

Visualize Tumor Response Data using ggplot2 R Package through Examples

Christine Teng, Merck & Co., Inc., Rahway, NJ USA

ABSTRACT

ggplot2 is a widely used graphics library in R. Plots created using ggplot2 are built using a layered approach. Users can view the work in progress without needing to complete the final work. The code can be developed and executed one layer at a time and adjusted as needed. The purpose of this paper is to utilize the ggplot2 package and few add-on packages to produce multi-layered plots. Most of the examples require just a few lines of code. Oncology tumor response data will be used as input. Users can easily visualize the data and the outcome of tumor assessment over time, using information such as tumor size, new lesion, overall response, etc. with different plot arguments (ex: symbol, color, size, etc.). Many default settings of ggplot2 functions can generate high-quality plots. The examples illustrated in this paper will show that ggplot2 syntax is intuitive and easy to follow.

SAS®9, Windows, Intermediate Level

Key Words: R, ggplot2, RECIST, Oncology

INTRODUCTION

A picture is worth a thousand words. Data visualization has been around for a long time, recently it has become more sophisticated due to new technology and various advanced data analytic tools. Using plots to visualize data is easier than comparing multiple numbers through spreadsheets or tables/listings. Data visualization is a simple way to convey concepts in a universal manner and it can effectively capture outliers, patterns, trends and in some cases, data/logic issues.

The ggplot2 package implements the **grammar of graphics**, a coherent system for describing and building graphs. With ggplot2, you can experiment with different scenarios by making slight adjustments. Useful defaults are provided by ggplot2, however the programming must provide the data, the mappings and the geom functions. You need to learn several functions and arguments of ggplot2 to customize specific requirements. It is recommended that you have a ggplot2 cheat sheet handy as a new beginner.

```
ggplot(data = <DATA>) +  
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

A layer combines data, aesthetic mapping, a geom (geometric object), a stat (statistical transformation), and a position adjustment. An aesthetic is a visual property of the objects in your plot. Aesthetics include things like the size, the shape, or the color of your data. A geom is the geometrical object that a plot uses to represent data such as line, point or boxplot. Use "+" to add layers but please note that it can not be at the beginning of the line.

The plot examples in this paper utilize different functions/layers to tell the data stories. These examples use very basic syntax as the author was learning how to adapt the packages, therefore, they may not be the most optimal approach. However, creating simple yet sophisticated plots is no longer a daunting task.

Below are the R libraries and functions used in this paper. For more information, please refer to the reference links at the end of this paper.

```
library(haven)  
library(ggplot2)  
library(directlabels)  
library(plotly)
```

Aesthetic mapping	aes()
Geometric objects - are the visual representations of observations	geom_rect() geom_line() geom_point() geom_bar() geom_text() geom_abline() geom_dl()
Coordinates of x, y	coord_cartesian() coord_flip()
Scales set default aesthetic value	scale_size_continuous() scale_color_brewer() scale_colour_discrete() scale_fill_manual() scale_y_continuous() scale_color_manual()
Guides for each scale can be set scale-by-scale	guides()
Facetting (multiple sub plots on same page)	facet_wrap()
Labels customize title/subtitle/caption and x, y labels, etc.	labs()
Themes customize the non-data components of plot	theme() theme_bw()

We will use dummy SAS datasets as input to data frame in ggplot2 for following examples.

R Syntax example: `trvd <- read_sas(paste0(folder, "trvdir.sas7bdat"))`

EXAMPLES

Based on RECIST (Response Evaluation Criteria in Solid Tumors) v1.1 rule for Target Response below (Table 1), users can visualize the response data change over time using ggplot2. This paper will not cover how to set up the input used by the plots. Please refer to the paper presented at PhamaSUG 2018 <https://www.lexjansen.com/pharmasug/2018/AD/PharmaSUG-2018-AD02.pdf> to understand the details of key derived variables used in solid tumor studies. This paper assumes readers have basic knowledge of the oncology response data to interpret the plots.

Table 1 – RECIST Target Lesion (TL) Response Criteria

Response	Definition
Complete Response (CR)	All non-nodal TLs disappeared; all lymph nodes short axis <10 mm
Partial Response (PR)	SOD decreased $\geq 30\%$ from baseline
Progressive Disease (PD)	SOD increased $\geq 20\%$ from nadir and the 20% has absolute increase ≥ 5 mm
Stable Disease (SD)	Not PR nor PD
Not Evaluable (NE)	Cannot determine target lesion response

Understanding the data variables/columns in the data frame of the examples

- SUBJECT – Patients
- WEEK – Assessment time points
- BASELINE – Baseline sum of diameter
- SOD – Sum of diameter (sum of longest diameters and sum of diameter of non-lymph-node tumor)

- PCBSD – Percent change from baseline in sum of diameter
- NADIR – Smallest sum of diameter prior to current time point
- PCNSD – Percent change from Nadir in SOD
- NLR – New lesion (Y, N)
- OVRLRESP – Overall response CR, PR, PD, SD, and NE
- RSPDUR – Response duration
- DTH – Death day from the reference start date (First dose date or randomization date)
- FPD – First progression day from the reference start date
- FCRPR – First response day from the reference start date
- LASTD – Last dose day from the reference start date
- EVNTDESC – The event description of the patient
- BESTCHG – minimal of tumor size assessment change from baseline (for water fall plot)
- BOR – Best confirmed response over time

RECIST Responders – Subjects whose best confirmed overall response (BOR) is CR or PR

irRECIST Responders – Subjects who has positive response after RECIST PD

Through data visualization, we can see how tumor shrinkage or overall response changes over time. When the new lesion indicator is Y, we can see the overall response status is set to PD per the RECIST rule.

RECIST Non-responders but irRECIST Responders

Pseudoprogression is defined as radiologic tumor progression from baseline that is not confirmed as progression on subsequent radiologic assessments. This is commonly seen in some immunotherapy studies. irRECIST is a sponsor defined approach to identify such potential cases. irRECIST algorithms will not be discussed in this paper but will be used as a data example for first three cases below to demonstrate possible cases of pseudoprogression through visualization.

For the 3 examples below, BOR is labeled with subjid and we see that subject 100002 has a BOR of PD but the tumor size actually decreased over time.

Figure 1 code using default without customization code

