Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Meredith N. Scarberry Date

Graphical Display Methods for Exploiting the Dimensionality of Flow Cytometry Data

By

Meredith N. Scarberry

Master of Science

Biostatistics

_____

Vicki Stover Hertzberg, Ph.D.

Advisor

_____

Paul S. Weiss, M.S.

Committee Member

Accepted:

_____

Lisa A. Tedesco, Ph.D.

Dean of the Graduate School

_____

Date

Graphical Display Methods for Exploiting the Dimensionality of Flow Cytometry Data

By

Meredith N. Scarberry

B.A., Boston University, 2006

Advisor: Vicki Stover Hertzberg, Ph.D.

An abstract of

A thesis submitted to the Faculty of the Graduate School of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science

in Biostatistics

2009

Abstract

Graphical Display Methods for Exploiting the Dimensionality of Flow Cytometry Data

By Meredith N. Scarberry

Flow cytometry is an emerging technology that measures characteristics of particles, such as blood cells, as they flow in a stream through a beam of light. This procedure is capable of producing data measurements on many variables simultaneously. In the past, scientists have limited their focus to a small subset of the data recorded by the flow cytometer to ease the tasks of understanding and analyzing the data. I propose the use of radar charts, windrose charts, star charts, and pie charts for visualizing such data. Flow cytometry data from the Protective Immunity Project (NIH NO1-AI-50025, PI: C. Larsen) is presented to exemplify the graphical displays. I discuss the adherence of the proposed graphics to principles of data visualization proposed by experts in the field and assess the effectiveness of each plot for conveying the intended information.

Graphical Display Methods for Exploiting the Dimensionality of Flow Cytometry Data

By

Meredith N. Scarberry

B.A., Boston University, 2006

Advisor: Vicki Stover Hertzberg, Ph.D.

A thesis submitted to the Faculty of the Graduate School of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science

in Biostatistics

2009

Table of Contents

Tables and Figures

Tables

Figures

Introduction

Flow cytometry is increasingly popular in many fields of study, including molecular biology and immunology. This technology is capable of recording measurements on many particle characteristics simultaneously. The dimensional capabilities of flow cytometry are often ignored due to the difficulty in comprehending and analyzing data of such magnitude. The aim of this report is to suggest data display options which exploit the dimensionality of flow cytometry. Such graphics convey patterns within the high-dimensional data that may be difficult to show numerically.

The graphics explored here are applied to data from the Protective Immunity Project, an ensemble of studies exploring immune function following renal transplantation. Though not typically applied to medical datasets, radar charts, windrose charts, star charts, and, in special cases, pie charts are proposed for exploring flow cytometry data. These visual tools enable comparisons between patients, visualizations of patients' change over time, and displays of average characteristics. I examine how well each graphic conveys the intended information and how well each adheres to principles of data visualization proposed by authorities in the field.

Background

Flow Cytometry

Flow cytometry measures multiple physical characteristics of particles, such as cells, as they flow in a fluid stream through a beam of light. The characteristics of each

particle are determined based on how incident laser light is scattered and emits fluorescence.[3]

There are currently a number of instruments available for sorting cells according to the molecules they display on their surface, and the breadth of flow cytometric analysis and sorting is expanding. For example, Mario Roederer of the US National Institutes of Health has simultaneously analyzed seventeen different intracellular and cell-surface markers in a single experiment.[6] Many of the commercial flow cytometers are capable of analyzing at least a dozen fluorescent parameters. Such technology creates large amounts of complex data.

Multidimensional Graphics

In the past, for ease of analysis, investigators have often narrowed their focus to a small number of parameters. Such an approach fails to exploit the high-dimensional capabilities of flow cytometry. One of the difficulties with multidimensional data arises during data visualization. Graphical presentations are useful for conveying the meaning of data when numerical descriptions are inadequate; however, as the number of dimensions increases, displaying the data in two dimensions becomes difficult. In this paper, I intend to convey the utility of multidimensional graphical displays commonly used in other scientific fields for visualizing an immunology dataset produced using flow cytometry.

The graphics explored here are the radar chart, the windrose chart, the star chart, and, in a few cases, the pie chart. These charts are all circular histograms that allow concise visualization, particularly of compositional data.

Radar charts display the relative frequencies of data measures. In a radar chart, each variable has its own axis, and the axes radiate from the center of the chart. These plots are commonly used in quality control, market research, analytical chemistry, and toxicology.[10, 11] Though this graphic is seldom used in medicine, there are numerous possible applications of the radar chart to this field. Saary proposes that radar charts are useful in clinical research: they can depict changes over time in multiple variables for both individuals and groups, multiple-treatment group differences on multiple-outcome measures, and differences between disease conditions on multiple variables.[10] She also posits that radar graphing would be effective for displaying biologic marker composition, as we exemplify in this study.

The windrose chart is a graphic tool typically used by meteorologists to visually display frequencies of wind speeds and directions. Like the radar chart, each axis radiates from the center and represents a particular category. In meteorology, the axis directions correspond to compass directions. For displaying data not associated with compass directions, the axes' orientations are arbitrary.

The star chart is very similar to the radar chart in its creation and appearance. It consists of lines or connected slices radiating from the center. Each slice represents a category, and magnitudes correspond to the length of each slice. Star charts are often thought to be simply a variation of the radar chart, but this variation of appearance may convey different aspects of the data.

Finally, in the commonly known pie chart, slices of the circle represent each category. The size of the slice corresponds to the frequency measure. This chart will be used when the others are made ineffective by the structure of the data being modeled.

Principles of Data Visualization

In creating each display, I have attempted to adhere to the guidelines proposed by experts in the fields of statistics and data visualization. I briefly discuss a few of these principles here.

Edward R. Tufte is possibly the most prominent authority in the arena of data visualization. In his 1983 book, The Visual Display of Quantitative Information, Tufte explains that the goal of creating statistical graphics is to communicate complex ideas with clarity, precision, and efficiency. Tufte posits the idea of graphical excellence. He says, "graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space." Additionally, as with all statistics, a quality graph should tell the truth about the data; it should not be deceptive. The size of the effect shown in the graphic should mirror the size of the effect in the data.

A few of Tufte's suggestions, for example, the minimization of ink used in a graphic, are not applied here. As the following plots were generated using SAS® software, implementing such changes would be difficult and diminish the practicality of the graphic.

In 1990, Tufte published Envisioning Information in which he broadens his focus from statistical graphics to all types of informational displays. He critiques previously published informational graphics and proposes effective design strategies. Tufte's previous work on statistical displays is more directly applicable to this study; however, his general advice concerning image layering, image separation, and the use of color

addressed in Envisioning Information is acknowledged in the creation of the graphics produced in this report.

William S. Cleveland also contributed greatly to the current literature and practices of data visualization. In 1985, he outlined new and lesser known graphical methods and principles for data communication in his book The Elements of Graphing Data. He discussed principles of graph construction, graphical methods, and graphical perception. Many of Cleveland's principles are similar to those of Tufte and include such rules as avoiding clutter in the data region and making overlapping datasets distinguishable. I will later discuss how well the figures used in this analysis adhere to some of the guidelines put forth by Tufte and Cleveland.

The Protective Immunity Project

The data used in this analysis comes from the Protective Immunity Project (PIP) (NIH NO1-AI-50025, PI: C. Larsen), an ensemble of studies examining immune function following renal transplantation.[7] This is an ongoing study at Emory University which began in 2005. PIP consists of three complementary studies. The first study aims to characterize the impact of immunosuppressive regimens on protective immunity over time. Sixty subjects were recruited from patients ages 18 to 59 who had undergone renal transplantation at Emory University. Additionally, a control group was recruited consisting of twenty age-, sex-, and race-matched healthy volunteers. Subjects were followed for two years. Blood samples were drawn at baseline, 3, 6, 9, 12, 18, and 24 months. The data used for the following analysis is a subset of the data gathered for this first of the PIP studies.

The variables used in this analysis were identified using flow cytometry. Whole blood was passed through the flow cytometer to determine blood composition (Fig. 1). Lymphocytes were isolated by visualization. The percentage of lymphocytes displaying the CD3+ surface marker, known as T lymphocytes, was recorded. The T lymphocytes were then broken into four categories: CD4+CD8- (T helper cells), CD4-CD8+ (cytotoxic T cells), CD4-CD8-, and CD4+CD8+. Percentages for each category were recorded. The T helper cells were further broken into four categories: CCR7+CD45RA- (central memory), CCR7+CD45RA+ (naïve), CCR7-CD45RA- (effector memory), and CCR7-CD45RA+. These percentages were recorded. Finally, the cytotoxic T cells were further broken into 4 categories: CCR7+CD45RA- (central memory), CCR7+CD56RA+ (naïve) CCR7-CD45RA- (effector memory), and CCR7-CD45RA+ (effector memory RA). These percentages were also recorded.

