

Implementing Reproducibility at MDACC: A Start

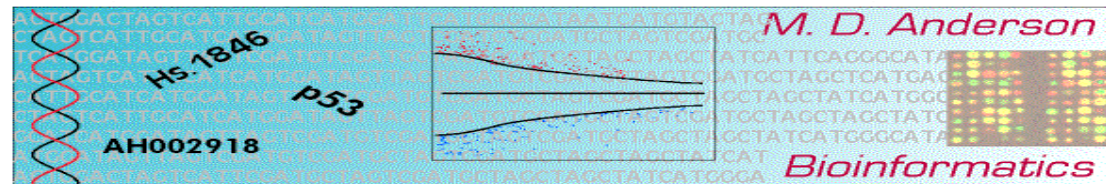
Keith A. Baggerly

Bioinformatics and Computational Biology

UT M. D. Anderson Cancer Center

kabagg@mdanderson.org

SISBID, July 17, 2018



Pressures and Rollout

Several core services at MD Anderson are supported by the Cancer Center Support Grant (CCSG).

Biostatistics is one such service.

In light of the recent NIH Initiative, some members of the Scientific Advisory Board asked how reproducibility of internal memoranda was pursued.

This led to discussions in June about whether some of the tools we've been discussing could be of use.

Could we have everything perfect by September?

Heh. No.

In talking about improving reproducibility, I'm all too aware I could often do a better job myself.

Improving reproducibility is a process.

It can require changing work habits, which is hard.

It's easier to do if we can make improvements in stages, and see benefits without undue burden.

An initial chunk of time to get familiar with a tool is fine, but a regular slowdown of common tasks is not.

So, what is a common task? Can we show it can work?

The Challenge


Take a typical memorandum produced by an analyst a few years ago (ms published, so sharing de-identified data ok), where analyses were performed in a mixture of SAS and STATA, and summarized in a MS Word writeup.

Produce the same results and “look and feel” using R and R markdown, with enough detail that this could be used as a template by others in the department.

I worked with three analysts (mixed R backgrounds) on this - we identified subtasks, divvied them up, and had at it.

Let's look at the target.

The Header

		Memorandum
Department of Biostatistics		Telephone: 713-563-4275
		Fax: 713-563-4242
To:	Amanda Cooper, MD, Matt Katz, MD Department of Surgical Oncology	
From:	Rebecca Slack, MS Department of Biostatistics	
Subject:	Final Report for Sarcopenia Measures from Trial 01-341 for the First of 2 Manuscripts.	
Date:	July 14, 2014	
<hr/>		
Goal: To report sarcopenia prevalence and its relationship with <u>resectability</u> , OS, and PFS among pancreatic cancer patients on trial 01-341. The sections below are designed to be pasted into a		

Note: logo, heading/preamble

Inline Results

Results. A total of 89 pancreatic cancer patients were available for analysis and had dates of initial CT from October 2001 to January 2006. Table 1 presents the patient characteristics and shows that patients had median age of 63 years, were primarily male (55%), White/non-Hispanic (87%), overweight/obese (59%), with a history of smoking (61%), and Sarcopenia present before treatment (52%).

Patients' change in muscle and adipose tissue measures are presented in Table 2. Skeletal muscle decreased on average 1.2 units ($p < 0.01$) after treatment. Visceral adipose decreased 4.0 units ($p = 0.01$). Subcutaneous adipose measures were only available pre and post for about half of the patients. Among those patients, subcutaneous adipose dropped on average 37 units after treatment ($p < 0.001$).

