapter 3. We apply multivariate analysis methods to investigate the PBMC data at each of the time points. We aim at the following:

1. to reduce the dimensionality of the PBMC data;
2. to partition patients and blood cells into clusters and explore the data structure;
3. to investigate the relationship between proportions of PBMCs and aGvHD.

4.1 Reducing the Dimension of PBMC Data

For each of the 31 patients, there were 123 types of PBMC cells. It is neither feasible nor efficient to use them directly in the analysis. A commonly used approach to reduce the data dimension is principal components analysis (PCA) .

PCA transforms the data to a new coordinate system, in which the first coordinate maximizes the variance of the data, the second coordinate is the direction of the second greatest variance, and so on. By keeping lower-order principal components and ignoring higher-order ones, we reduce the dimensionality and keep the most important aspects of the PBMC data.

Denote the $n = 31$ observation vectors by $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n}$, each with dimension $p = 123$. These form a swarm of points in a p-dimensional space. By multipling each $\mathbf{y_i}$ with an orthogonal matrix $\mathbf{A}$, $\mathbf{y_i}$ is transformed to a point $\mathbf{z_i}$ without changing the distance to the origin. With an appreciate choice of $\mathbf{A}$, the covariance matrix of $\mathbf{z_i}, \ldots, \mathbf{z_n}$, $\mathbf{S_z} = \mathbf{ASA'}$, is a diagonal matrix, where $\mathbf{S}$ is the sample covariance matrix of $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n}$. The principal components (PCs) are the transformed variables $z_j,\ j = 1, \ldots, p$, the components of $\mathbf{z} = \mathbf{Ay}$. The proportion of variance explained by the first k components is:

$$Proportion\ of\ Variance = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\sum_{j=1}^{p} s_{jj}}$$

where λ_j is the j^{th} largest eigenvalues of $\mathbf{S}$. If the variables are highly correlated, the essential dimensionality is much smaller than p, and the first few eigenvalues will be relatively large (Rencher, 2002). We can then use a space with lower-dimension to represent the original space without losing much information.

Because there were many missing values at each time point, we chose two subsets of data to conduct PCA. One (Set A) was chosen to attain the maximum number of patients, but sacrificing the number of PBMCs. Another (Set B) was chosen to attain the maximum number of PBMCs. Dataset A and dataset B were roughly consistent over different time points. We used statistical software R to obtain principal components and dimension-reduced data for further analysis.

According to the results of PCA, the first four principal components counted for over 80% of the variation no matter which data set we used and at which time point. Therefore, we can reduce dimensions of the data by projecting data on a space that is formed by first 4 or 6 most important principal components. Figure 4.1 was randomly chosen from 28 plots of PCs cumulative weight. "d28A" means data set A on 28^{th} day post transplant. It shows that the first four PCs cumulative weight is about 83%. Figure 4.2 is a plot of the first two PCs. The plot can visualize high dimensional data on a plane. This will help to identify outliers and any tendency toward clusters of points. Figure 4.2 shows that patient $^{\#}20$ is a potential outlier. Here, a patient ID with suffix "a" indicates that the corresponding patient has aGvHD, and a patient ID with suffix "u" indicates that the corresponding patient was not used in Lee's study.

Data dimensions can also be reduced by selecting most important PBMCs. Weights of variables were used to select PBMCs. The weight of PBMC x in the i^{th} principal component

(PC_i) is defined as

$$weight(x) = \text{percentage of explained variance of } PC_i \times (\text{coefficient of x})^2$$

The cells with large sum of weight(x) can likely be the important ones. After processing the data with this method, we find that cells with greater proportion values such as TCR CD3+, leukocyte CD45+CD33-, and 1Activation CD3+ have higher probabilities being selected as the important variables. This method appears not appropriate to deal with data with such imbalanced values.

Analysis can focus on data with each aliquot. Then the dimension of data of interest is at most 18. However blood cells in an aliquot are in the same blood cell category. If cells from different categories predict aGvHD, we can not identify them.

PCA was also conducted with data only from patients diagnosed with aGvHD and in the time scale with day 0 being the onset time of aGvHD. The first five principal components account for over 80% of the total variance. This holds for all data sets. However the PCA outcomes with the PBMC data at different times are not in agreement. They provide different principal components which is not desirable.

4.2 Partitioning Patients

Clustering partitions a data set into subsets (clusters) according to some defined distance measure. There are two common approaches for clustering: hierarchical clustering and partitioning. Cluster analysis can find an optimal grouping for which the observations or objects within each cluster are similar, but the clusters are dissimilar to each other (Rencher, 2002).

To explore the structure of the PBMC proportions, we mainly consider hierarchical clustering in partitioning patients or PBMCs. Figure 4.3 gives two dendrograms from the cluster analysis. The left plot is the dendrogram from unsupervised clustering. it shows there are three big clusters and one outlier (P20). It is consistent with Figure 4.2. A heat map can show clusters of patients and PBMCs simultaneously (Figure 4.4), with different colors for different clusters.

Classical clustering groups multivariate data based on the distance between observations only. If there is considerable overlap of two groups, objects could be assigned to the wrong

group. Since the proportion values of PBMCs are likely correlated with the clinical outcomes, the information on clinical outcomes can be a guide to group better the PBMC data. This type of clustering, called supervised-clustering, can improve the accuracy of clustering (Bair and Tibshirani 2004).

We conduct cluster analyses with the raw data and with the Box-Cox transformed data. There is no consistent pattern over time for either raw data or transformed data. The plots in Figure 4.3 show there are some changes compared to the unsupervised-clustering. Nonetheless, the improvement is not significant.

4.3 Relationship between PBMC and Clinical Outcomes

We conduct MANOVA and generalized linear regression analysis to explore the association PBMCs and the disease aGvHD.