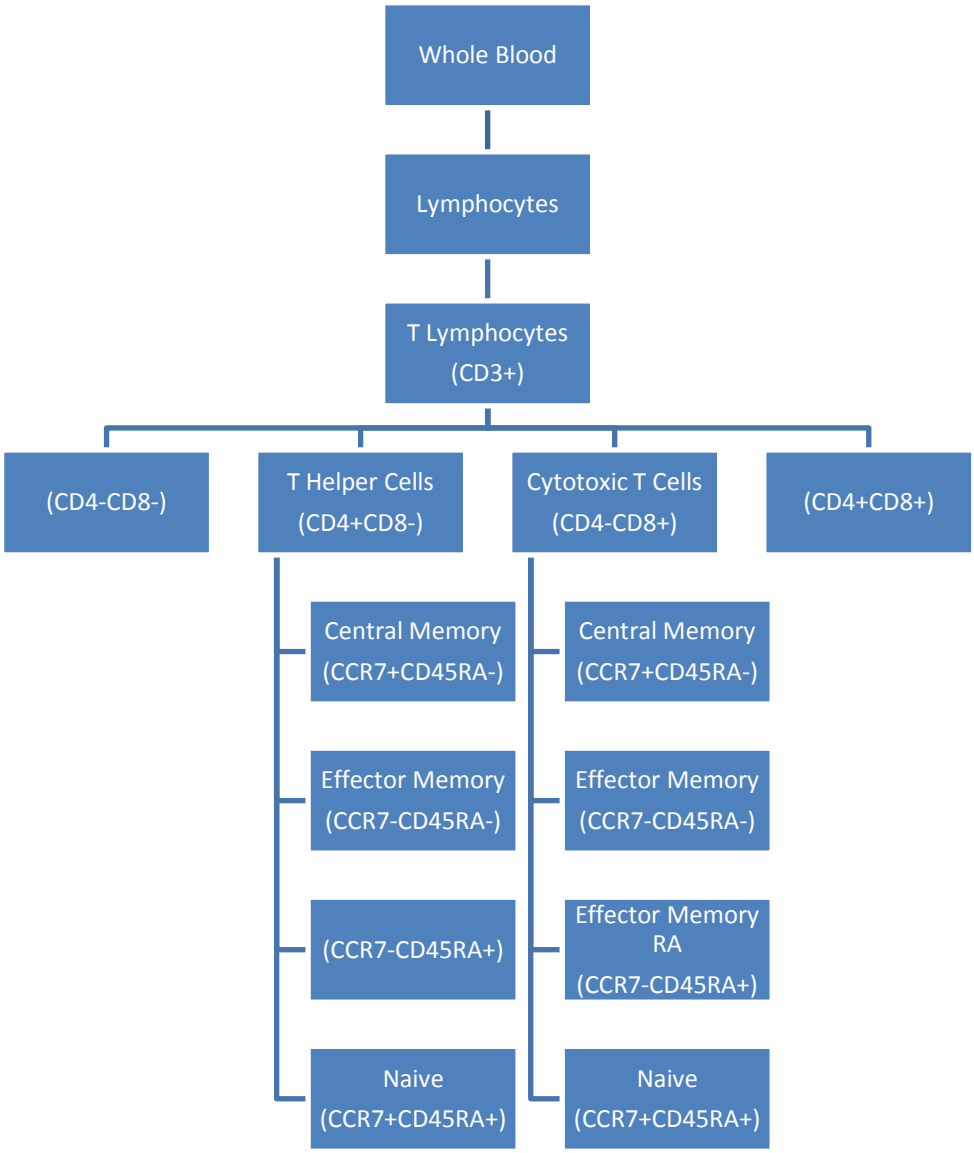


Figure 1. Hierarchy of blood cell types in PIP dataset.
Note: There are more blood cell types than are displayed here. Only cells that are isolated in the PIP study by flow cytometry are included in this figure.

Blood Composition

I provide below a brief explanation of blood composition, specifically the variables contained in the PIP dataset.

Human blood is composed of red blood cells, white blood cells, platelets, and other substances suspended in plasma.[9] White blood cells are immune cells and help the body fight infection and disease. A lymphocyte is a type of white blood cell. There are two types of lymphocytes: B cells and T cells. T lymphocytes (T cells) help control immune responses and fight foreign or diseased cells. Cytotoxic T cells, identified by the CD3+CD4-CD8+ marker combination, kill foreign or infected cells. T helper cells, identified by the CD3+CD4+CD8- marker combination, recognize foreign cells and secrete substances that activate killer T cells.[8] Memory T cells can recognize foreign or infected cells that were encountered previously, during infection or vaccination.[15] They exist among both cytotoxic and helper T cells. Central memory T cells may represent memory stem cells. Both effector memory subtypes, effector memory and effector memory RA, express genes for molecules essential to the functioning of the T cells. Naïve T cells are those that have not encountered a foreign cell.[2] They are able to respond to novel foreign cells, and, therefore, fight new viruses or diseases.

Compositional Data

The nature of this dataset requires statistical methodology tailored to compositional data. In compositional data, the variables take on values that are proportions of some whole. Therefore, the sum of the variables is constrained to be a constant. In this case, that constant is one hundred percent. This characteristic of the data causes statistical methods designed for unconstrained data to be inappropriate. In this report logratio analysis, a statistical methodology for compositional data, is used.

This method involves a logarithm of ratios transformation, as logratios are mathematically more tractable than ratios.[1]

Application

Methods

Analyses were conducted for seven different subsets of the data: (1) all eight memory categories (the four types of T helper cells and the four types of cytotoxic T cells), (2) the four categories of T helper cells, (3) the four categories of cytotoxic T cells, (4) the four types of T lymphocytes, (5) the two central memory T cells, (6) the two naïve T cells, and (7) the two effector memory T cells (refer back to Fig.1).

Subset 1: All memory categories

To make comparisons among all eight memory categories, new component values were calculated. That is, for each cell type, a new percentage was calculated out of the population consisting of all T helper cells and all cytotoxic T cells. This was done for each time point (0, 3, 6, 9, 12, 18, and 24 months). Next, compositional means were calculated for each time point using logratio analysis. As the data is compositional, arithmetic means were not appropriate for summarizing the blood composition variables.

The method for calculating the means is as follows. The data was first transformed into logratios: the ratio of each variable was taken with respect to one arbitrary reference variable, and the logarithms of these ratios were calculated. The arithmetic means for each time point were then calculated. Finally, the means were transformed back to their original scale by exponentiation, division by the sum of the

exponentials, and multiplication by one hundred. The value of the exponential of the reference category is defined to be one. Details of this calculation can be found in Aitchison's Concise Guide to Compostional Data Analysis.

The data could now be displayed graphically. SAS® version 9.2 was used to generate all graphics.[11] First, radar plots were created using the GRADAR procedure. Three types of radar plots were created in order to show different aspects of the data. The first plots were grouped by month, and plots for each patient were tiled horizontally. This type of chart allows for comparisons between patients at each time point. Plots were tiled rather than overlaid, as the number of patients caused overlaid lines to be indistinguishable. The second type of radar diagram visualizes patient progression over time. One plot was created for each patient, and plots for each month were overlaid. Finally, mean plots for each month were overlaid to create a single radar plot which summarizes the data.

Next, two types of windrose plots were generated by the GRADAR procedure with the *windrose* option. SAS® does not allow windrose charts to be tiled or overlaid. Therefore, a single display was created for each patient at each timepoint. These could then be grouped by either patient or month to view patient progression or comparisons between patients, respectively. Mean windrose plots were also created for each time point.

The last plots created for this subset of the data were star charts. The *star* statement of the GCHART procedure generated these displays. Like windrose charts, star charts cannot be tiled or overlaid. Plots were created for each patient at each time point, and a mean plot was created for each month.

Subset 2: T helper cell categories

In the original data, the values for each of the T helper cell categories - central memory, effector memory, naïve, and the unnamed category - were expressed as percentages of all T helper cells; therefore, these values did not need to be transformed. Logratio analysis (as previously explained) was used with these original values to calculate the compositional means.

Radar and star charts were used to visualize this data subset. Windrose plots could not be used here, as they require a minimum of eight components. Three types of radar charts and two types of star charts were made as they were for the first subset of the data: radar charts grouped by month with the patient plots horizontally tiled, radar charts for each patient with time point measurements overlaid, a radar chart with the mean plots for each month overlaid, star charts for each patient for each time point, and mean star charts for each time point.

Subset 3: Cytotoxic T cell categories

Graphics displaying the composition of cytotoxic T cells (central memory T cells, effector memory T cells, effector memory RA T cells, and naïve T cells) were created just as they were for T helper cells (subset 2).

Subset 4: T lymphocyte categories

Like the data for subsets 2 and 3, the original data for the four lymphocyte categories (T helper cells, cytotoxic T cells, double negative CD4-CD8-, and double positive CD4+CD8+) were expressed as percentages of all T lymphocytes. Therefore, means were calculated and graphics were created for this dataset as they were for subsets 2 and 3.

Subsets 5, 6, 7: Central memory T cells, Naïve T cells, Effector memory T cells

To create the graphical displays for these subsets, the transformed data values from subset 1 (all memory categories) were used. In this subset, the magnitudes of each type of T helper cell were comparable to their cytotoxic T cell counterparts. Likewise, the means calculated for subset 1 were comparable.

For each subset, only two categories were compared. Neither the radar chart, nor the windrose chart, nor the star chart is capable of effectively plotting two categories. The windrose chart requires at least eight categories. Both the radar chart and the star chart attempt to display the two categories as single spokes in opposite directions, and all that is apparent is an axis line.

As all previous charts are inadequate, simple pie charts were created using the *pie* statement of the GCHART procedure. Plots were created for each patient at each time point, and mean plots were created for each time point. These plots allow the viewer to see whether each memory category is more heavily represented in the T helper cells or cytotoxic T cells.

Results

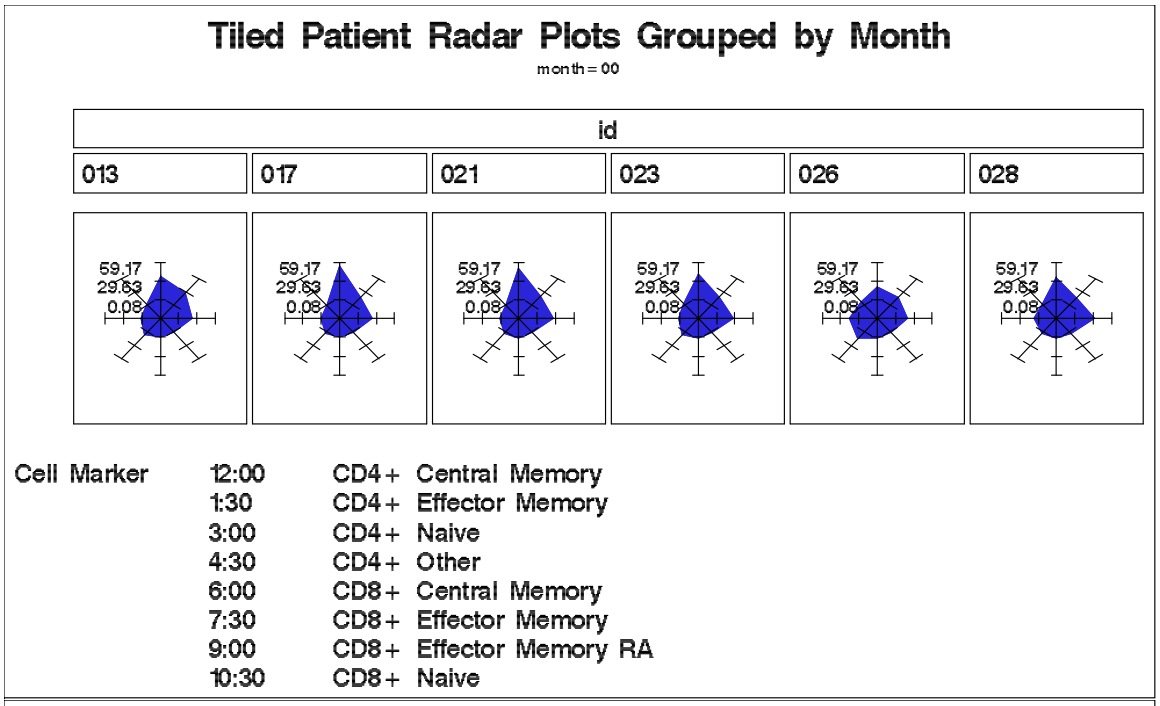
Eighteen of the sixty renal transplant patients were used in this analysis. These patients were chosen for their relatively complete follow-up information. Three of the patients had no missing data. Four patients had observations for five of the six time points. Ten patients had data for four time points, and one patient had observations for three time points. All patients provided blood samples at baseline. Fifteen returned for follow-up at 3 months, thirteen at 6 months, fifteen at 9 months, sixteen at 12 months,