Note: these numbers were initially read off from separately generated tables

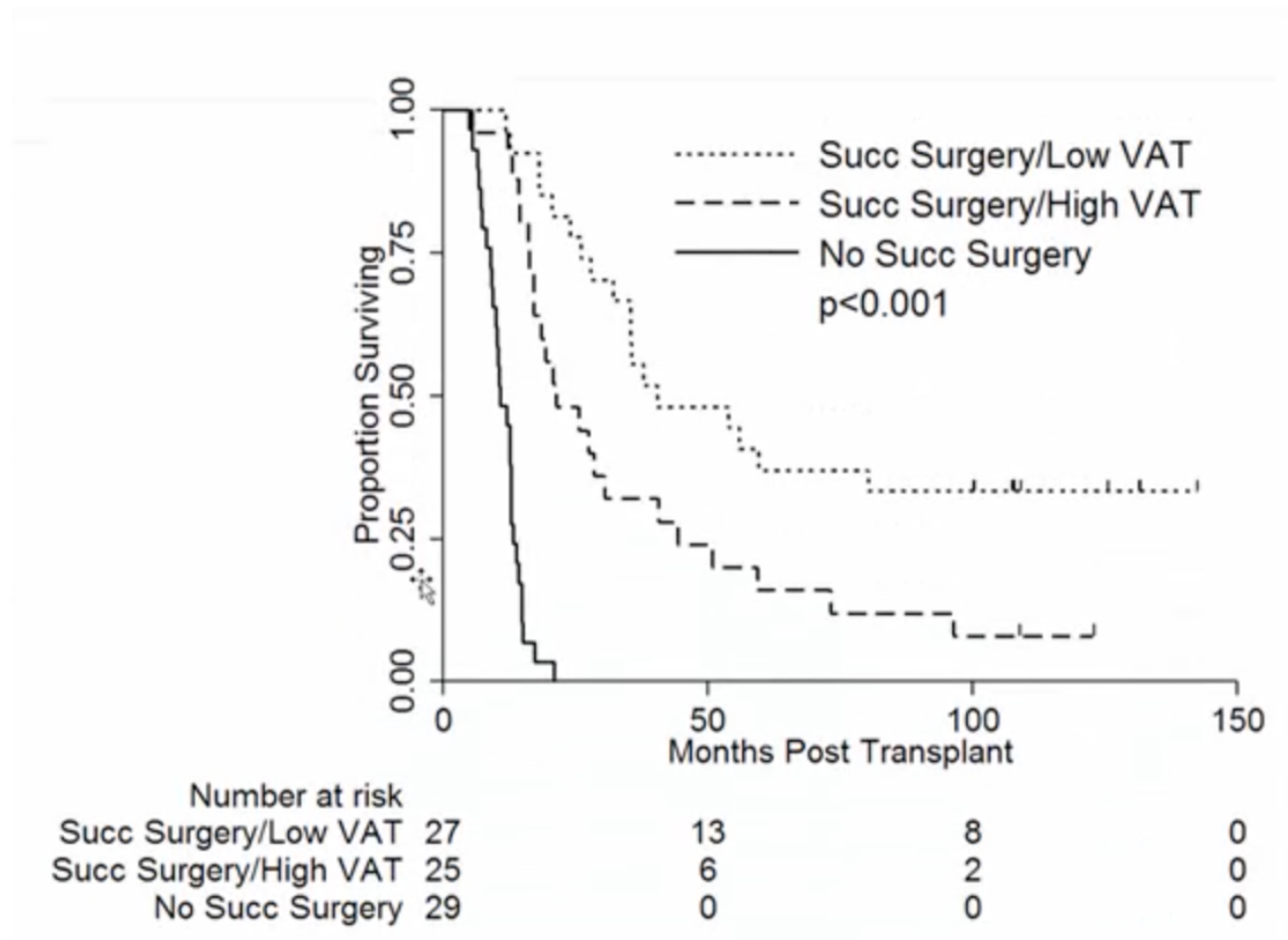
Tables

⊕ Table 1. Patient Characteristics Prior to Treatment

Patient Characteristics	All N (%)
All Patients	89 (100%)
Age - median (<u>min,max</u>)	
N=90	63 (38, 79)
Gender	
F	40 (45%)
M	49 (55%)
Race Ethnicity	
White/Non-Hispanic	77 (87%)
White/Hispanic	6 (7%)
Asian/Non-Hispanic	3 (3%)
Black/Non-Hispanic	3 (3%)
BMI	