4.3.1 Multivariate One-Way Analysis of Variance (MANOVA)

Patients were divided into two groups: aGvHD and non-aGvHD. Since the original data have dimensions larger than the sample size, we carried out one-way MANOVA with the dimension-reduced data obtained from the PCA. The first 5, 7 and 9 PCs were used in the analysis one by one. The Lawley-Hotelling test only detected a significant difference at 0.05 level based on the proportions from the 91^{th} day with nine dimensions. It may be too late to predict aGvHD. MANOVA on each aliquot found eight subsets of the cells with significant difference between the two groups. Table 4.1 provides details about the analysis.

4.3.2 Stepwise Selection of Variables

We conduct a data-directed search for the variables that can best separate the groups. This approach is called as stepwise discriminant analysis. The forward selection method uses the value of Wilk's Λ statistic to form the criteria for selection. At each step, we select the variable with minimum $\Lambda(y|.)$ or maximum F, where the dot in the parenthesis denotes previously selected variable(s). Continue the process until the Λ or F reaches some predetermined threshold value. A stepwise procedure follows a similar sequence, except that after a variable has entered, the variables previously selected are reexamined to see if each still contributes significantly. Here we use the stepwise procedure. The variable selection

is performed on two data sets at each fixed time point. The output of variable selection is provided in Table 4.3.3 There is no cell that is important across the all data sets at different time. Of 37 selected cells, 13 cells are nature killer cells, and 10 cells belong to leukocytes.

4.3.3 Generalized Linear Regression

We view aGvHD onset as an event and the PBMC data or or dimension-reduced PBMC data as predictors, and consider a logistic regression model. None of the groups of PBMCs is significantly associated with the event of aGvHD onset.The output of analysis seemed strange because the most residual deviances were 0. This needs further investigation.

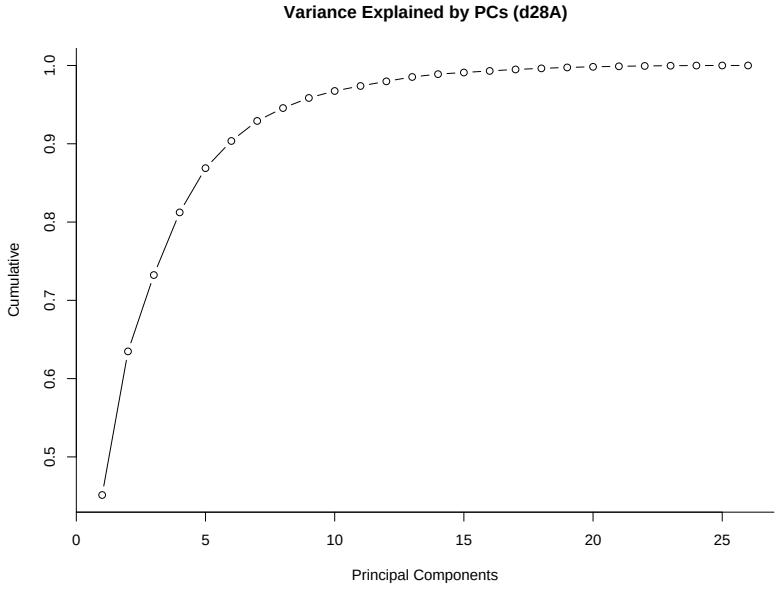

Figure 4.1: Cumulative Weight of Principal Components for Day 28 Data Set A

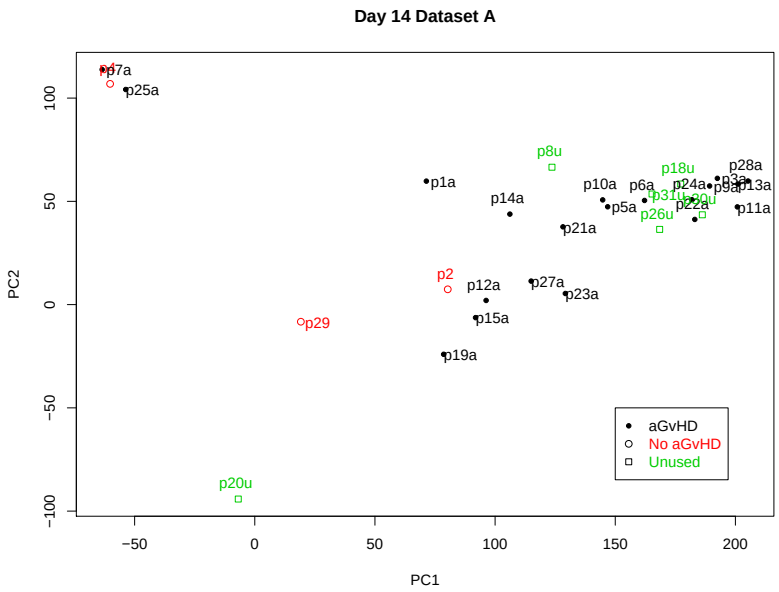

Figure 4.2: First Two Principal Components for Day 14 Data Set A.

Table 4.1: Summary of Fixed Time Points MANOVA

Data Set	Lawley-Hotelling Test P-Value	Comment Comment
day 91 A	0.039	use 9 PCs
day 7, NK cell	0.028	per aliquot
day 14, T cell	< 0.001	per aliquot
day 21, Leukocytes	0.015	per aliquot
day 28, NK cell	0.010	per aliquot
day 56, 1Activation	0.036	per aliquot
day 63, 1Activation	0.045	per aliquot
day 63, T cell	< 0.001	per aliquot
day 84, NK cell	0.018	per aliquot