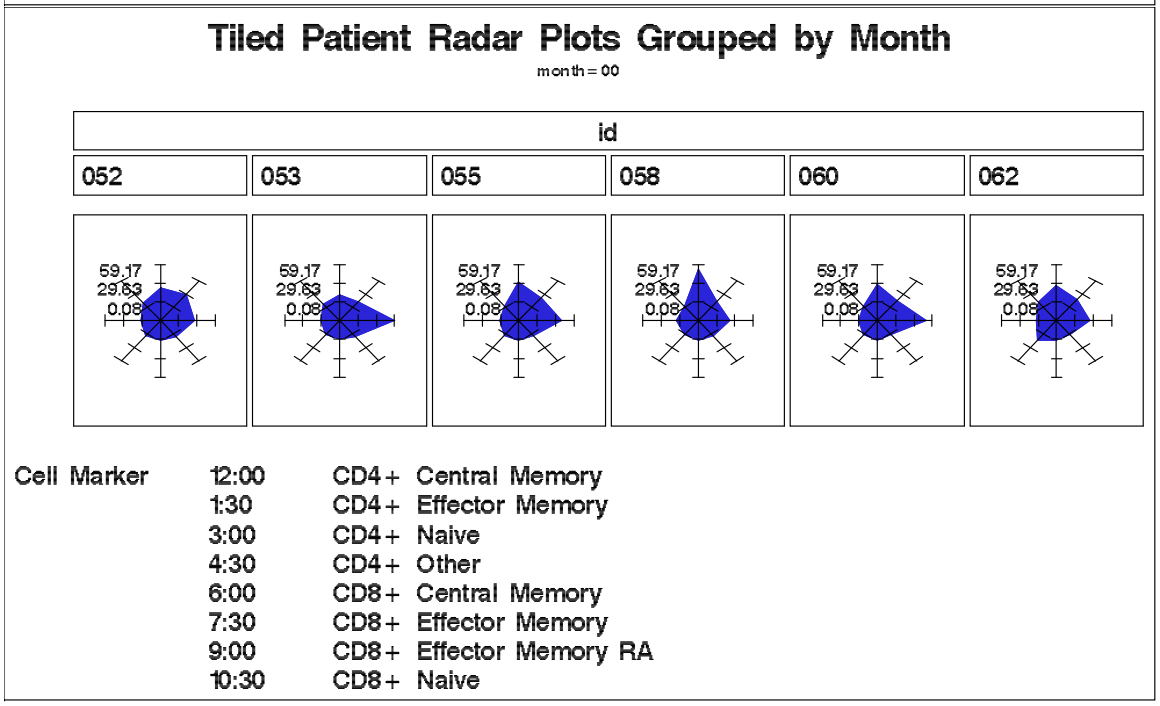
three at 18 months, and only one at 24 months. Though the data is compositional and theoretically should sum to one hundred, measurement error caused some deviation from this total. Therefore, there is some error in the original data and all subsets that were created.

The effector memory T helper cell component was used as the reference variable when calculating the means for all data subsets except for subset 3, the categories of cytotoxic T cells. For this subset, the effector memory RA cytotoxic T cell was used, as the effector memory T helper cell was not a member of this subset.

Subset 1: All memory categories

I first created patient radar plots for the dataset with all eight memory cell categories as variables. The first graphic arranges patient plots side-by-side for comparison, and groups the plots by time point. For example, Figure 2 displays all patients' blood composition at baseline. All patients could not be displayed in a single row, for such plots were too small to read. I judged six plots per row to be the most effective. A legend is used to prevent the clutter of axis labels, and each axis is identified by its position on a clock. Note that on radar plots the vertex is not necessarily zero. By default, the first tick mark is the lowest observed value, the last is the highest, and the middle tick mark is halfway between these two values.


Tiled Patient Radar Plots Grouped by Month
month = 00
id
013
017
021
023
026
028
59.17
29.63
0.08
Cell Marker
12:00 CD4+ Central Memory
1:30 CD4+ Effector Memory
3:00 CD4+ Naive
4:30 CD4+ Other
6:00 CD8+ Central Memory
7:30 CD8+ Effector Memory
9:00 CD8+ Effector Memory RA
10:30 CD8+ Naive


Tiled Patient Radar Plots Grouped by Month
month = 00
id
052
053
055
058
060
062
59.17
29.63
0.08
Cell Marker
12:00 CD4+ Central Memory
1:30 CD4+ Effector Memory
3:00 CD4+ Naive
4:30 CD4+ Other
6:00 CD8+ Central Memory
7:30 CD8+ Effector Memory
9:00 CD8+ Effector Memory RA
10:30 CD8+ Naive

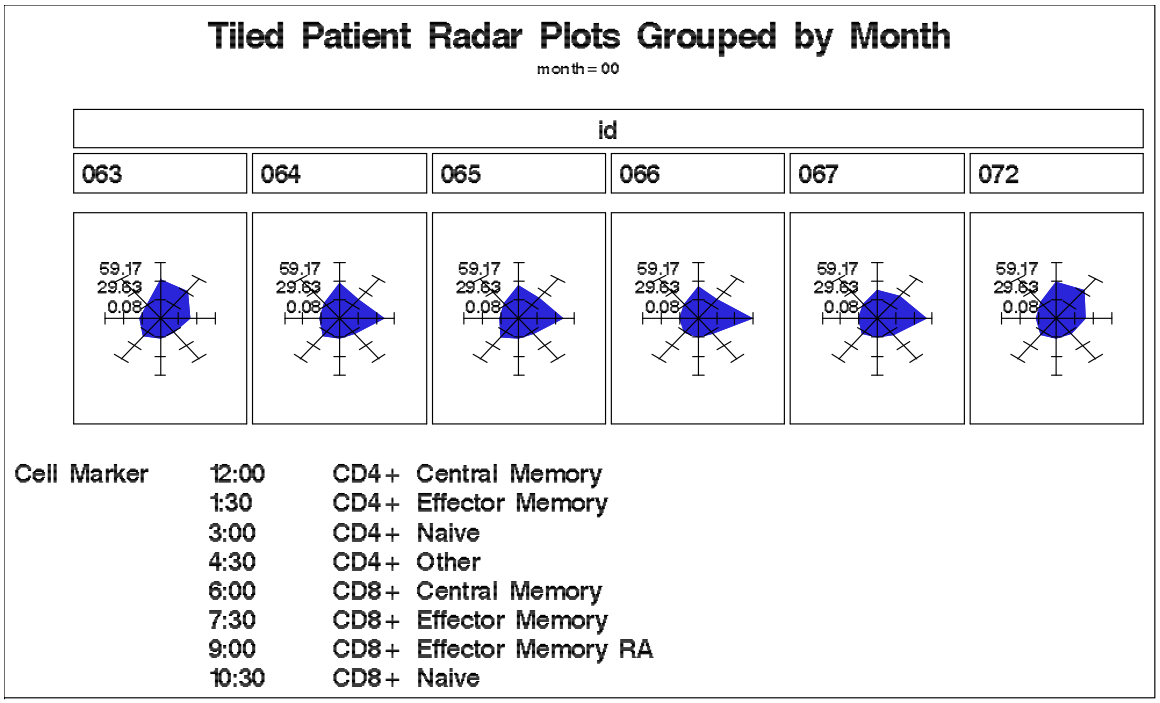


Figure 2: Tiled patient radar charts for baseline subset 1 data.
Note: CD4+ denotes T helper cells and CD8+ denotes cytotoxic T cells.

The next type of radar plot is useful for viewing changes over time. In this type of plot, one chart is created for each patient. Observations for each time point are overlaid on the same axes. This allows easy comparison between time points and conveys how the patient's blood cell counts changed at intervals following transplantation. Figure 3 displays the radar plot for patient 13 as an example. Note that only a few examples of the graphics produced are included in this report, as the graphical tool is of interest here, not the data itself.