```
ggplot(trvd, aes(x = WEEK, y = PCBSD, color=SUBJECT)) +
  geom_line(aes(color = SUBJECT)) +
  geom_point(aes(shape=OVRLRESP, size=SOD)) +
  geom_point(aes(color=NLR))
```

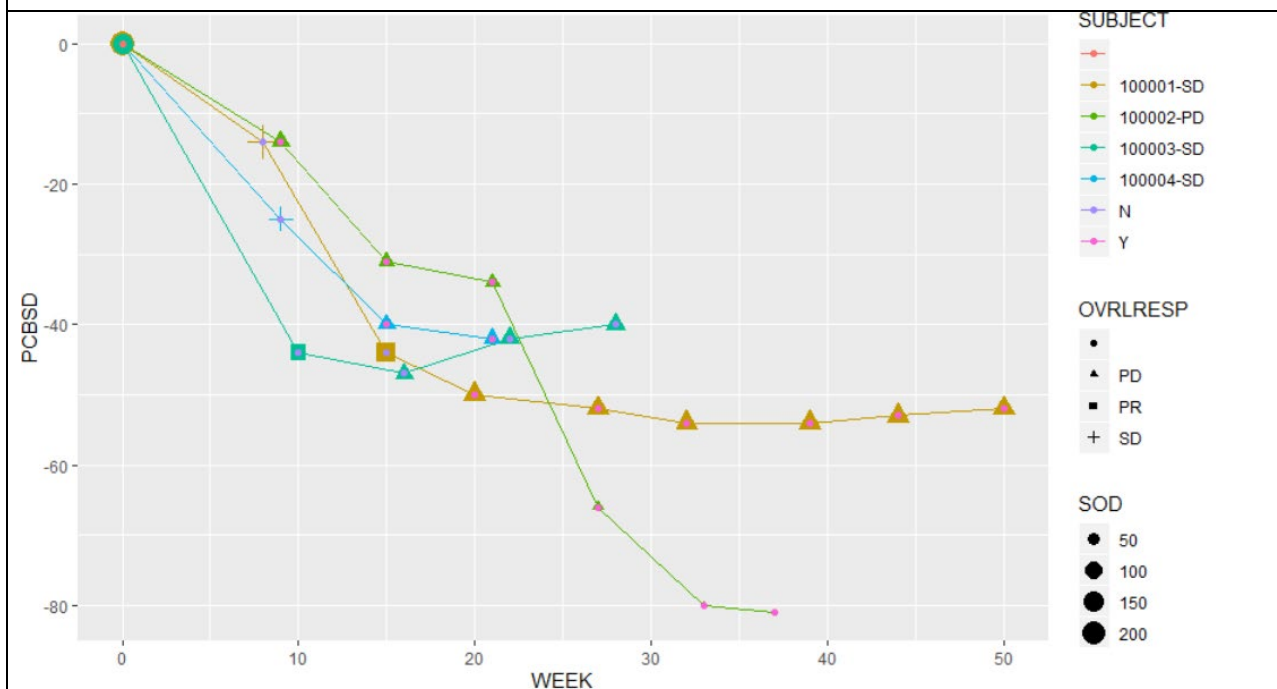
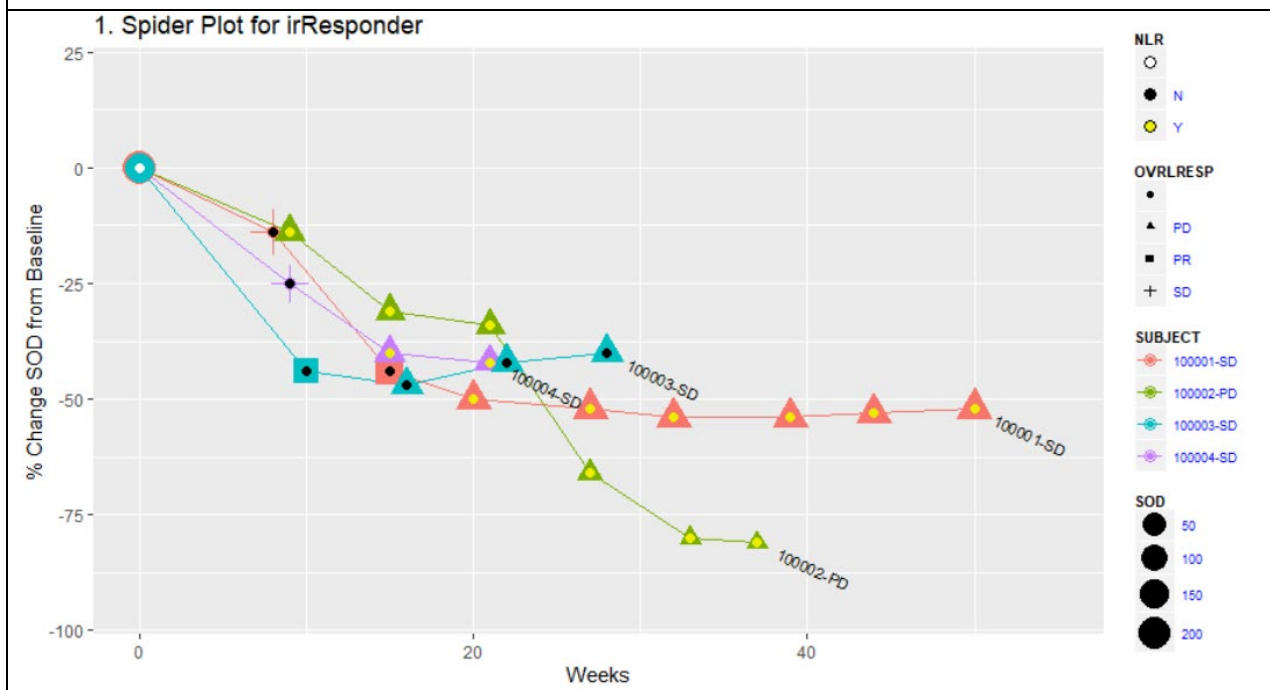


Figure 1 code with additional customizations for readability:

```
ggplot(trvd, aes(x = WEEK, y = PCBSD, color = SUBJECT)) +
  geom_line(aes(color = SUBJECT)) +
  geom_point(aes(shape=OVLRESP, size=SOD)) +
  geom_point(aes(fill=NLR, size=3, shape=21)+
  geom_dl(aes(x = WEEK,
    y = PCBSD, label = SUBJECT),
    color="black",
    method = list("last.points", cex = 0.7, rot = -25, hjust = -.3)) +
  coord_cartesian(xlim=c(0,55), ylim=c(-95,20)) +
  scale_size_continuous(range = c(4, 8))+
  scale_fill_manual(values=c("white", "black", "yellow2")) +
  labs(y="% Change SOD from Baseline",
    x="Weeks",
    title="1. Spider Plot for irResponder") +
  theme(legend.title = element_text(color = "black", size = 8, face="bold"),
    legend.text = element_text(color = "blue", size = 7),
    legend.position = 'right',
    legend.spacing.y = unit(0.01, 'cm'))
```



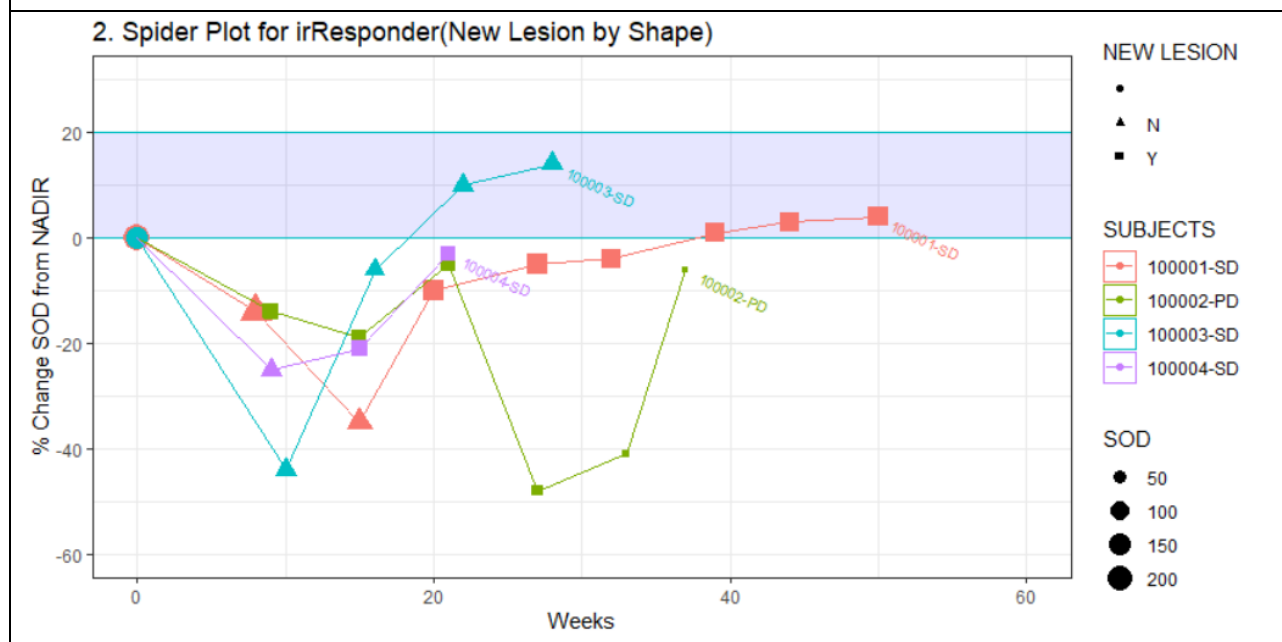
In Figure 1

1. %Change of tumor size from baseline over time in weeks (Y – axis and X – axis)
2. Subject ID and its BOR – shown through color and text annotation
3. Overall response at each assessment – shown through symbols
4. The tumor size (SOD) at each timepoint (increased or decreased) – shown through symbol size
5. Which assessment has new lesion (which leads to PD) – through color scheme in colored circle (N-black, Y-yellow)

The below plot uses the **directlabels** package to label the subject information. The label can be adjusted for different rotations and heights.

Figure 2 code:

```
ggplot(trvd, aes(x = WEEK, y = PCNSD, color = SUBJECT)) +
  geom_rect(aes(xmin = -Inf, xmax = Inf, ymin = 0, ymax = 20), fill = "blue", alpha=0.005) +
  geom_point(aes(shape=NLR, size=SOD)) +
  geom_line() +
  geom_dl(aes(x = WEEK,
              y = PCNSD,
              label = SUBJECT),
          method = list("last.points", cex = 0.6, rot = -25, hjust = -.2)) +
  coord_cartesian(xlim=c(0,60), ylim=c(-60,30)) +
  labs(y="% Change SOD from NADIR",
       x="Weeks",
       title= "2. Spider Plot for irResponder(New Lesion by Shape)",
       color = "SUBJECTS",
       shape="NEW LESION") +
  theme_bw()
```



In Figure 2,

1. %Change of tumor size from nadir over time in weeks (Y – axis and X – axis)
2. Subject ID and its BOR – shown through color and text annotation (non responder only) using the add-on directlabel package
3. The tumor size (SOD) at each timepoint (increased or decreased) – shown through symbol size
4. Which assessment has New Lesion (which leads to PD) – shown through symbol shape

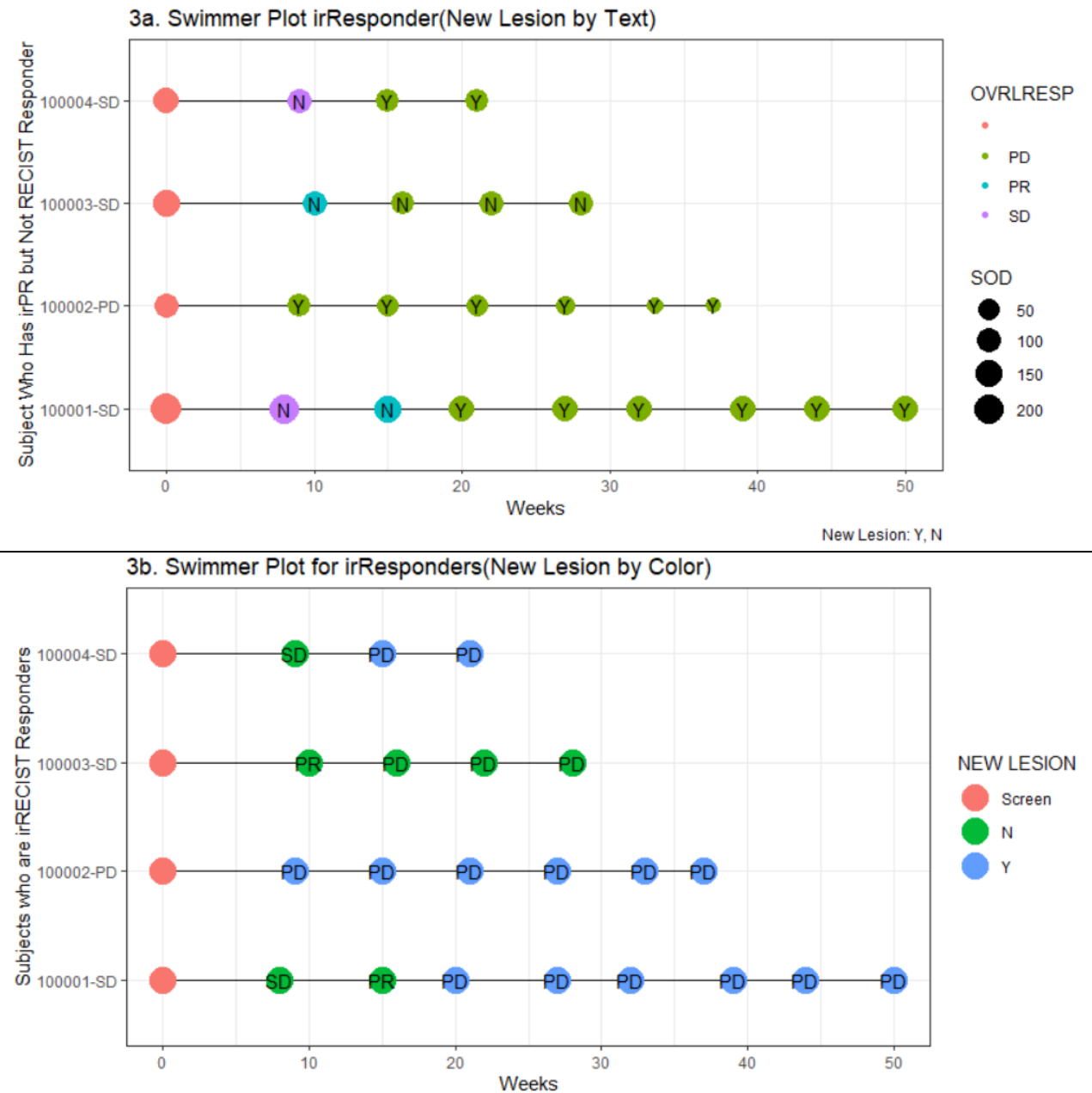
The below plots demonstrate how to identify new lesion using text or color scheme. Use the guides() function to provide text description instead of using default variable names in legend section. Please note that since ggplot2 uses layered approach, text is drawn after symbols/points in order to be seen on top of the symbols.

Figure 3 code:

```
#3a
ggplot(trvd, aes(x = WEEK, y = SUBJECT)) +
  geom_line()+
  geom_point(aes(color = OVRLRESP, size=SOD)) +
  geom_text(aes(label=NLR), color="black") +
  labs(y="Subject Who Has irPR but Not RECIST Responder",
       x="Weeks",
       title="3a. Swimmer Plot irResponder(New Lesion by Text)",
       caption = "New Lesion: Y, N") +
  scale_size_continuous(range = c(4, 8)) +
  coord_cartesian(xlim=c(0,50)) +
```