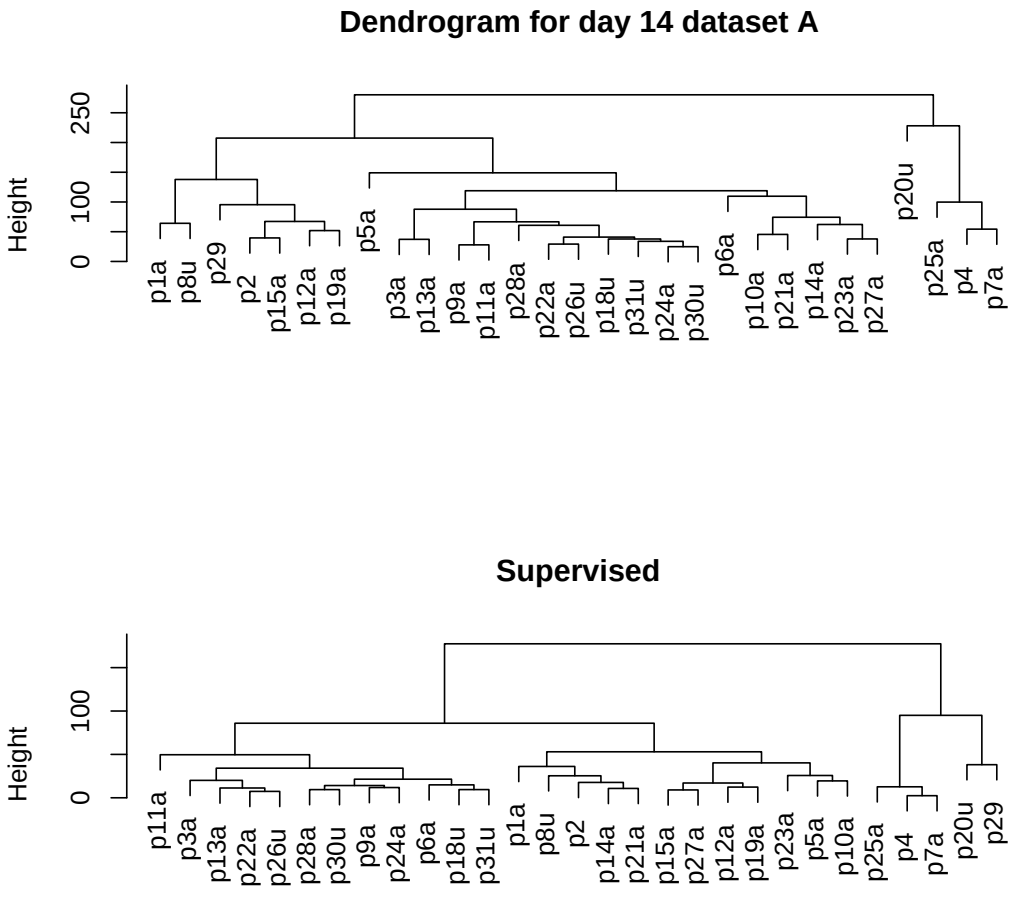

Figure 4.3: Dendrogram for Data Set A on Day 14.

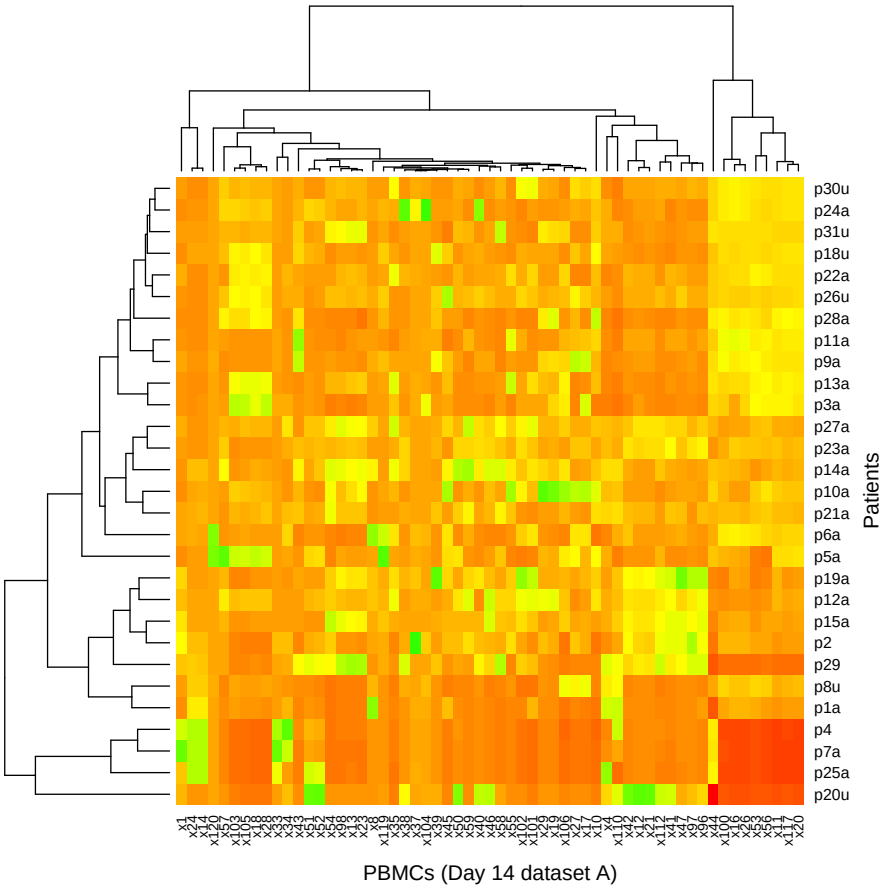

Figure 4.4: Heat Map for Data Set A on Day 14.