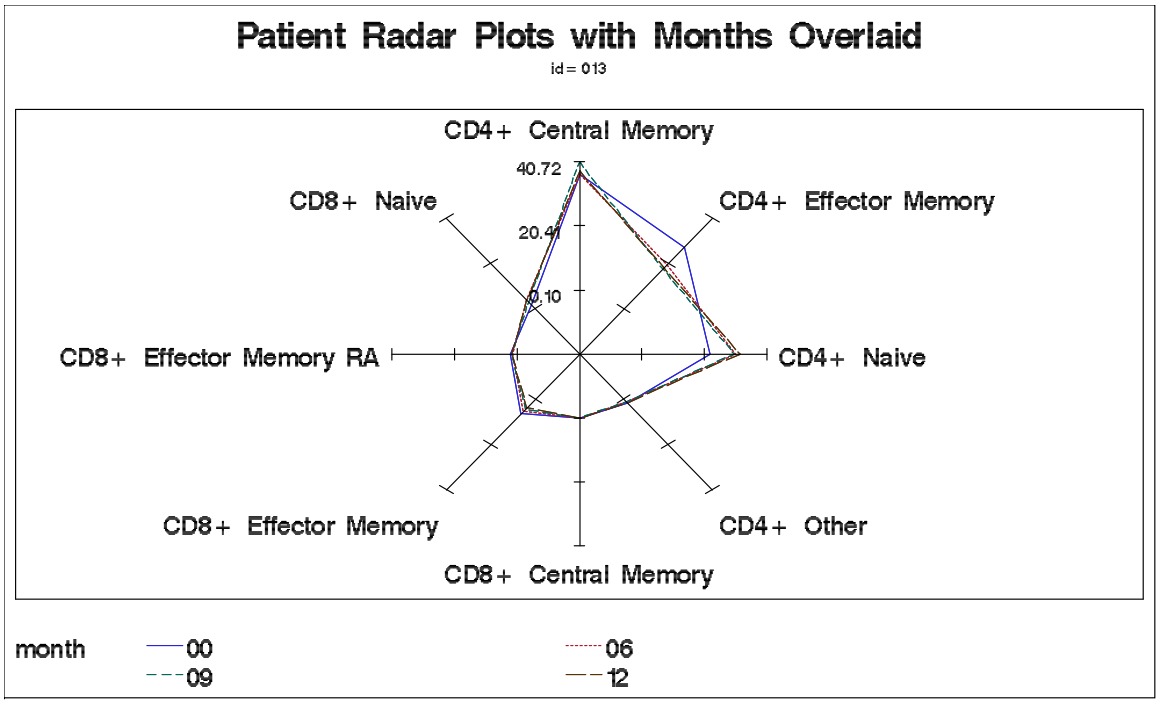
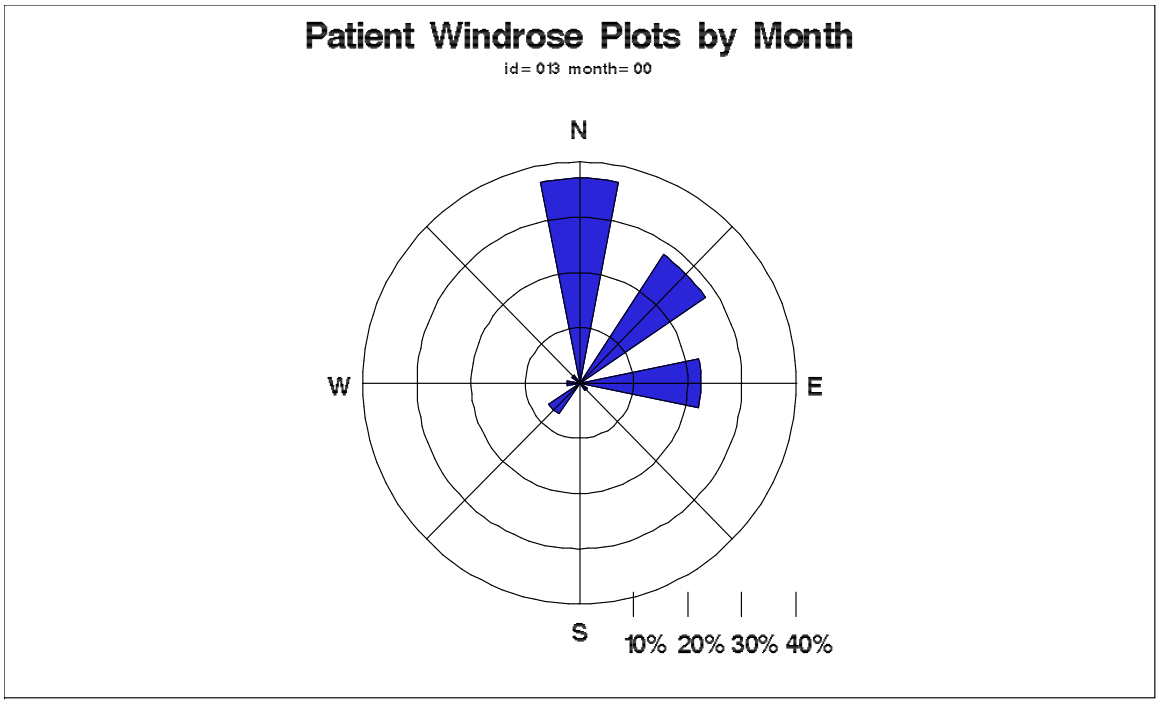


Figure 3: Radar chart for patient 13.
Measurements from each available follow-up blood sample are overlaid.

Next, one type of windrose chart was produced that displays the same information as the preceding two types of radar charts. A single windrose chart was generated for each patient for each time point. Figure 4 gives an example of such a plot. Groups of these charts could be viewed together to convey the desired information. All patients at a single time point could be viewed for comparison (similar to the tiled radar charts), or all time points for a single patient could be viewed to observe patient progression (similar to the overlaid radar charts).

Note that on windrose charts the vertex represents zero, and concentric circles mark proportionally spaced values on the axes. Also, as windrose charts are typically used for displaying wind direction, SAS® labels axes with compass directions. As such, a legend must be provided to relate each axis with a data category.



- N CD4+ Central Memory
- NE CD4+ Effector Memory
- E CD4+ Naïve
- SE CD4+ Other
- S CD8+ Central Memory
- SW CD8+ Effector Memory
- W CD8+ Effector Memory RA
- NW CD8+ Naive

Figure 4: Windrose chart for patient 13 at baseline.

The final patient plot is the star chart. Like windrose charts, star charts cannot be tiled or overlaid in SAS®. Therefore, a single plot for each patient at each time point must be created. Figure 5 shows the star chart that displays the same information as the windrose chart of Figure 4. Note, however, that the categories are not displayed in the same directions as in the windrose chart. The star chart conveniently provides both the category labels and values next to each slice on the graphic.

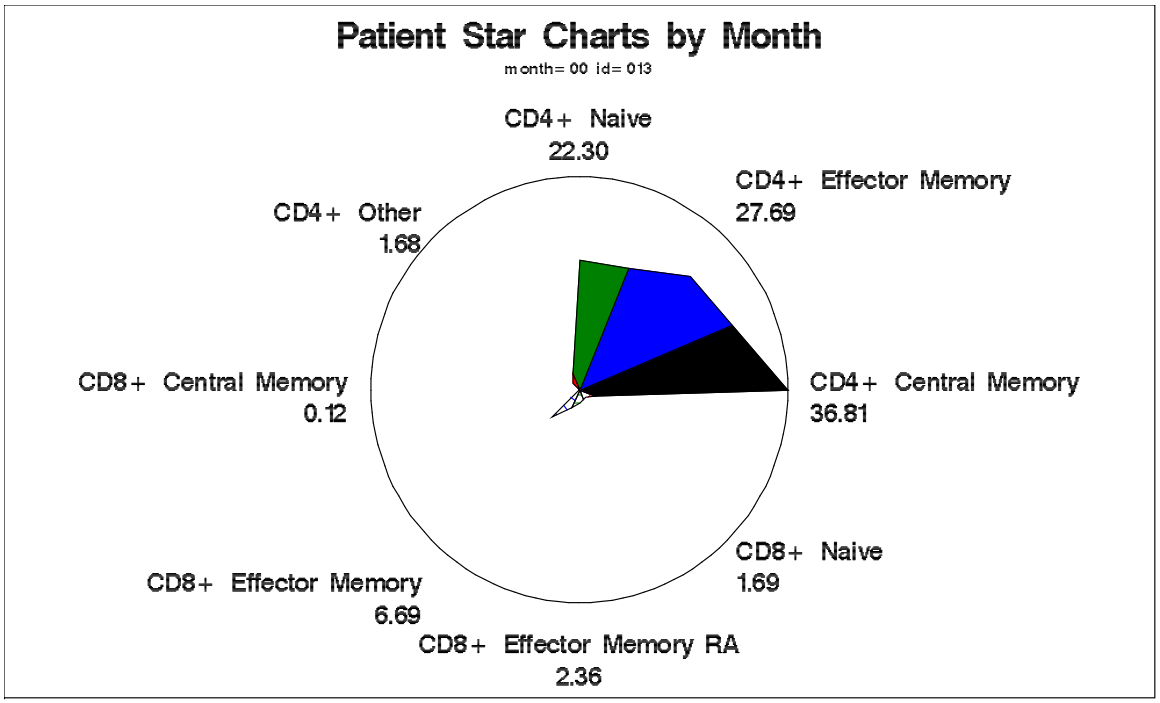


Figure 5: Star chart for patient 13 at baseline.

Finally, graphics were produced which summarize this subset of the data. One type of each chart – radar, windrose, and star – was created to display the mean component values. The compositional means for subset 1 are reported in Table 1. Generally, the largest percentages of cells are central memory T helper cells or naïve T helper cells. For all time points, central memory cytotoxic T cells have the lowest percentage.

Table 1: Compositional means for subset 1 data

	Cytotoxic T cells				**T Helper Cells**			
Month	Central Memory	Naïve	Effector Memory	Effector Memory RA	Central Memory	Naïve	Effector Memory	Other
0	0.391	3.048	4.844	1.885	31.924	33.930	21.082	2.895
3	0.494	4.928	7.433	2.750	29.356	24.124	27.746	3.171
6	0.434	3.238	5.446	3.267	35.403	25.011	24.725	2.476
9	0.369	3.872	7.126	3.559	31.121	25.674	25.239	3.041
12	0.342	3.402	6.983	4.467	29.438	22.138	29.828	3.402
18	0.325	2.502	3.863	2.086	39.891	29.401	20.610	1.322
24	0.330	1.697	1.001	3.128	38.476	41.385	11.074	2.909

First, a single radar chart depicts all values from Table 1 (Fig. 6). For windrose and star charts, separate plots must be created for each time point (Fig. 7 and Fig. 8, respectively). To convey changes in the mean composition over time, all seven time point plots must be provided.