Note: initially generated in SAS, some polishing in Word

Figures



Note: standard publication type K-M plot, with numbers at risk

Formatted References

References

1. Mehta CR, Patel NR: A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables. *Journal of the American Statistical Association* 78:427-434, 1983
2. Kaplan EL, Meier P: Nonparametric estimator from incomplete observations. . *J Am Stat Assoc* 53:457-481, 1958
3. Cox DR: Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 34:187-+, 1972
4. Lausen B, Sauerbrei W, Schumacher V: Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. *Computational Statistics*:483-496, 1994

Arranged and cited in the style requested by the target journal

Headers: MS Word is a Pain

It's not my preference, but it is the format most of our collaborators prefer.

Customizing Word output is not, at present, something which lends itself to complete scripting within RStudio.

It involves the interface between R, pandoc, and MS Word

Need to interact with MS Word to generate docx files whose settings can be “inherited”

The Process

Start with a “vanilla” Rmd file (stuff01.Rmd)

From RStudio, produce a default docx file (stuff01.docx)

In Word, open this docx file and shift a few parameter settings (e.g., telling it to allow for page headers, and include an image)

Save the modified docx file as a new “reference” (ref01.docx) in the same folder as the Rmd file, or be prepared to specify a path

Go to a new copy of the Rmd file (stuff02.Rmd), and edit the YAML

The YAML

```
## stuff01.Rmd
```

```
output:
```

```
  word_document: default
```

```
  html_document: default
```

```
## stuff02.Rmd
```

```
output:
```

```
  word_document:
```

```
    reference_docx: ref01_addLogo.docx
```

```
  html_document: default
```

iterate, adding new tweaks each time

The Output



The logo is cosmetic, but nice. The same approach is also used for more basic tweaks, such as adding page numbers.

Contact: Tables in Markdown

```
` `` {r, echo=FALSE}
```

The table below is needed as is to stretch across the whole page. The end of the telephone number must line up with the far right dashes

```
` ``
```

```
-----
```

```
- - - - -
```

Department of Biostatistics

Telephone: 713-563-4275

```
-----
```

```
- - - - -
```

Address Fields: Markdown Tables 2

To: Amanda Cooper, MD

 Matt Katz, MD

 Department of Surgical Oncology

From: Rebecca Slack, MS

 Department of Biostatistics

What about the specified YAML author and title fields?

Keep Only What Works

```
---  
title: ''  
author: ''  
date: ''  
output:  
  word_document:  
    reference_docx: ref02_addTextBox.docx  
  pdf_document: default  
  html_document: default  
header-includes: \usepackage{caption}  
cs1: nature-genetics.cs1  
bibliography: becky.bib  
---  
\captionsetup[table]{labelformat=empty}  
  
***
```

A kludge, but it works...

The Output

Department of Biostatistics

Telephone: 713-563-4275



To: Amanda Cooper, MD
Matt Katz, MD
Department of Surgical Oncology



From: Rebecca Slack, MS
Department of Biostatistics

Subject: Final Report for Sarcopenia Measures from Trial 01-341 for the First of 2 Manuscripts

Date: July 14, 2014



Inline Values in the Abstract?

```
```{r, labeledChunkBeforeAbstract, echo=FALSE}  
...
```
```

my inline value `r sum(x)` here

```
```{r, labeledChunkAfterAbstract, eval=FALSE}  
<<labeledChunkBeforeAbstract>>
```
```

Where the code is displayed need not match where it is run.

The Data Tables

Tables are hard to get right. There are a lot of user-tunable parameters, which Word often hides. Some of this is visible if you try working with tables in \LaTeX ...

There are a few different R packages for this

kable (easiest to use “out of the box”)

xtable (often allows for more customization)

pander (can take summary objects directly)

ascii

Some Links for Comparisons

`https://rpubs.com/benmarwick/tables-rmarkdown`

`http://kbroman.org/knitr_knutshell/pages/figs_tables.html`

`https://gist.github.com/benmarwick/7797391`

Our final choice here:

kable allowed for the most consistent output when working with MS Word.