Table 4.2: Output of Stepwise Selection

Day	Data Set	Cell Name
0	A	1Activation_CD3-DN,3Activation_CD3-DN
0	B	Leukocytes_CD33+CD45dimP,NKcells_CD2+CD16-CD56+CD3-, TCR_CD3+TCRab+CD5+TCRabgdCD5+
7	A	3Activation_CD3+CD8dim,Leukocytes_CD33+CD45dimP
7	B	3Activation_CD3+CD8dim,Leukocytes_CD33+CD45dimP
14	A	2Activation_CD3-CD8low,NKcells_CD2-CD16+DN
14	B	NKcells_CD2-CD16+DN,NKcells_CD2+CD16+CD56+CD3-, NKcells_CD2+CD16+DN
21	A	NKcells_CD2dimCD16+CD56+CD3-
21	B	NKcells_CD2-CD16+CD56+CD3-,restactTh_45RACD3+CD4low
28	A	Leukocytes_CD33+CD45dimCD15+CD14-, NKcells_CD2+CD16-CD56+CD3-
35	A	Leukocytes_CD33+CD45dimCD15+CD14-, NKcells_CD2-CD16+CD56+CD3-, NKcells_CD2+CD16+CD3+CD56-
35	B	NKcells_CD2-CD16+P,Tcells_CD3+CD8+CD8b-,TCR_CD3-5+
42	A	Leukocytes_CD33+CD45dimCD15loCD14lo, Leukocytes_CD33+CD45dimCD15+CD14-
42	B	NKcells_CD2-CD16+P,restactTs_45RACD3-CD8, restactTs_CD8+CD3-P
49	A	NKcells_CD2-CD16+CD56+CD3-,Tcells_CD3-CD8+CD8b-, Tcells_CD3+CD8b+CD8low
56	A	3Activation_CD3-CD8low,Leukocytes_CD33+CD45dimP
70	A	Leukocytes_CD33+CD45dimCD15+CD14-
70	B	Leukocytes_CD33+CD45dimCD15+CD14-

Chapter 5

Longitudinal Analyses of PBMC Outcomes

Chapter 4 separately analyzes data collected at different times. No particular PBMC or subset of PBMCs is found to distinguish patients with aGvHD from patients without aGvHD at a fixed time point. Change of PBMC proportions over time may disclose a certain relationship between PBMCs and aGvHD. In this chapter, we apply longitudinal data analysis methods to analyzing the "windowlized" data and the raw (the original) data.

5.1 MANOVA of Repeated Measures

We consider MANOVA models with the "windowlized" data. In order to carry out the MANOVA procedure, we imputed the missing data based on a non-parametric regression analysis (Yokoyama 1995). For each patient, the proportion of a peripheral blood mononuclear cell was measured over time. We consider these measurements together as a response vector, and take status of disease, and PBMC types as the factors. Let y_{ijkl} be the transformed proportion of PBMC k at time l of patient i in disease group j. We assume proportions of PBMCs are independent. A two-way fixed effects MANOVA model is, for $\mathbf{y_{ijk}}$ with components y_{ijkl},

$$\mathbf{y_{ijk}} = \mu + \alpha_{\mathbf{j}} + \gamma_{\mathbf{k}} + (\alpha\gamma)_{\mathbf{jk}} + \boldsymbol{\epsilon}_{\mathbf{ijk}},$$

where μ is the grand mean vector for the whole population, α_j and γ_k are the main effects of disease, and PBMC respectively, $(\alpha\gamma)_{jk}$ is corresponding interaction effect, and the random

error vector $\boldsymbol{\epsilon}_{ijk}$'s are independently distributed as $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Here p equals the number of repeated measures of a peripheral blood mononuclear cell, and Σ is the symmetric variance-covariance matrix.

Since most aGvHDs occurred within 6 weeks post transplant, observations on day 0, 7, 14, 21, 28, 35, 42 were included in the analysis. To make the analysis feasible, we fitted the model with each aliquot data. The MANOVA table (Table 5.1) shows that the main effect of aGvHD for T_cells is significant (p-value = 0.045). Proportions of T cells over time significantly separate the patients with aGvHD and without aGvHD. This is in agreement of the founding in Lee (2006). In addition, it appears the GvHD outcomes are significantly different across PBMC types within each aliquot.

The independence of PBMCs is not a reasonable assumption because of the sequence gating of flow cytometry data. A MANOVA model with random effects may describe our data more appropriately. Yokoyama (1994) provides detail on statistical inference of mixed MANOVA models.

5.2 Analysis of the Raw Data

To accomplish the goal of developing a model to predict aGvHD, we model the transformed proportions of PBMCs as a function of the disease status z_i (a factor with two levels), PBMC type (a factor with K levels) and the collection time t_{ikl} (a continuous variable). Let y_{ikl} be patient i's l^{th} transformed PBMC proportion of PBMC k. The linear regression model is

$$y_{ikl} = \mu + \alpha z_i + \beta_k + \gamma t_{ikl} + \epsilon_{ikl}, \qquad \text{(Model 1)}$$

where t_{ikl} is the measurement time of l^{th} measurement of PBMC k for patient i, z_i is an indicator with $z_i = 1$ or 0 for patient i having aGvHD or not, β_k is the fixed effect of PBMC k, and ϵ_{ikl} are independent random errors with identically-distributed as $N(0, \sigma^2)$.

It is reasonable to regard the patients as a random sample from a population. So a random intercept model is considered as

$$y_{ikl} = U_i + \alpha z_i + \beta_k + \gamma t_{ikl} + \epsilon_{ikl}, \qquad \text{(Model 2)}$$

where U_i are the random effects of patient grouping factor and independent and identically-distributed as $N(\mu, \sigma_1^2)$, ϵ_{ikl} and U_i are independent of each other.

We also consider a mixed effects model with a random intercept and a random slope, that is

$$y_{ikl} = U_i + \alpha z_i + \beta_k + V_i t_{ikl} + \epsilon_{ikl}, \qquad \text{(Model 3)}$$

$$\text{with } U_i \overset{iid}{\sim} N(\mu, \sigma_1^2), \text{ and } V_i \overset{iid}{\sim} N(\gamma, \sigma_2^2),$$

where V_i are the random effects of time, and γ is the overall slope. We have fitted this model with V_i and U_i are independent or correlated. Of the ten aliquots, only 1Activation associated data indicate that U_i and V_i may be independent (p-value of Log-likelihood Ratio test (LRT) is 0.57).

The data structure shows two different classifications, Patient ID and PBMC. A two-level model would describe data better. LRTs indicate that it is significant better adding two-level structures in the corresponding models. If we only consider a random intercept, then a two-level mixed effects model with a random intercept is

$$y_{ikl} = U_i + B_{ik} + \alpha z_i + \gamma t_{ikl} + \epsilon_{ikl}, \qquad \text{(Model 4)}$$

$$\text{with } U_i \overset{iid}{\sim} N(\mu_{,}\sigma_1^2), \text{ and } B_{ik} \overset{iid}{\sim} N(\beta_{ik}, \tau_1^2),$$

where B_{ik} are the random effects of PBMC:patient grouping factor, and independent of U_i. Random effects in different levels are assumed independent.