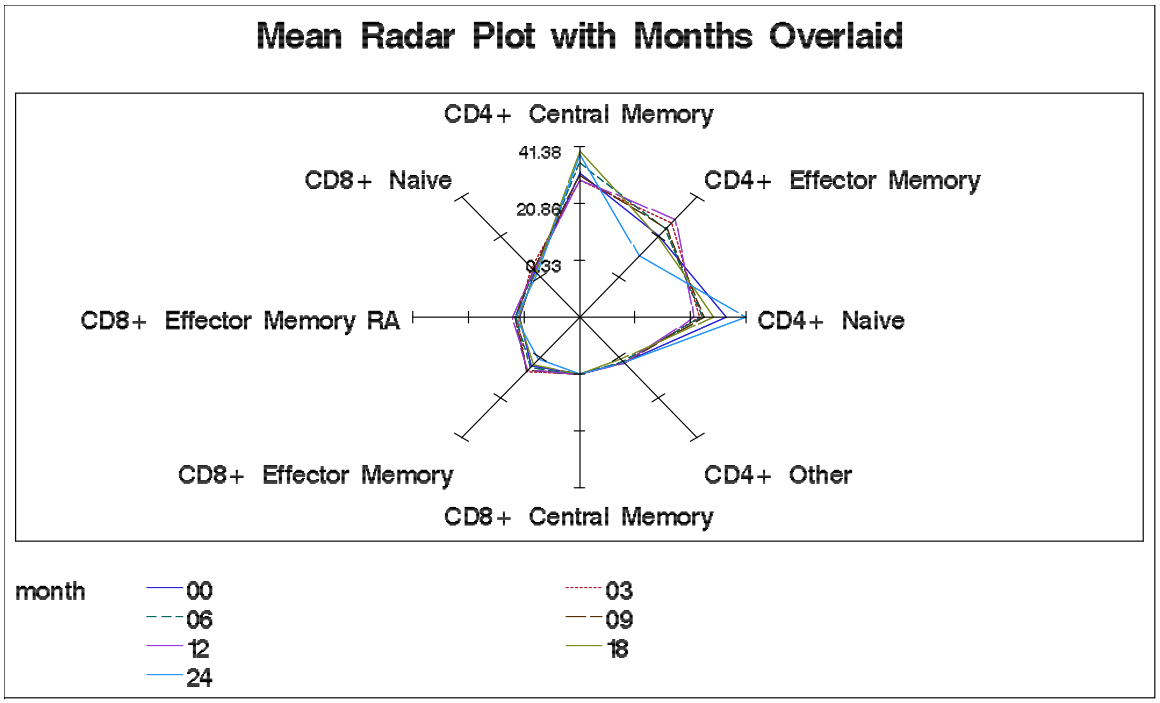
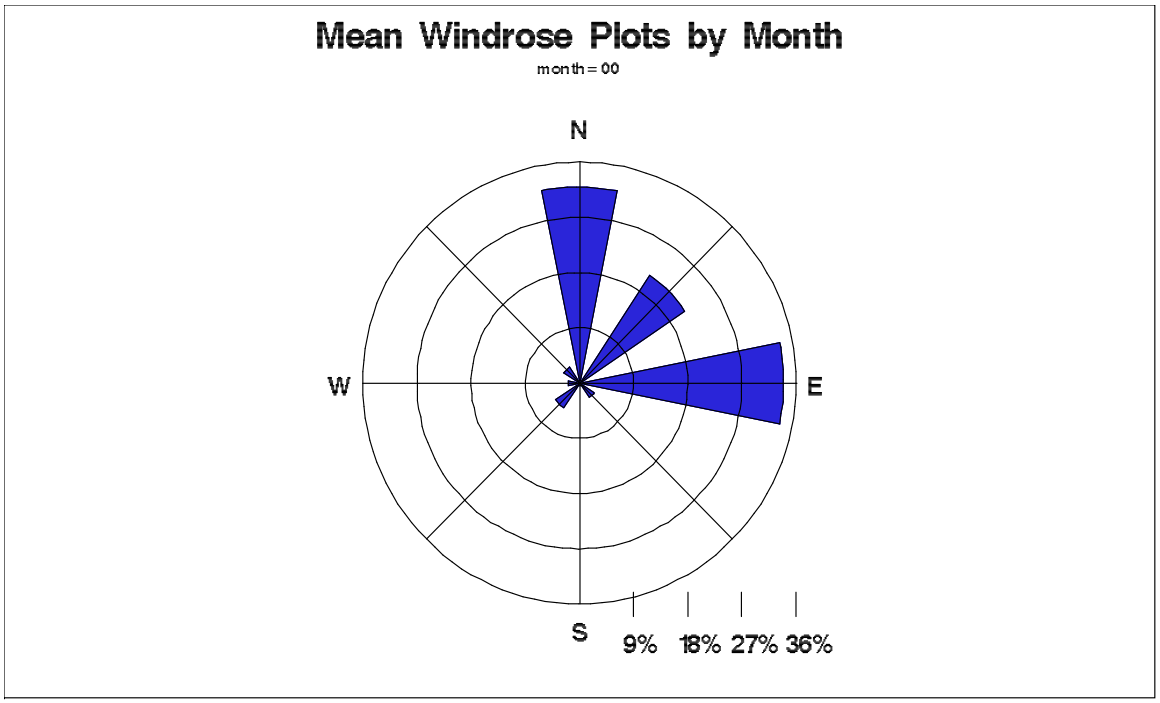


Figure 6: Radar chart of compositional means with each time point overlaid.
Note: Month 24 data was available for only one patient.



- N CD4+ Central Memory
- NE CD4+ Effector Memory
- E CD4+ Naïve
- SE CD4+ Other
- S CD8+ Central Memory
- SW CD8+ Effector Memory
- W CD8+ Effector Memory RA
- NW CD8+ Naive

Figure 7: Windrose chart of component means at baseline.

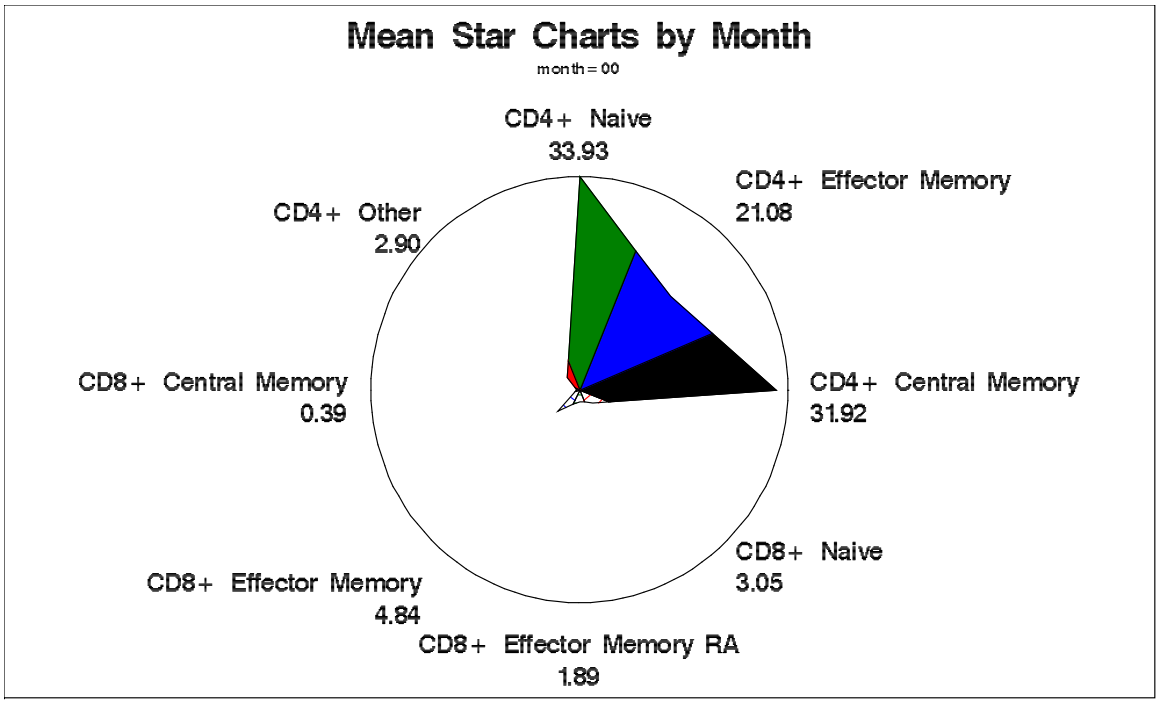


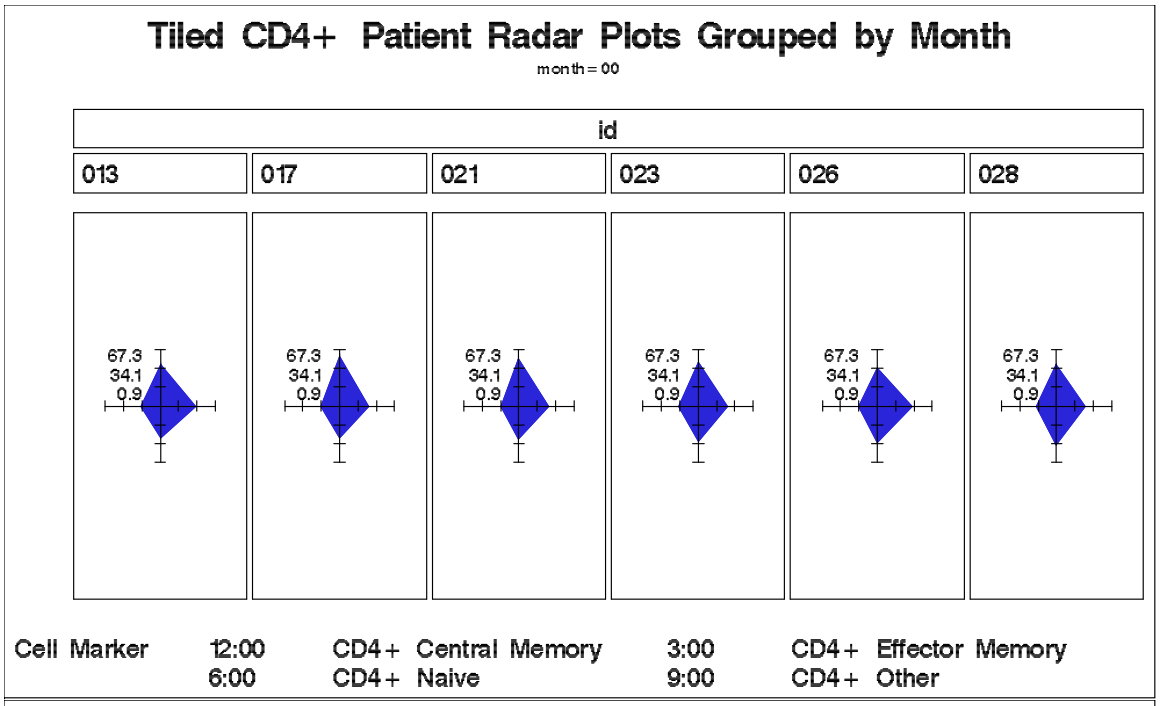
Figure 8: Star chart of component means at baseline.