```
theme_bw()
```

```
#3b
ggplot(trvd, aes(x = WEEK, y = SUBJECT)) +
  geom_line()+
  geom_point(aes(color = NLR ), size=7) +
  geom_text(aes(label=OVLRESP))+
  scale_colour_discrete(name="NLR", labels=c("Screen", "N", "Y")) +
  labs(y="Subjects who are irRECIST Responders",
       x="Weeks",
       title="3b. Swimmer Plot for irResponders(New Lesion by Color)") +
  theme_bw() +
  guides(color=guide_legend("NEW LESION"))
```



In Figure 3a,

1. Subject ID and its BOR and associated overall response over time (Y – axis and X – axis)
2. Overall response – shown through symbol colors

3. SOD – shown through symbol sizes
4. New lesion – shown through text within circle symbol

In Figure 3b,

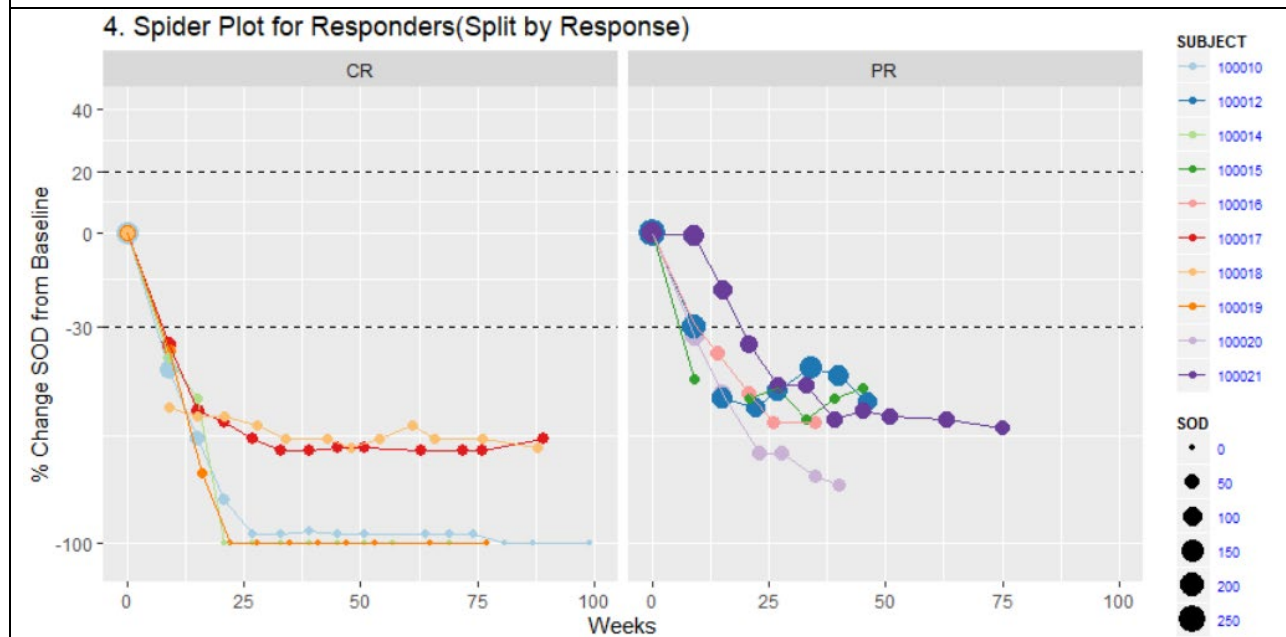
1. Subject ID and its BOR and associated overall response over time (Y – axis and X – axis)
2. Overall response – shown through text inside symbol circle
3. New lesion – shown through symbol colors

RECIST Responders

The below two examples use responder data to demonstrate two methods (side by side or up and down) to display split plots. Please note that each sub-plot shares the same legend by default.

Figure 4 code:

```
ggplot(trvdc, aes(x = WEEK, y = PCBSD, color = SUBJECT)) +
  geom_abline(slope=0, intercept=20, col = "black", lty=2) +
  geom_abline(slope=0, intercept=-30, col = "black", lty=2) +
  geom_point(aes(size=SOD), na.rm = TRUE) +
  geom_line() +
  labs(y="% Change SOD from Baseline", x="Weeks",
       title="4. Spider Plot for Responders(Split by Response)") +
  scale_y_continuous(breaks=c(-100,-30,0, 20, 40)) +
  coord_cartesian(xlim=c(0,100), ylim=c(-105,40)) +
  facet_wrap(~BOR) +
  scale_color_brewer(palette="Paired") +
  theme(legend.title = element_text(color = "black", size = 8, face="bold"),
        legend.text = element_text(color = "blue", size = 7),
        legend.position = 'right',
        legend.spacing.y = unit(0.001, 'cm'))
```



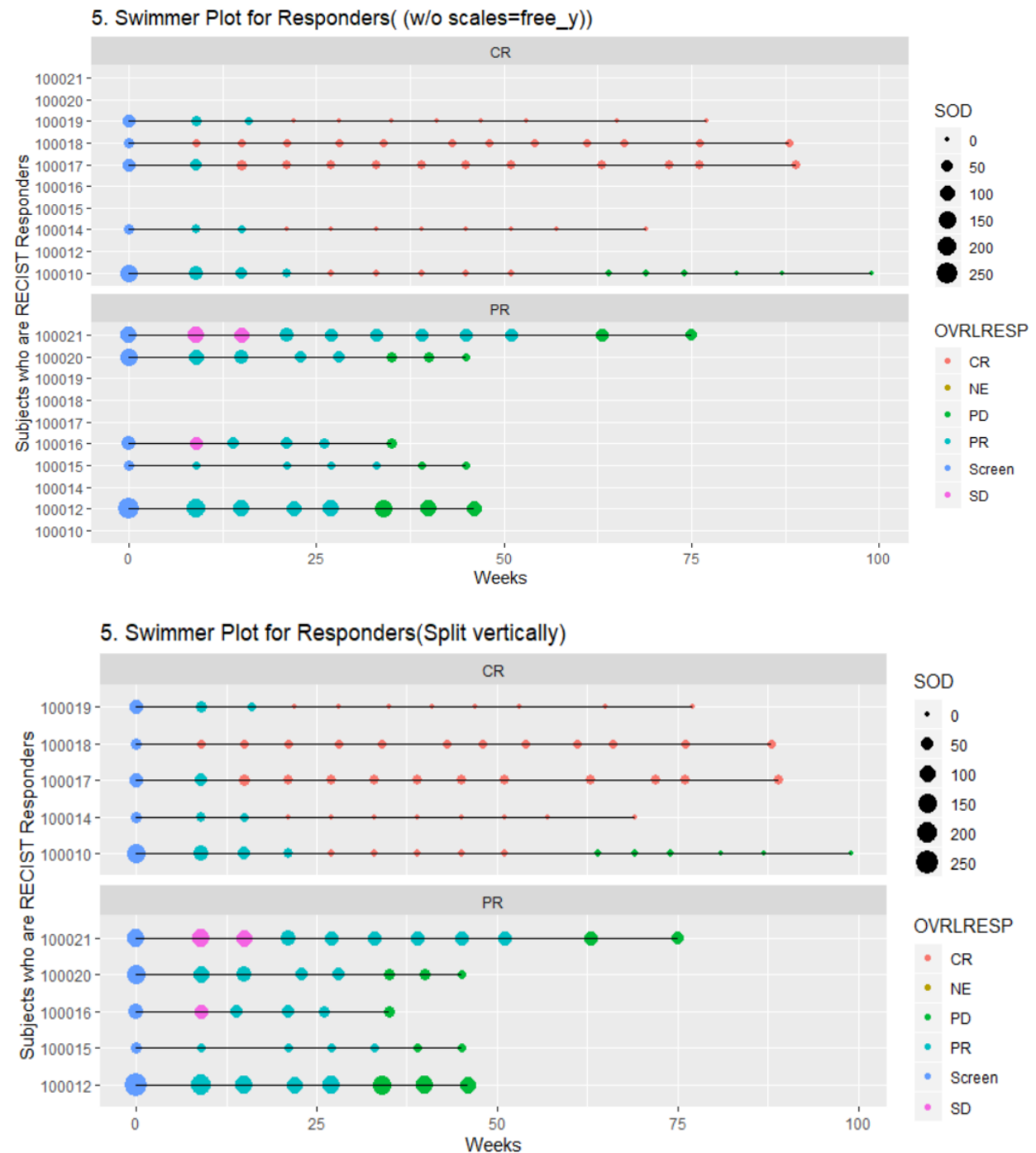
In Figure 4,

1. %Change of tumor size from baseline over time in weeks (Y – axis and X – axis), split data by responder CR and PR in two columns
2. Subject ID – shown through colors
3. The tumor size (SOD) at each time point – shown through symbol size

Facet_wrap's scale='free_y' in below code removes empty subject data from each sub-plot.

Figure 5 code:

```
ggplot(trvdcr, aes(x = WEEK, y = SUBJECT)) +
  geom_point(aes(size=SOD, color = OVRLRESP), na.rm = TRUE) +
  geom_line() +
  labs(y="Subjects who are RECIST Responders",
       x="Weeks",
       title="5. Swimmer Plot for Responders(Split vertically)") +
  facet_wrap(~BOR, ncol=1, scales = 'free_y')
```



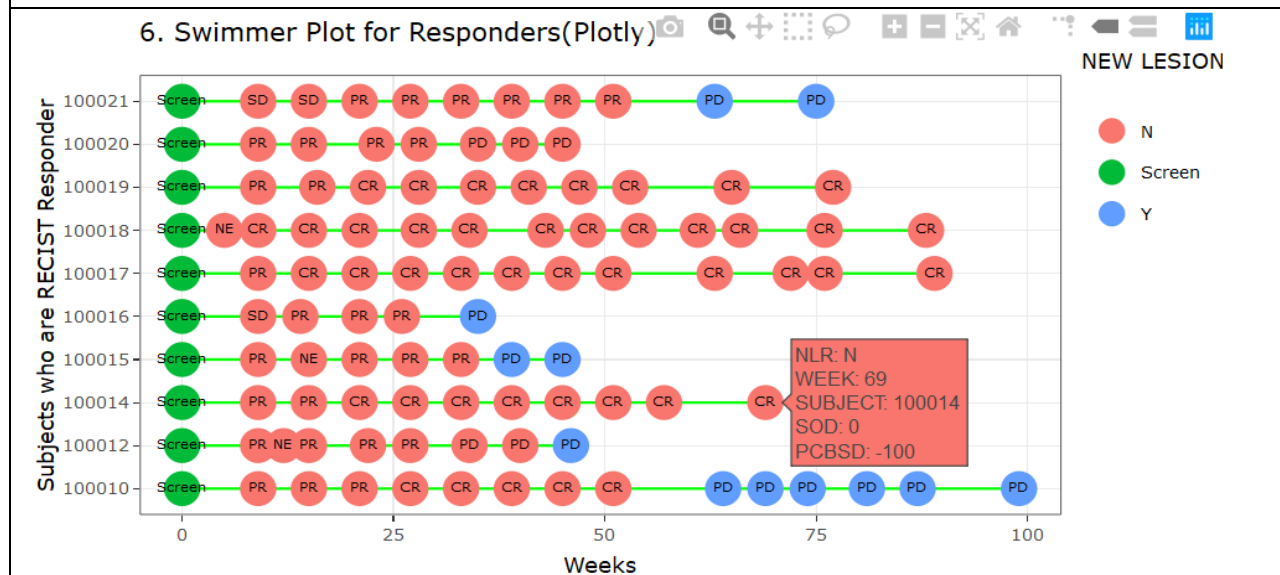
In Figure 5,

1. Subjects overall response over time (Y – axis and X – axis), Split data by responder CR and PR in two rows
2. Overall response – shown through colors
3. The tumor size (SOD) at each time point – shown through symbol size

The below plot was saved as an object before using the “plotly” function.

Figure 6 code:

```
p<- ggplot(trvdcr, aes(x = WEEK, y = SUBJECT, name=SOD, name1=PCBSD)) +
  geom_line(color = "green") +
  geom_point(aes(color =NLR ), size=6, na.rm = TRUE) +
  geom_text(aes(label=OVRLRESP), size=2.5) +
  labs(y="Subjects who are RECIST Responder",
       x="Weeks",
       title="6. Swimmer Plot for Responders(Plotly)",
       color = "NEW LESION")+
  theme_bw()
ggplotly(p)
```



In Figure 6,

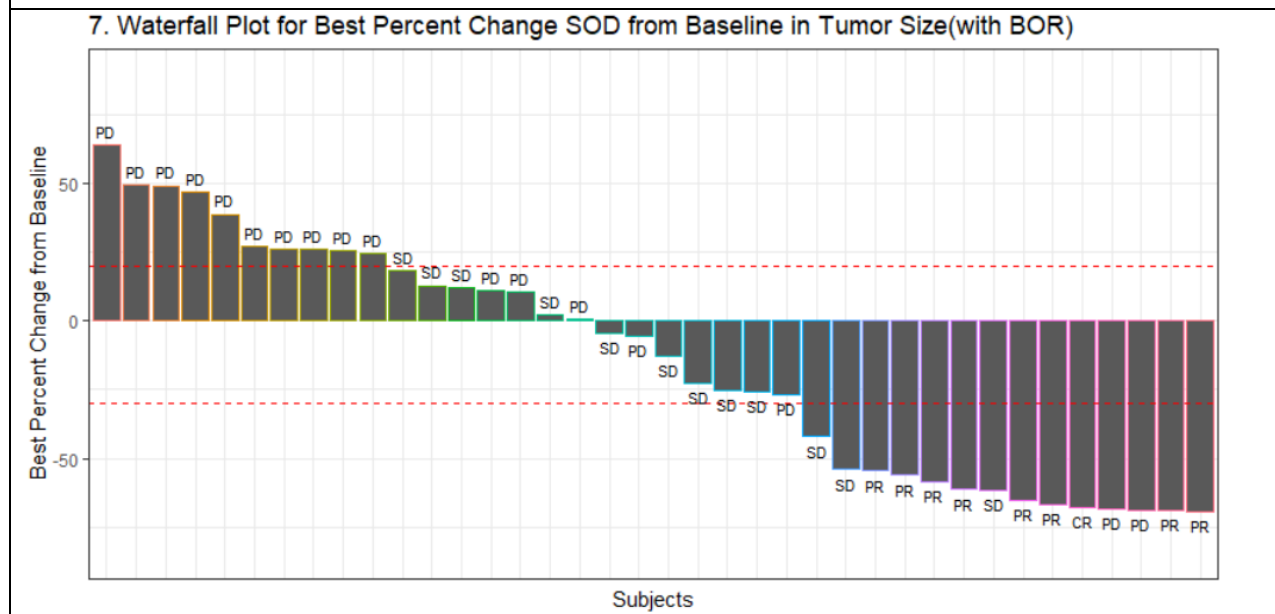
1. Subjects overall response over time (Y – axis and X – axis) for responders (CR or PR)
2. Overall response at each assessment week – shown through text annotation
3. New lesion – through color scheme
4. Additional information when users hover over the response assessment – through interactive box using add-on gplotly function

Other Common Oncology Plots

The below bar plot added BOR label for each bar/subject. The distance/position BESTCHGX is calculated in the dataset based on the BESTCHG result. The input data is sorted from the worst to the best result. Legend and X-tics are suppressed.

Figure 7 code:

```
ggplot(subjw, aes(x = SUBJECT, y = BESTPCHG, col = SUBJECT)) +
labs(title = "7. Waterfall Plot for Best Percent Change SOD from Baseline in Tumor Size(with BOR)",
x = "Subjects",
y = "Best Percent Change from Baseline") +
coord_cartesian(ylim = c(-85,90)) +
geom_bar(stat = "identity", width = 0.9) +
geom_abline(slope=0, intercept=20, col = "red", lty=2) +
geom_abline(slope=0, intercept=-30, col = "red", lty=2) +
geom_text(aes(x = SUBJECT,
y = BESTPCHX, label=BOR),
color="black", size=3) +
theme_bw() +
theme(
axis.text.x = element_blank(),
axis.ticks.x = element_blank(),
legend.position="none")
```



In Figure 7,

1. Best %change of tumor size from baseline by subjects (Y – axis and X – axis)
2. Best overall response for each subject – shown through text annotation

Example 8 and 9 provide similar plots by adding different features and customizations

The below bar plot was flipped, and the symbol and color were customized instead of using the default.

Figure 8 code:

```
ggplot(subjs, aes(x = SUBJECT, y = RSPDUR)) +
geom_bar(stat = "identity", fill="grey") +
geom_point(aes(y = FPD, col="First PD", fill="red", shape=24, size=3) +
geom_point(aes(y = FPRCR, col="First CR/PR", fill="black", shape = 21, size = 3) +
geom_point(aes(y = LASTD, col="Last Dose"),fill="blue", shape=23,size=3) +
geom_point(aes(y = DTH, col="Death"), fill = "red", shape = 22, size=3) +
coord_flip() +
geom_text(aes(x = SUBJECT,
y = 20, label=EVNTDESC),
color="black", size=2) +
labs(x = "Subjects",
title = "8. Swimmer Plot for Responders(Add Event Description)") +
scale_y_continuous("Months", limits = c(0,23),
breaks=seq(0,23,1),
expand = c(0,0.1)) +
```

8. Swimmer Plot for Responders(Add Event Description)



- A more typical oncology swimmer plot is ordered by response duration. Instead of hard coding the text description with distance from end of duration, NUDGE is used to move labels to a certain distance for each piece of data. The theme() function is used to update the title position and to modify the colors for different symbols. Since not all subjects have a death or PD date, na.rm=TRUE is used to remove warning about missing value.

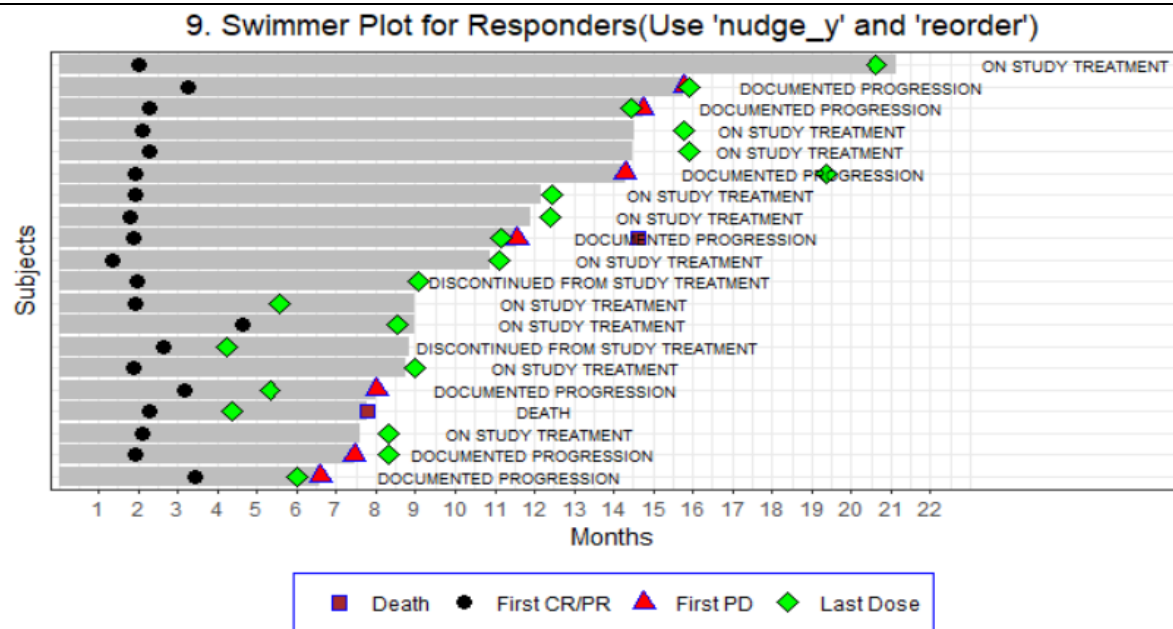
```
ggplot(subjs, aes(x = reorder(SUBJECT, RSPDUR), y = RSPDUR)) +
  geom_bar(stat = "identity", fill="grey") +
  geom_point(aes(y = FPD, col="First PD"), fill="red", shape=24, size=3, na.rm = TRUE) +
  geom_point(aes(y = FPRCR, col="First CR/PR"), fill="black", shape = 21, size = 3) +
  geom_point(aes(y = DTH, col="Death"), fill = "brown", shape = 22, size=3, na.rm = TRUE) +
  geom_point(aes(y = LASTD, col="Last Dose"), fill="green", shape=23, size=3) +
  coord_flip() +
  theme_bw() +
  geom_text(aes(x = SUBJECT,
                y = RSPDUR, label=EVNTDESC,
                nudge_y = 4.5,
                color="black",
                size=2.5)) +
```

11

```

fill = c("brown", "black", "red", "green")))) +
theme(plot.title = element_text(hjust = 0.5),
      legend.position = "bottom",
      legend.background = element_rect(linetype = "solid", color = "blue"),
      axis.text.y = element_blank(), axis.ticks.y = element_blank())

```



In Figure 9,
1. Similar to the figure 8 but result is displayed in descending order

SUMMARY

Using R ggplot2, programmers can quickly produce data plots without creating complicated programs, and the plots can be easily modified. R is a free software for statistical computing and graphics. The ggplot2 package is part of the popular tidyverse package which can be downloaded for free as well. For SAS programmers who are new to R, ggplot2 is a great starting point to learn R programming. The above examples demonstrate ggplot2 layer graphical programming. ggplot2 creates informative and elegant graphs with relatively simple, short and readable code. Deciphering and cleaning the data as well as selecting the right statistical methods are important first step in creating meaningful plots to convey the data story. The challenges of data visualization involve preparing a proper data structure which can be analyzed easily.

REFERENCES

Garrett Golemund, Hadley Wickham. "R for Data Science". O'Reilly, January 2017.
<https://r4ds.had.co.nz/index.html>

ggplot2 Cheat Sheet by RStudio®.
<https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>

A ggplot2 grammar guide
https://evamaerey.github.io/ggplot2_grammar_guide/ggplot2_grammar_guide

ggplot2 Reference
<http://ggplot2.tidyverse.org/reference>

ggplot2 Extensions/Gallery:
<https://www.ggplot2-exts.org/gallery/>

directlabel using geom_dl()

https://www.rdocumentation.org/packages/directlabels/versions/2018.05.22/topics/geom_dl

Getting Started with Plotly and ggplot2
<https://plot.ly/ggplot2/getting-started/>

New Response Evaluation Criteria in Solid Tumors: Revised RECIST Guideline (Version 1.1), European Journal of Cancer, 45:228-247
https://ctep.cancer.gov/protocoldevelopment/docs/recist_guideline.pdf

ACKNOWLEDGEMENTS

The author would like to thank her management team for their encouragement and review of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Christine Teng Merck & Co., Inc. Rahway, NJ 07065 christine_teng@merck.com	
---	--



Brand and product names are trademarks of their respective companies.