If we consider random effects of measurement time at the patient level, then another model is

$$y_{ikl} = U_i + B_{ik} + V_i t_{ikl} + \alpha z_i + \epsilon_{ikl}, \qquad \text{(Model 5)}$$

$$\text{with } U_i \overset{iid}{\sim} N(\mu, \sigma_1^2),\ B_{ik} \overset{iid}{\sim} N(\beta_{ik}, \tau_1^2), \text{ and } V_i \overset{iid}{\sim} N(\gamma, \sigma_2^2),$$

where V_i is the random effect of time at the patient level,

If we consder random effects of time at the PBMC level in addition to Model 5, it gives

$$y_{ikl} = U_i + B_{ik} + V_i t_{ikl} + V_{ik} t_{ikl} + \alpha z_i + \epsilon_{ikl}, \qquad \text{(Model 6)}$$

$$\text{with } U_i \overset{iid}{\sim} N(\mu, \sigma_1^2),\ B_{ik} \overset{iid}{\sim} N(\beta_{ik}, \tau_1^2),\ V_i \overset{iid}{\sim} N(\gamma_i, \sigma_2^2), \text{ and } V_{ik} \overset{iid}{\sim} N(\gamma_{ik}, \tau_2^2),$$

where U_i and V_i, B_{ik} and V_{ik} are correlated.

We can use the log-likelihood ratio test to approximately test whether B_{ik} is independent of V_{ik}, the p-value of LTR is less than 0.001 in aliquot 9 (T_cells). The intercept and slope at the PBMC level is correlated for most aliquots.

The above models assume that the within-group observations are uniformly correlated. It rarely holds in practice. The correlation between two successive observations is likely larger than others. Some statistical softwares incorporate a flexible mechanism for specifying correlation structures. Most of these correlation structures are appropriate only for equally spaced observations. A continuous first-order autoregressive process in the error can be used to describe continuous time. We consider Model 6 with first-order autoregressive correlation structure, i.e. Model 7.

The model selection is based on the log-likelihood ratio test and with reference to Akaike's Information Criterion (AIC) and Bayesian Information Criteria (BIC). When we fit the above mixed effects models, we also want to consider the interactions between PBMC, disease status and the data collection time. The three-factor interaction term in the model is not significant for all aliquot data. So we do not include it in the final model. The final model is the selected model pluses two-factor interactions. The final model is

$$y_{ikl} = U_i + B_{ik} + V_i t_{ikl} + V_{ik} t_{ikl} + \alpha z_i + \theta_1 z_i * t_{ikl} + \theta_2 z_i * x_k + \theta_3 t_{ikl} * x_k + \epsilon_{ikl}, \text{ (Model 8)}$$

$$\text{with } U_i \overset{iid}{\sim} N(\mu, \sigma_1^2),\ B_{ik} \overset{iid}{\sim} N(\beta_{ik}, \tau_1^2),\ V_i \overset{iid}{\sim} N(\gamma_i, \sigma_2^2), \text{ and } V_{ik} \overset{iid}{\sim} N(\gamma_{ik}, \tau_2^2),$$

where U_i and V_i, B_{ik} and V_{ik} are correlated. x_k is a PBMC indicator variable, if k is specified, $x_k \equiv 1$. θ_1 is the interaction effect of measuring time and aGvHD. θ_2 is the interaction effect of the k^{th} PBMC and aGvHD. θ_3 is the interaction effect of the k^{th} PBMC and time.

Because of the large number of PBMC types, we fit the models with each aliquot one by one. The comparison of linear regression model and random effects models for T_cells is given in Table 5.3. It shows that the two-level model is significantly better than one level model. The proportion data of four aliquots reached the iteration limit reached without convergence (B cell, Leukocyte, T helper, and T suppressor) when we fitted Model 7. For the other aliquots, Model 6 is the best model based on log-likelihood ratio test. Model 8 is a two-level hierarchical model with correlated intercepts and slopes at each level. Residual plot (Figure 5.1) also shows Model 8 has the best fitting. The p-values of testing significance of interaction between fixed effects of T_cells CD3-CD8bdimCD8-, CD3-CD4lowCD8low and CD3-CD8+CD8b- and aGvHD are 0.018,0.002 and 0.019 respectively, when using Model 8. Table 5.3 shows interaction of aGvHD and PBMCs in the T_ cells aliquot is significant at 0.05 level. For PBMCs in the other aliquots, we did not find a strong association of PBMC with aGvHD.

The likelihood ratio test (LTR) provides an evidence that variances of the random effect for the patient and PBMC grouping factors are not ignorable: the LTR is significant when adding grouping factors of patient and PBMC in the models. We also use simulation methods to test the random effects of the patient. The p-value of the test is less that 0.001. The estimates associated with the random effects for aliquot 9 are given in Table 5.4.

5.3 Trends of PBMC Data

Proportion of PBMC changes over time. The slopes of the linear regression lines of the proportions against time may indicate the trends of PBMCs. The trends of PBMC data may be different between the patients with aGvHD and the patients without aGvHD. We conduct an ANOVA to investigate whether the mean of slopes are different among the patient group and PBMC group. The ANOVA model is

$$b_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk},$$

where b_{ijk} is the slope of the patient i with PBMC k in j^{th} disease group, α_j is the main effect of the aGvHD, β_k is the main effect of the PBMC k, $(\alpha\beta)_{jk}$ is the interaction between aGvHD and the PBMC, and $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$.

Table 5.5 provides the ANOVA outcomes. The mean slopes for T_helper and T_suppressor are different between aGvHD and non-aGvHD groups. But when we look at the each cell in these two aliquots, we cannot find a PBMC that separates the slopes of the two patient groups (Figure 5.2).