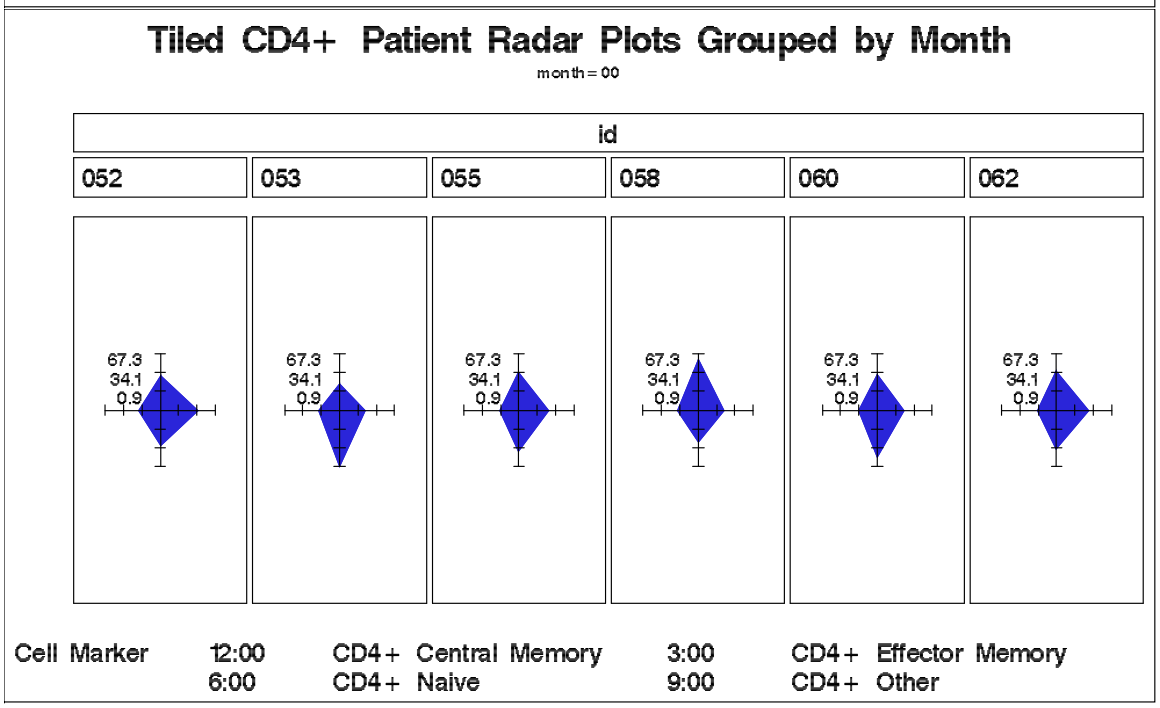
Subsets 2, 3, and 4: T helper cell, cytotoxic T cell, and T lymphocyte categories

Similar plots were made for subsets 2, 3, and 4. Each of these subsets consisted of four variables; therefore, the charts were all constructed the same way but had different data values. As the data is not of interest here, I present example charts for the T helper cell dataset only.

The first graphic, similar to Figure 2, enables comparisons between patients at a given time point (Fig. 9). Separate plots for each patient are placed side-by-side, and the plots are grouped by month. The second type of radar chart for these datasets is similar to that of Figure 3 (Fig. 10). One plot was generated for each patient. The flow cytometry measurements for each time point are overlaid on the same axes. Such a plot reveals patient progression over the course of the study. The final type of individual patient graphic for these data subsets is the star chart (Fig. 11). A star chart was created

for each patient at each time point. To reveal the data patterns shown in the previous two radar plots, the star charts may be grouped by either time point or patient. Once again, windrose plots are not generated for these smaller datasets, as SAS® requires at least eight variables for this type of graphic.


Tiled CD4+ Patient Radar Plots Grouped by Month
month = 00
id
013
017
021
023
026
028
67.3
34.1
0.9
Cell Marker
12:00 CD4+ Central Memory
3:00 CD4+ Effector Memory
6:00 CD4+ Naive
9:00 CD4+ Other


Tiled CD4+ Patient Radar Plots Grouped by Month
month = 00
id
052
053
055
058
060
062
67.3
34.1
0.9
Cell Marker
12:00 CD4+ Central Memory
3:00 CD4+ Effector Memory
6:00 CD4+ Naive
9:00 CD4+ Other

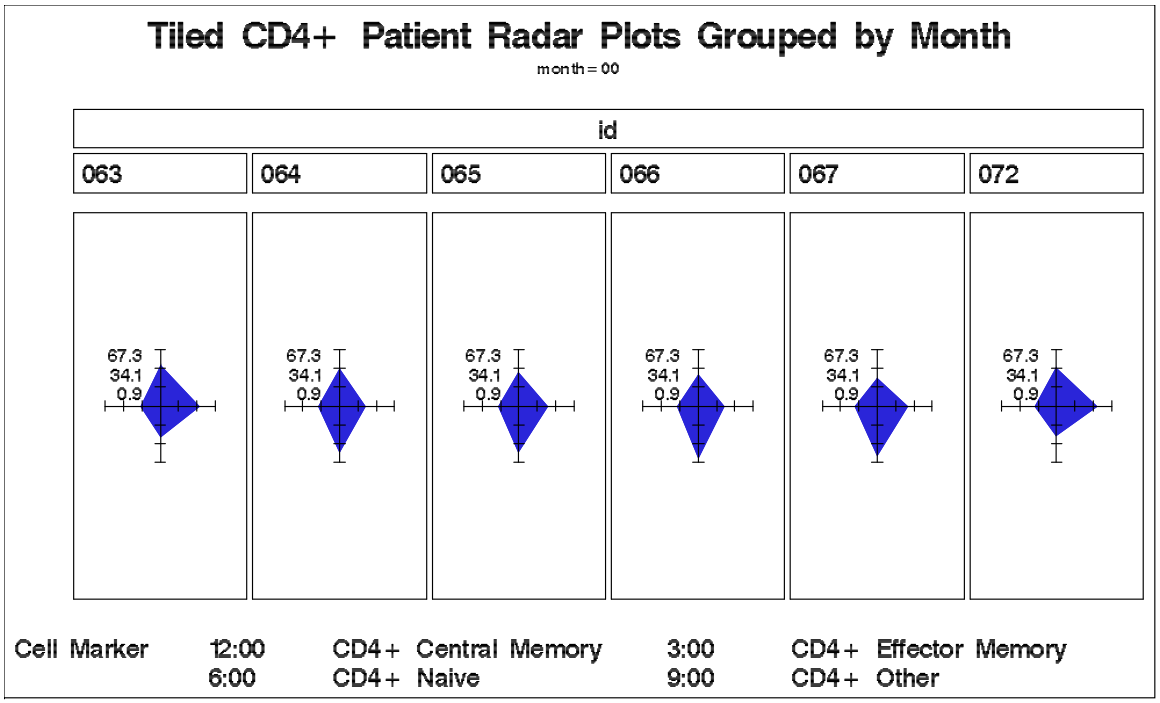


Figure 9: Tiled patient radar charts for baseline data showing composition of T helper cells.

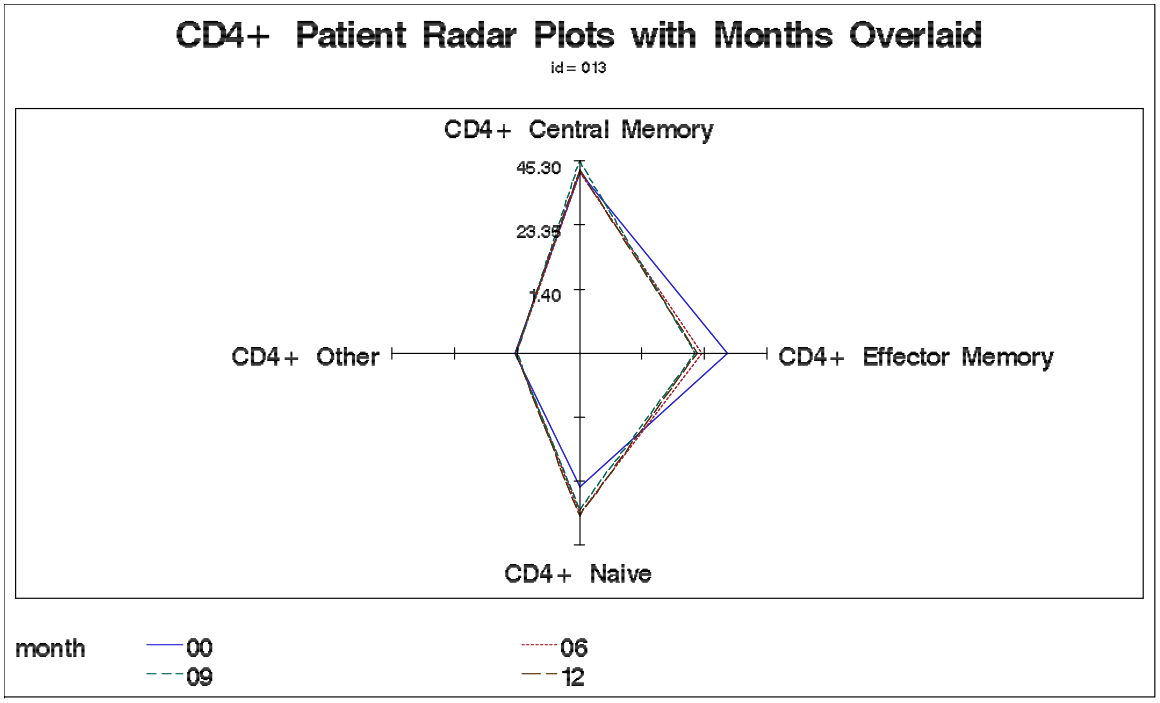


Figure 10: Radar chart of T helper cell composition for patient 13 with plots for each time point overlaid.

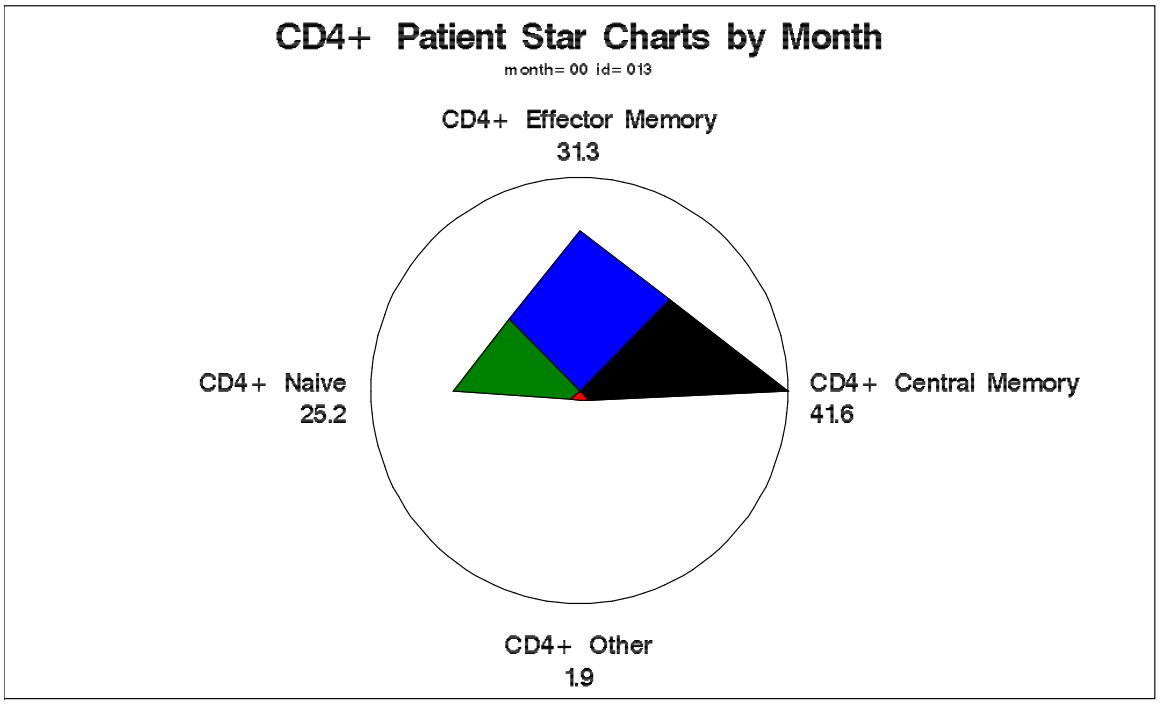


Figure 11: Star chart for patient 13 at baseline showing composition of T helper cells.