The Code, Part 1

```
Table1 <-  
  chrTable(  
    numerical = c('agedx', 'adjustedSATpre',  
                  'adjustedVATpre'),  
    categorical = c('Gender', 'raceethnicity',  
                   'BMIgrp', 'Smoking_Hx', 'PreTxSarcopenia'),  
    data = myData,  
    missingInd = c('.', ''),  
    label = c('Age', 'Pretreatment Adjusted SAT',  
              'Pretreatment Adjusted SKM',
```

The Code, Part 2

```
'Pretreatment Adjusted VAT',  
'Gender', "Race Ethnicity", 'BMI',  
'Smoking History',  
'Pretreatment Sarcopenia*'))
```

```
kable(Table1, caption = "Table 1. Patient  
  Characteristics Prior to Treatment",  
  align = c(rep("l", 2), rep("c", 1)))
```

The actual kable invocation is pretty basic!

The Output

Table 1. Patient Characteristics Prior to Treatment

| Patient Characteristics | | N (%) |
|--|--------------------|-------------------|
| All Patients | | 89 (100%) |
| Age - median (min,max) | N = 89 | 63 (38, 79) |
| Pretreatment Adjusted SAT - median (min,max) | N = 53 | 48.9 (16.8, 121) |
| Pretreatment Adjusted SKM - median (min,max) | N = 84 | 46.1 (29.3, 69.9) |
| Pretreatment Adjusted VAT - median (min,max) | N = 84 | 39.6 (0.9, 115.7) |
| Gender | F | 40 (45%) |
| | M | 49 (55%) |
| Race Ethnicity | Asian/Non-Hispanic | 3 (3%) |
| | Black/Non-Hispanic | 3 (3%) |
| | White/Hispanic | 6 (7%) |
| | White/Non-Hispanic | 77 (87%) |
| BMI | Normal | 36 (40%) |
| | Obese | 22 (25%) |
| | Overweight | 30 (34%) |
| | Underweight | 1 (1%) |
| Smoking History* | No | 34 (39%) |
| | Yes | 54 (61%) |
| Pretreatment Sarcopenia* | No | 38 (45%) |
| | Yes | 46 (55%) |

SKM=Skeletal Muscle; VAT=Visceral Adipose Tissue; SAT=Subcutaneous Adipose Tissue *Numbers do not sum to the total because 1 patient is missing Smoking History and 5 patients are missing Pretreatment Sarcopenia status.

The Kaplan-Meier Plot

There's nothing all that hard about adding figures to an R markdown file.

The hardest thing here is adding the numbers at risk below the plot, since this isn't built into the base survival package.

Fortunately it is an option in relatively recent survminer package, which is available from either CRAN or GitHub

The GitHub version now includes a cheatsheet.