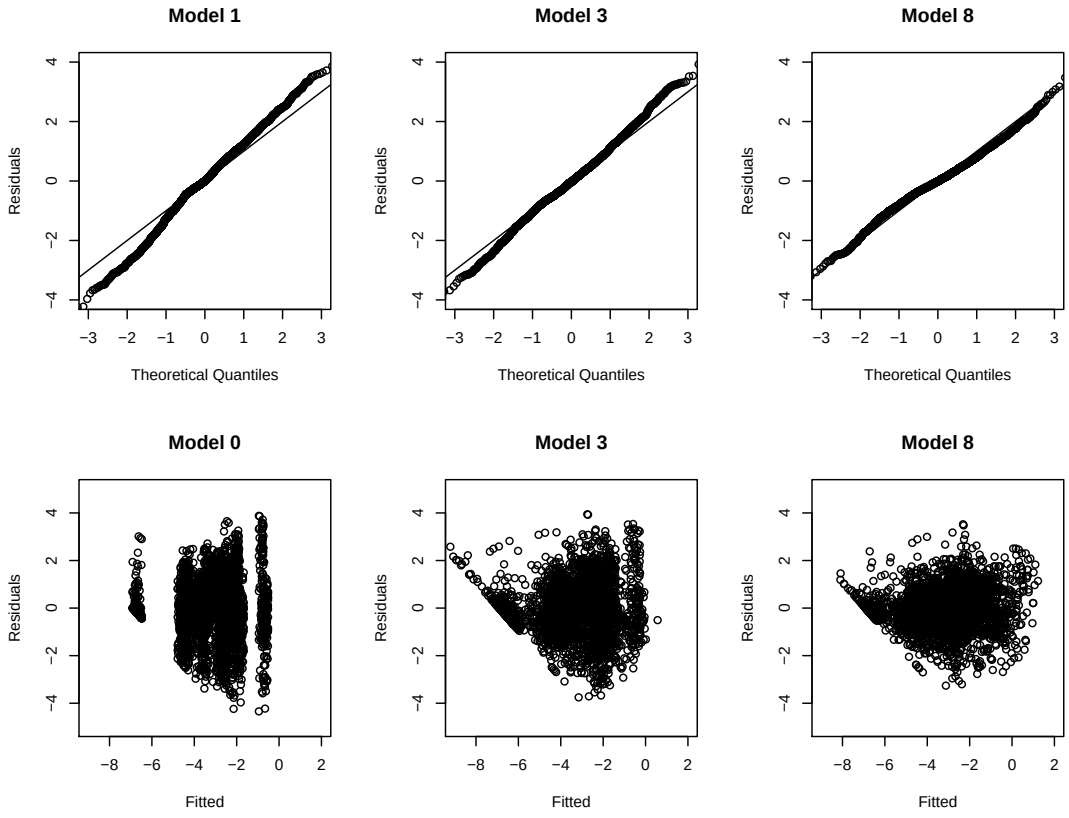

Figure 5.1: Diagnosis Residual Plots

Table 5.1: MANOVA Table for Repeated Measures

Aliquot	Source	Df	Pillai	approx F	num Df	den Df	Pr(>F)
1Activatation	agvhd	1	0.022	0.544	7	168	0.800
	pbmc	5	1.194	7.704	35	860	< .001
	agvhd:pbmc	5	0.148	0.749	35	860	0.855
	Residuals	174					
2Activation	agvhd	1	0.004	0.138	7	226	0.995
	pbmc	7	1.351	7.928	49	1624	< .001
	agvhd:pbmc	7	0.140	0.676	49	1624	0.958
	Residuals	232					
3Activation	agvhd	1	0.014	0.388	7	197	0.909
	pbmc	6	1.226	7.413	42	1212	< .001
	agvhd:pbmc	6	0.130	0.641	42	1212	0.964
	Residuals	203					
B cells	agvhd	1	0.093	0.758	7	52	0.625
	pbmc	1	0.275	2.814	7	52	0.015
	agvhd:pbmc	1	0.056	0.444	7	52	0.870
	Residuals	58					
Leukocytes	agvhd	1	0.005	0.176	7	226	0.990
	pbmc	7	1.400	8.283	49	1624	< .001
	agvhd:pbmc	7	0.189	0.921	49	1624	0.630
	Residuals	232					
NK cells	agvhd	1	0.015	0.824	7	371	0.568
	pbmc	12	1.298	7.151	84	2639	< .001
	agvhd:pbmc	12	0.179	0.824	84	2639	0.874
	Residuals	377					
T cells	agvhd	1	0.047	2.088	7	298	**0.045**
	pbmc	11	1.568	7.977	77	2128	< .001
	agvhd:pbmc	11	0.246	1.006	77	2128	0.465
	Residuals	304					
TCR	agvhd	1	0.036	0.561	7	104	0.786
	pbmc	4	1.474	8.916	28	428	< .001
	agvhd:pbmc	4	0.182	0.729	28	428	0.844
	Residuals	110					

Table 5.2: Comparing Regression Models for Aliquot 9 (T Cells)

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
Model3*	18	9061.9	9087.7	−4512.9	3 vs 2	116.4	< .001
Model2*	16	9174.3	9197.2	−4571.2	2 vs 1	438.9	< .001
Model1	15	9611.1	9700.4	−4790.6	1 vs Null	2834.6	< .001
Model4	17	9108.7	9113.0	−4551.4	4 vs 2	89.0	< .001
Model2	16	9195.7	9198.5	−4595.9			
Model5	19	8972.4	8979.6	−4481.2	5 vs 3	111.7	< .001
Model3	18	9082.1	9078.9	−4537.1			
Model6	21	8685.1	8695.1	−4335.6	6 vs 5	291.3	< .001
Model5	19	8972.4	8979.6	−4481.2	5 vs 4	140.4	< .001
Model4	17	9108.7	9113.0	−4551.4			
Model7	20	8766.4	8775.0	−4377.2	6 vs 7	83.3	< .001
Model6	21	8685.1	8695.1	−4335.2			
Model8	44	8478.4	8488.4	−4232.2	8 vs 6	206.7	< .001
Model6	21	8685.1	8695.1	−4335.2			