Next, plots of component means were created for these three subsets: T helper cell categories, cytotoxic T cell categories, and T lymphocyte categories. One radar plot was generated to summarize each data subset (Fig. 12). These charts present all compositional means by overlaying the plots for each time point. This image depicts the progress of the "average" patient over the course of follow-up. Finally, star charts of the compositional means were produced for each time point. As before, star charts from each time point must be viewed together to visualize change over time. A single plot, however, is effective in communicating a profile of an "average" patient at a given time.

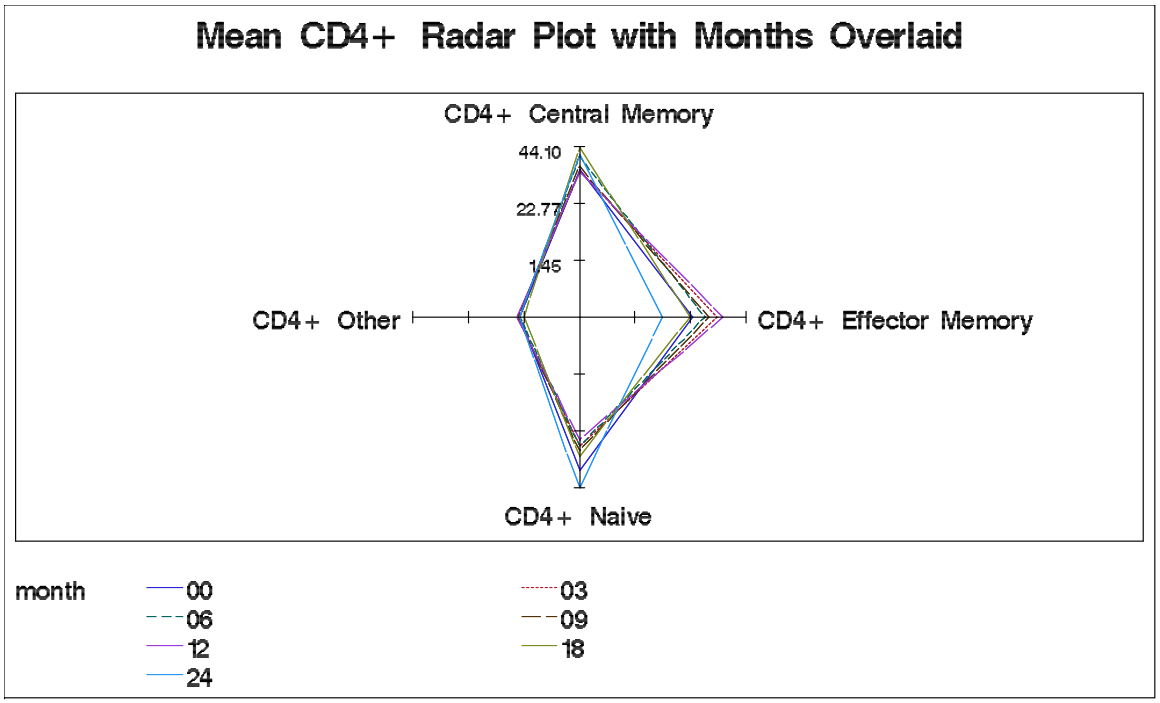


Figure 12: Radar chart of mean T helper cell components with data from each time point overlaid.

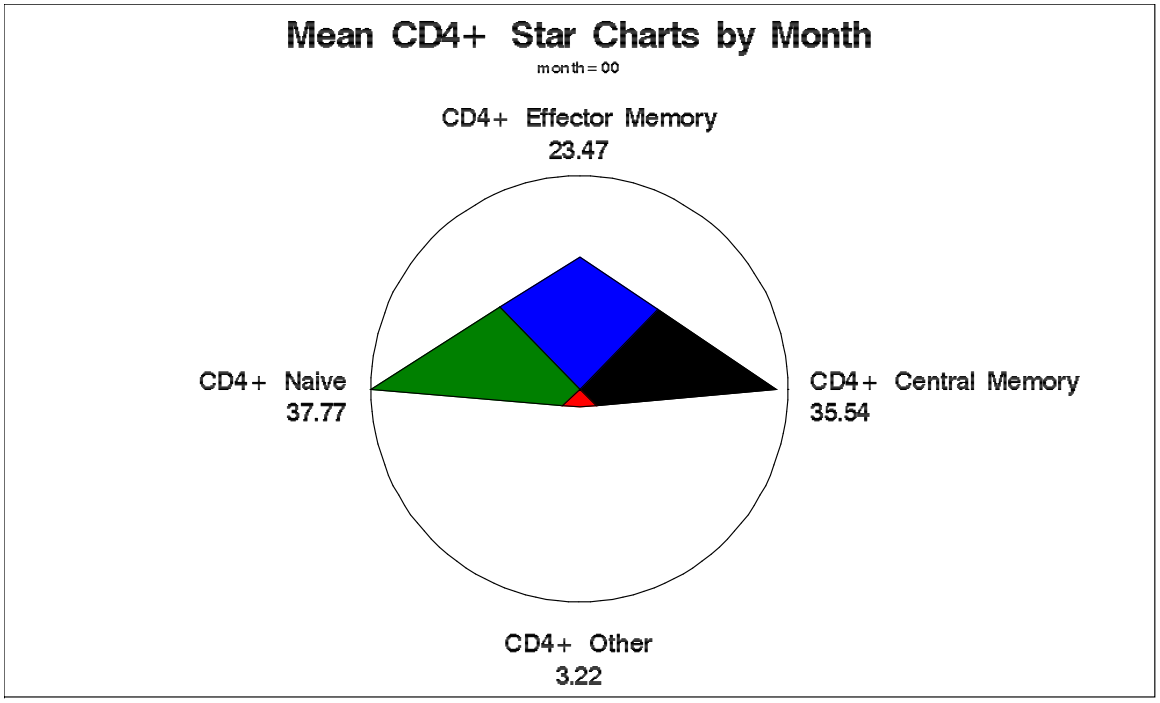


Figure 13: Star chart of mean T helper cell components at baseline.

Subsets 5, 6, and 7: Central memory T cells, Naïve T cells, and Effector memory T cells

Data subsets 5, 6, and 7 are very similar. Each dataset contains patient id, time point, and two T cell categories. The plots created for these datasets visualize the relative frequencies of helper T cells and cytotoxic T cells for each type of memory cell. Only central memory T cells, naïve T cells, and effector memory T cells exist in both cell types, so these are the variables presented. The examples that follow are for central memory T cells but are representative of the types of plots created for all three subsets.

First, pie charts are created for each patient at each available time point (Fig. 14). Remember that the two components do not sum to one, as the values are from the dataset containing all eight memory categories. The relative sizes of the pie slices, however, reveal the relative proportions of memory cells of each cell type, T helper and cytotoxic T. Depending on the objective, these may be grouped by month or patient. Figure 15 gives an example of a graphical display of means when there are only two categories.

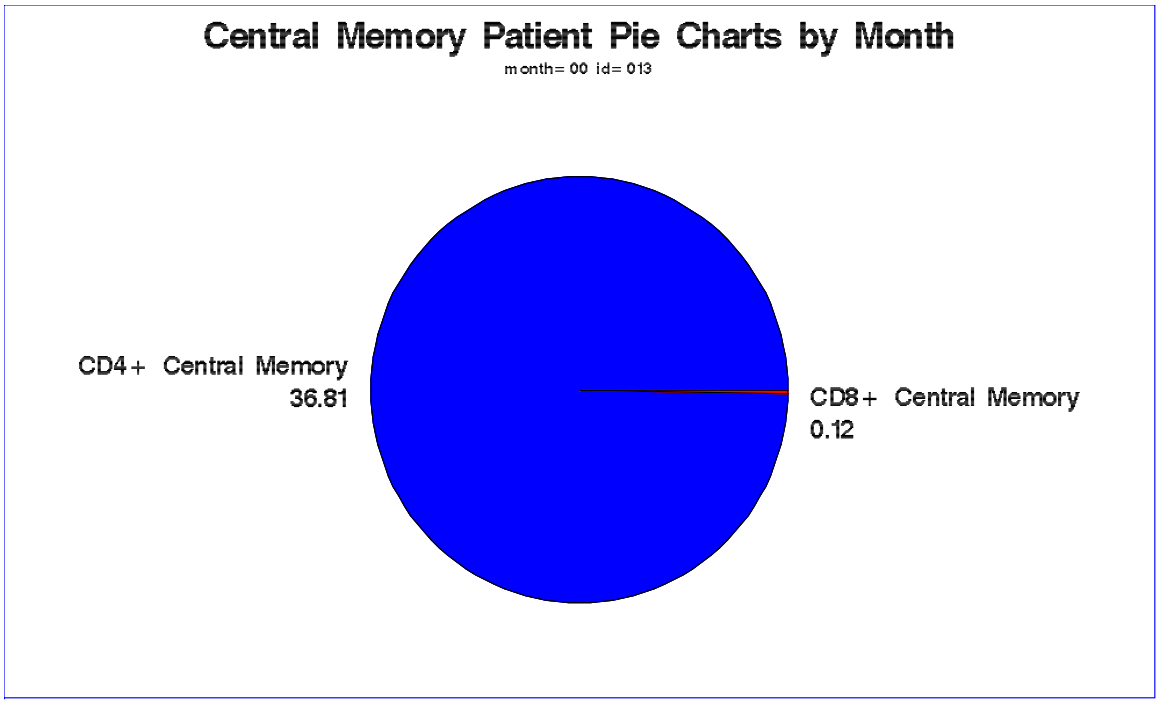


Figure 14: Pie chart of central memory T cells for patient 13 at baseline.
Note: CD4+ denotes T helper cells and CD8+ denotes cytotoxic T cells.

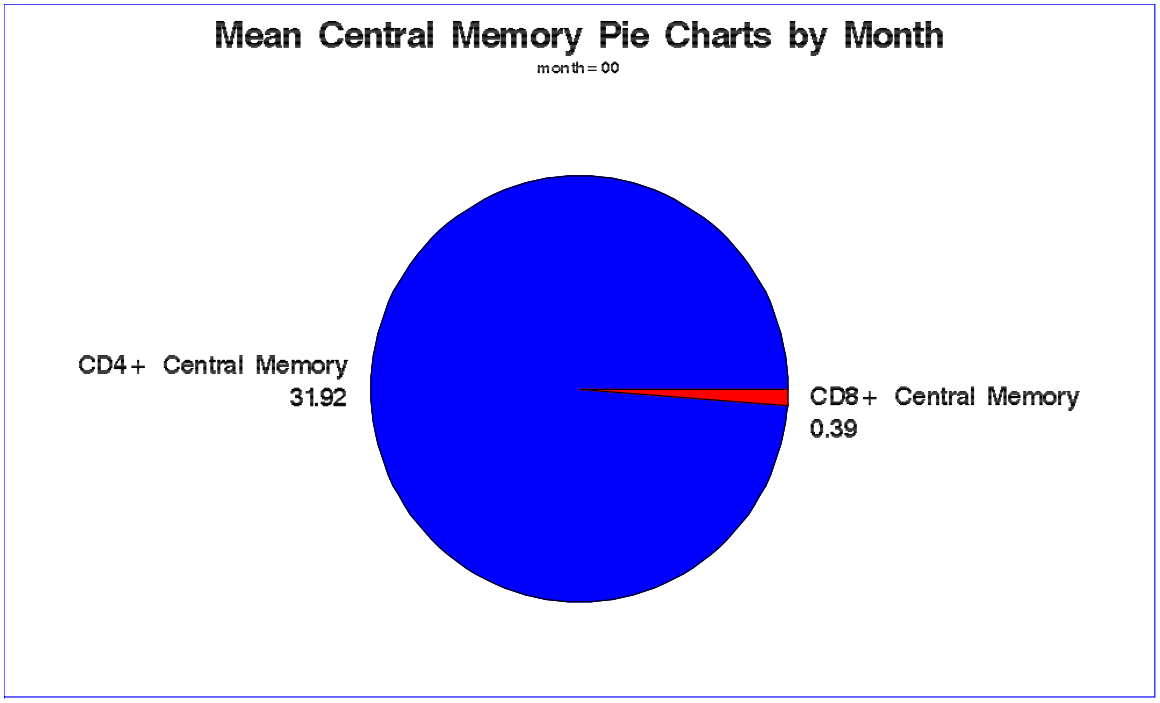


Figure 15: Pie chart of mean central memory T cells at baseline.

Discussion

The application of radar, windrose, star, and pie charts to the PIP dataset has exemplified the utility of these graphical tools for displaying flow cytometry data. Such charts enable faster comprehension than do tables of numerical values. Often statisticians and scientists are inhibited by relying solely on the common tools of their respective fields. Here, data graphics common to fields as diverse as market research and meteorology facilitate the display of immunology data. Though the focus here is on compositional longitudinal clinical data, such plots may be useful for many types of categorical data.

I believe that for the PIP dataset, the radar chart best coincides with Tufte's idea of graphical excellence. It is the most versatile and allows for conveying a large amount of information in a small space. The ability to tile and overlay radar charts permits quick comparisons between patients or time points. The overlaid plots use space most efficiently, but observations are indistinguishable when too many plots are displayed on the same axes. Cleveland warns against such indistinguishable graphics. When there are many plots, tiled charts make the most efficient use of space. Additionally, the radar chart of the mean data summarizes of the entire dataset in a single graphic.

Radar charts produced in SAS®, however, have one major downfall: the scale of the axes. As previously mentioned, data values must be interpolated from only three identified values, the minimum, the maximum, and the value halfway between these two values. Identifying precise values for a category or observation of interest is very difficult. As such, the radar charts lack Tufte's ideals of clarity and precision. The area

between the vertex and the first tick mark is essentially meaningless and wasted space. Though the distance between each tick mark and the next is proportional, the wasted space between the vertex and the first tick mark crams the data values together. The appearance of differences between groups is thus diminished.

A nice feature of windrose charts is the clear distinction between groups. The white spaces between the categories cause data values to stand out. On radar charts and star charts where the values of each category are connected, the placement of categories around the circle may distort the data. Altering the placement of the groups will alter the shape of the enclosed figure. Different figures may cause different perceptions of the data. Windrose charts are immune to this problem.

Additionally, windrose charts do not have the scale distortion characteristic of radar charts. The vertex is zero and the axes are scaled according to the data maximum. Since there is more space to present the range of the data, there is better distinction between the values. Though the axes of windrose charts must be labeled by compass directions, this is no more difficult to read than the clock legend of the radar chart. (There are other options for labeling the axes of radar charts, for example degrees and integers; however, I believe the hours on a clock to be the most easily understood.)

A disadvantage of windrose charts is their inability to display fewer than eight data categories. SAS® requires between eight and sixteen variables for this type of graphical display. Also, perhaps more importantly, windrose charts cannot be tiled or overlaid. Therefore, a separate chart must be created for each data point, and multiple images must be viewed to visualize patterns in the data. The same is true of star charts.

Star charts, however, also have some redeeming qualities. By default, SAS® includes both the data values and the category labels adjacent to each spoke on the chart. This format allows the viewer to take in the information more quickly than if a legend is used. Also, on star charts, each category is displayed with a unique color-pattern combination. Depending on the data, the choice of colors and patterns may be meaningful and facilitate better or faster comprehension of the dataset.

I would like to reemphasize that the purpose of this report is to demonstrate the utility of the visualization tools and not to attribute scientific meaning to the particular dataset. In fact, the dataset contains so few patients that any analyses are likely to be invalid. The original data was used to generate seven different subsets so that plots with different numbers of variables could be presented. The primary examples of each type of chart involved the dataset containing all eight memory categories. These examples were chosen in order to construct high dimensional figures and exemplify the utility of the plots for high dimensional flow cytometry data. These eight variable graphics, however, do not have much clinical meaning. Additionally, the relatively small amount of cytotoxic T cells causes these categories to be greatly overshadowed by the T helper cell categories. The four variable plots for T helper cells and cytotoxic T cells likely provide more relevant clinical information. The pie charts are presented as an option when only two categories are of interest; however, the examples provided in this report are clinically meaningless.

It may be argued that the preceding displays do not make the most efficient use of space or ink. However, these plots were chosen as they are relatively easy to produce. SAS® is commonly used by statisticians. The typical statistician is likely to have easy

access to this software and should have little trouble coding the built-in GRADAR and GCHART procedures. As much as possible, SAS® defaults were used in the generation of the plots. I have tried to balance practicality with good data visualization principles.

This application of radar charts, windrose charts, star charts, and pie charts to data from the Protective Immunity Project has raised a few questions for further investigation. The capabilities of flow cytometry are increasing. Will the graphics presented here be effective in displaying many more variables? SAS® will not create windrose charts with more than sixteen categories. Perhaps a different software is more flexible or can produce graphics that better adhere to Tufte's principles of ink minimization and efficient use of space. Another aspect of data visualization that I did not thoroughly explore is the use of color. The graphics could be enhanced by color and pattern combinations that optimize visual perception and understanding.

Overall, I hope to have shown that there are a number of graphical display options available for high dimensional data. As new technologies capable of producing large amounts of data emerge, we must be creative and innovative in our presentation and explanation of such datasets. We must not neglect the capabilities of technologies such as flow cytometry simply because our typical descriptive and analytical tools are inadequate. Often the necessary methods already exist; we must merely look beyond the confines of our scientific discipline.

References

1. Aitchison, J. A Concise Guide to Compositional Data Analysis, University of Glasgow.
2. Biology-Online.org. (2005). "Naive T-cell." Retrieved February 24, 2009, from http://www.biology-online.org/dictionary/Naive_t-cell.
3. BD Biosciences. (2000). Introduction to Flow Cytometry: A Learning Guide, Becton, Dickinson and Company.
4. Cleveland, W. S. (1985). The Elements of Graphing Data. Monterey, California, Wadsworth Advanced Book Program.
5. Cleveland, W. S. (1993). Visualizing Data. Summit, New Jersey, Hobart Press.
6. Eisenstein, M. (2006). "Cell sorting: Divide and conquer." Nature **441**(7097): 1179-+.
7. Larsen, C. P. and Ahmed, R. (2005). Immune Function and Biodefense in Children, Elderly, and Immunocompromised Populations. Emory University. Department of Health and Human Services, Public Health Service, National Institutes of Health. NIH NO1-AI-50025.
8. MedicineNet, Inc. (2009). "Definition of T-helper Cell." Retrieved February 23, 2009, from http://www.medterms.com/script/main/art.asp?articlekey=11306.
9. National Cancer Institute, U. S. National Institutes of Health. "Dictionary of Cancer Terms." Retrieved February 23, 2009, from http://www.cancer.gov/dictionary/.

10. Saary, M. J. (2008). "Radar plots: a useful way for presenting multivariate health care data." Journal of Clinical Epidemiology **61**(4): 311-317.
11. SAS Institute Inc., SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc., 2000-2004.
12. Tufte, E. R. (1983). The Visual Display of Quantitative Information. Cheshire, Connecticut, Graphic Press.
13. Tufte, E. R. (1990). Envisioning Information. Chesire, Connecticut, Graphic Press.
14. Tufte, E. R. (1997). Visual Explanations: Images and Quantities, Evidence and Narrative. Cheshire, Connecticut, Graphics Press.
15. Willinger, T., Freeman, T., Hasegawa, H., McMichael, A.J., Callan, M.F.C. (2005). "Molecular signatures distinguish human central memory from effector memory CD8 T cell subsets." Journal of Immunology **175**(9): 5895-5903.