*: Using Maximum Likelihood method to estimate parameter

Table 5.3: Test of Fixed Effects for Aliquot 9

Source	numDF	denDF	F-value	p-value
PBMC	11	319	173.65	< .001
agvhd	1	2111	0.31	0.581
time	1	29	2.16	0.152
PBMC:agvhd	11	2111	1.96	0.029
time*cn	11	330	40.19	< .001
time*agvhd	1	2111	0.01	0.918

Table 5.4: Summary of Random Effects (aliquot 9)

Cov Parm	Subject	Estimate	SE	Z-Value	p-value
σ_1^2	pid	0.5359	0.1596	3.36	0.0004
$COV(U_i, V_i)$	pid	−0.00771	0.003206	−2.41	0.0161
σ_2^2	pid	0.000215	0.000086	2.51	0.0061
τ_1^2	pid(PBMC)	0.5498	0.09191	5.98	< .001
$COV(B_{ik}, V_{ik})$	pid(PBMC)	−0.01878	0.002849	−6.59	< .001
τ_2^2	pid(PBMC)	0.000942	0.000108	8.73	< .001

Table 5.5: Some Results from ANOVA of Slopes

Aliquot	Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
1Activation	agvhd	1	0.00002	0.00002	0.052	0.821
	PBMC	7	0.01431	0.00204	5.191	< .001
	agvhd:PBMC	7	0.00061	0.00009	0.223	0.979
2Activation	agvhd	1	0.00031	0.00031	0.714	0.401
	PBMC	7	0.01380	0.00197	4.475	< .001
	agvhd:PBMC	7	0.00116	0.00017	0.377	0.913
3Activation	agvhd	1	0.00048	0.00048	1.244	0.272
	PBMC	6	0.01138	0.00190	4.906	0.001
	agvhd:PBMC	6	0.00065	0.00011	0.278	0.943
B_cells	agvhd	1	0.00022	0.00022	0.528	0.473
	PBMC	2	0.00067	0.00034	0.810	0.455
	agvhd:PBMC	2	0.00007	0.00004	0.086	0.918
Leukocyte	agvhd	1	0.00044	0.00044	1.185	0.278
	PBMC	7	0.01074	0.00153	4.148	< .001
	agvhd:PBMC	7	0.00358	0.00051	1.382	0.218
NK_cells	agvhd	1	0.00002	0.00002	0.054	0.816
	PBMC	12	0.02473	0.00206	4.951	< .001
	agvhd:PBMC	12	0.00173	0.00014	0.346	0.979
T_helper	agvhd	1	0.00234	0.00234	5.039	**0.026**
	PBMC	12	0.02587	0.00216	4.650	< .001
	agvhd:PBMC	12	0.00370	0.00031	0.664	0.784
T_suppressor	agvhd	1	0.00253	0.00253	6.717	**0.011**
	PBMC	11	0.01784	0.00162	4.300	< .001
	agvhd:PBMC	10	0.00070	0.00007	0.185	0.997
T_cells	agvhd	1	0.00028	0.00028	0.601	0.439
	PBMC	11	0.01835	0.00167	3.538	< .001
	agvhd:PBMC	11	0.00322	0.00029	0.621	0.809
TCR	agvhd	1	0.00003	0.00003	0.098	0.754
	PBMC	11	0.00953	0.00087	2.909	0.002
	agvhd:PBMC	9	0.00092	0.00010	0.345	0.958

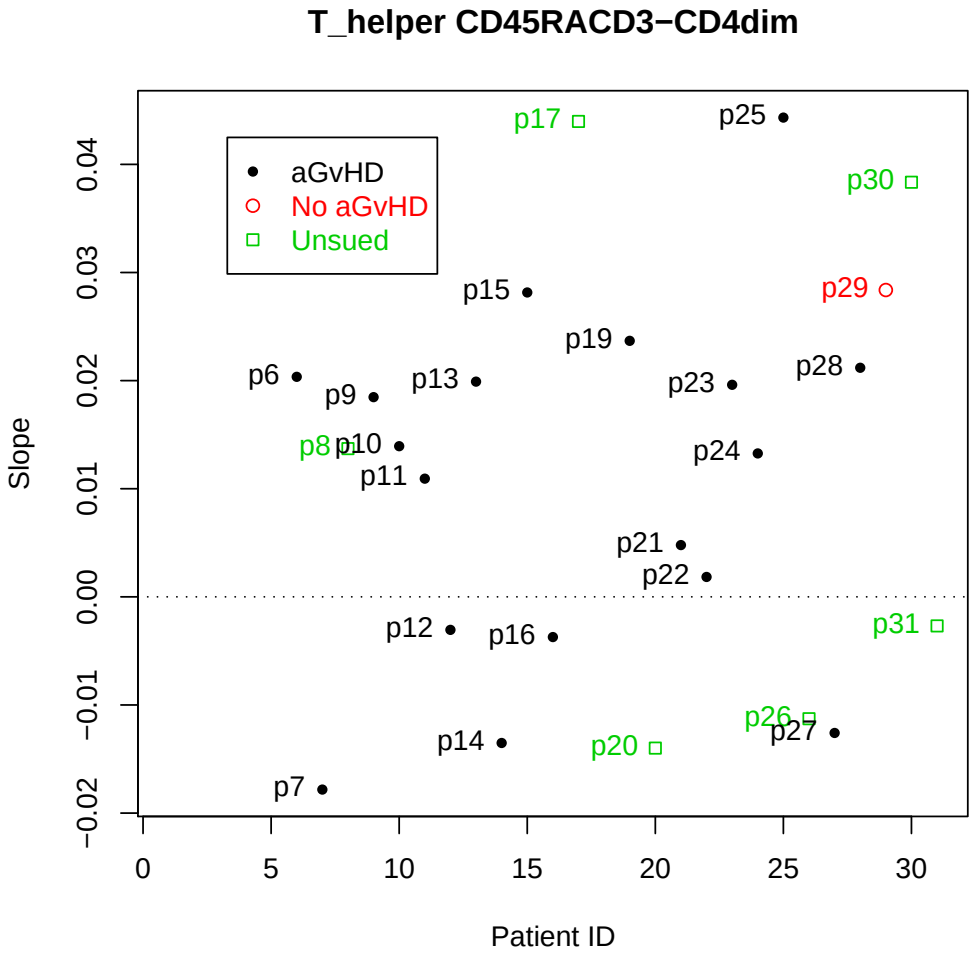

Figure 5.2: Scatter Plot of Slopes for a Subtype of T Helper Cells

Chapter 6

Final Remarks

We analyzed the peripheral blood mononuclear cell data, together with their clinical outcomes. The probability of having aGvHD for patients who received the bone marrow from their siblings is much lower than it for those with non-sibling donors. However, in a long time period, the two groups have no significant difference in death or having chronic GvHD. The conducted PCA tended to select the PBMCs with larger proportion as the important variables. Analysis of "windowlized" data at a fixed time point could not find one PBMC or a subset of PBMCs that is important in the prediction of aGvHD. MANOVA of longitudinal data shows that proportions of T cells over time significantly separate the patients with aGvHD and patients without aGvHD. The ANOVA of slopes of linear regression lines of the proportion data against time tells that the trends of "T_helper" and "T_suppressor" are different between aGvHD patients and non-aGvHD patients in general.

The sample size of the study data is relatively small. Manual gating, missing values and some typos in the data also make analyses of this study data difficult. Currently there are few sample size guidelines referenced in the literature. Some people (Hox,1998; Maas & Hox 2002) suggest that the minimum of units at each level of the analysis is 30. The sample sizes at the PBMC level are about 9 in our study.

We did not take into account the status of disease changing over time in the analysis. More sophistic models, such as a multi-state model, need be applied to analyzing the clinical outcomes with status of disease.

We have used Box-Cox and logistic transformation to transform data. The histograms of the transformed data are given in Figure 6.1 They both led the original data closer to normality.

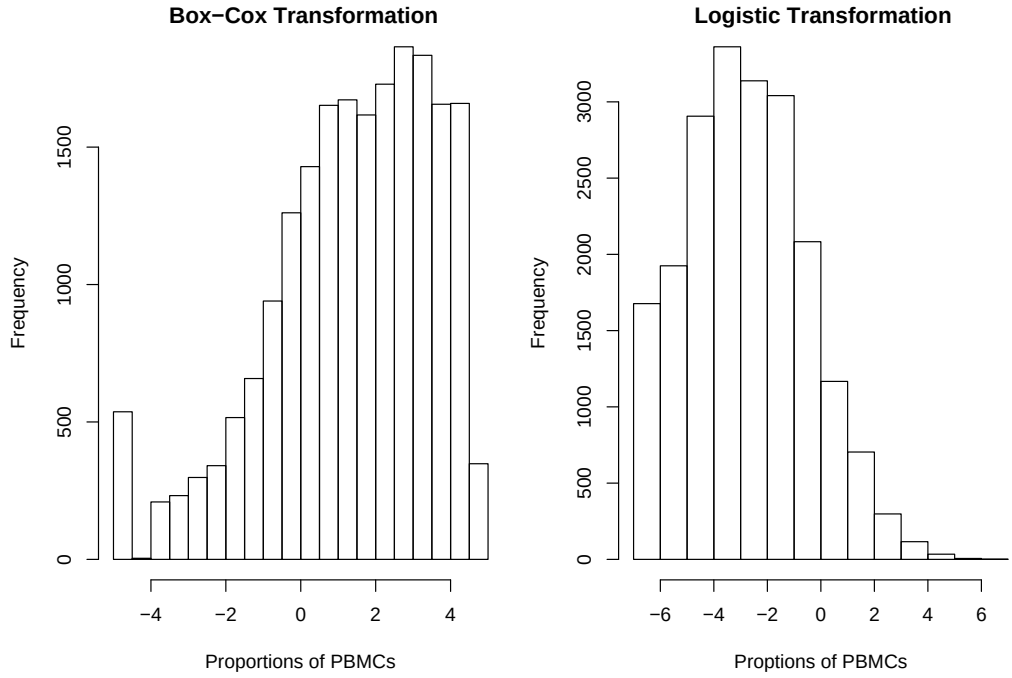

Figure 6.1: Histograms of Transformed Data

Bibliography

Andrew Gelman, J. H. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Bair, E., & Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*.

Berger, G. C. . R. L. (2002). *Statistical Inference*. Murray Hill, New Jersey: The Wadsworth Group.

Cleveland, W. S. (1993). *Visualizing Data*. Murray Hill, New Jersey: AT&T Bell Laboratories.

Faraway, J. J. (2006). *Extending the Linear Model with R*. 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL: Chapman and Hall/CRC.

Hox, J. J. (1998). Multilevel modeling: when and why. *Classification, data analysis, and data highways*.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Hoboken, New Jersey: John Wiley & Sons, INC.

Lee, J. (2006). Prediction of graft-versus-host disease based on supervised temporal analysis on high-throughput flow cytometry data (master thesis). *Medical Genetics, University of British Columbia*.

Lesaffre, D., E Rizopoulos, & Tsonaka, R. (2007). The logistic-transfor for bounded outcome scores. *biostatistics*.

Peter J. Diggle, K.-Y. L. . S. L. Z., Patrick J. Heagerty (2003). *Analysis of Longitudinal Data*. Hoboken, New Jersey: Oxford University Press.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*. New York: John Wiley & Sons, INC.

Shapiro, H. M. (2003). *Practical Flow Cytometry*. Hoboken, New Jersey: John Wiley & Sons, INC.

Yokoyama, T. (1995). Statistical inference on some mixed manova-gmanova models with random effects. *HIROSHIMA MATH*.