The Code, Part 1

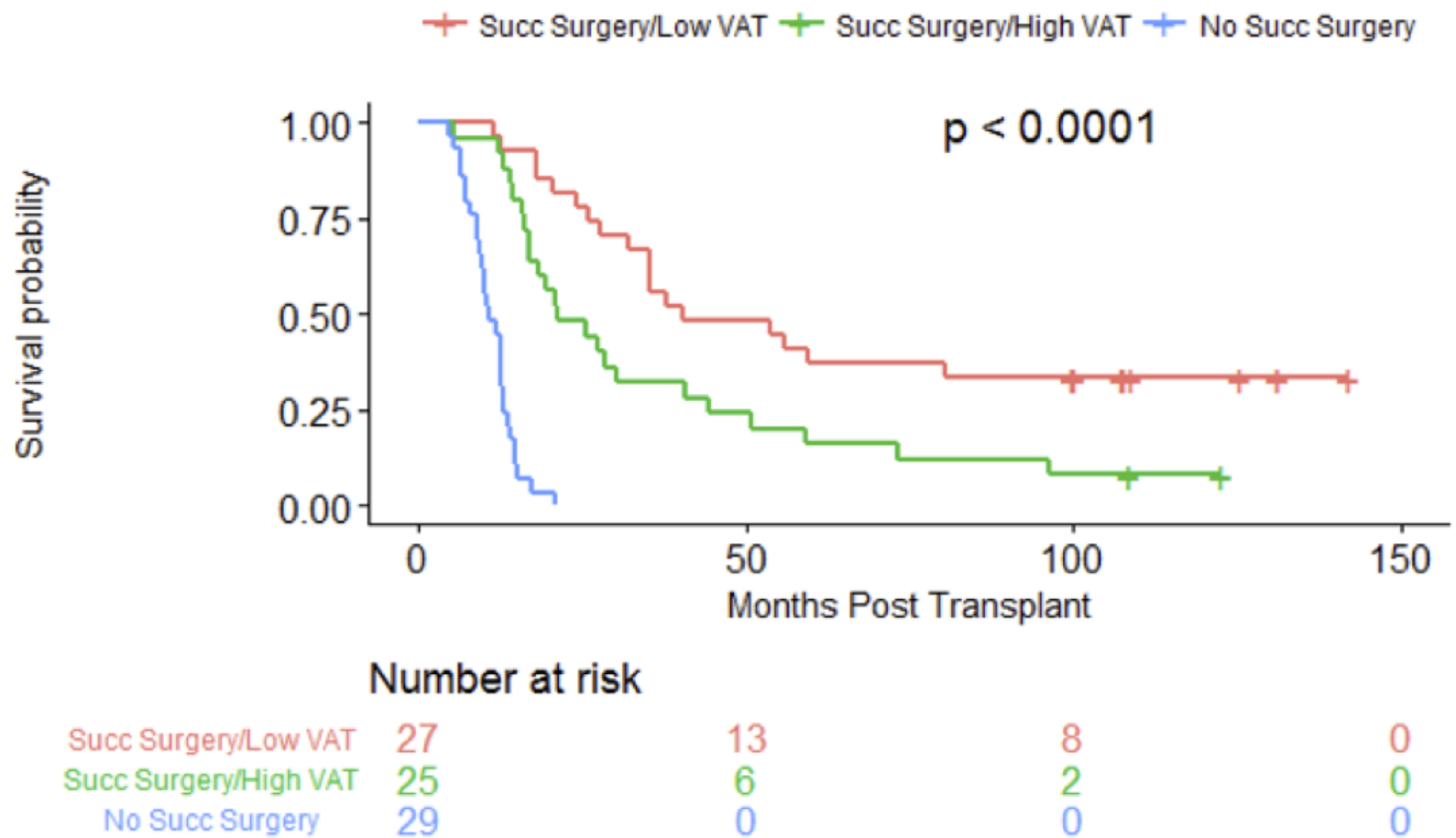
```
d1 <- relevel(factor(d$node), '2')
d2 <- relevel(d1, '3')
d$node <- d2
fit <- survfit(Surv(Overall_Survival,
  osevent)~node, data=d)
plot<-ggsurvplot(fit, data = d, size = 1, # change size
  #palette = c("#E7B800", "#2E9FDF"), # custom color palette
  #conf.int = TRUE, # Add confidence interval
  xlab = "Months Post Transplant",
  xlim = c(0,150),
  pval = TRUE, # Add p-value
  pval.coord = c(80,1),
  strata.order = 3:1,
```

The Code, Part 2

```
risk.table = TRUE, # Add risk table
risk.table.col = "strata", # Risk table color by
legend.title = NULL,
legend.labs = c("Succ Surgery/Low VAT",
  "Succ Surgery/High VAT", "No Succ Surgery"),
risk.table.height = 0.25, # change if many groups
ggtheme = theme_classic(), # Change ggplot2 theme
tables.theme = theme(axis.text.x=element_blank(),
axis.title.x = element_blank(),
axis.ticks.x = element_blank(),
axis.line = element_blank(),
axis.text.y = element_blank(),
axis.ticks.y = element_blank(),
axis.title.y = element_blank()) )
```

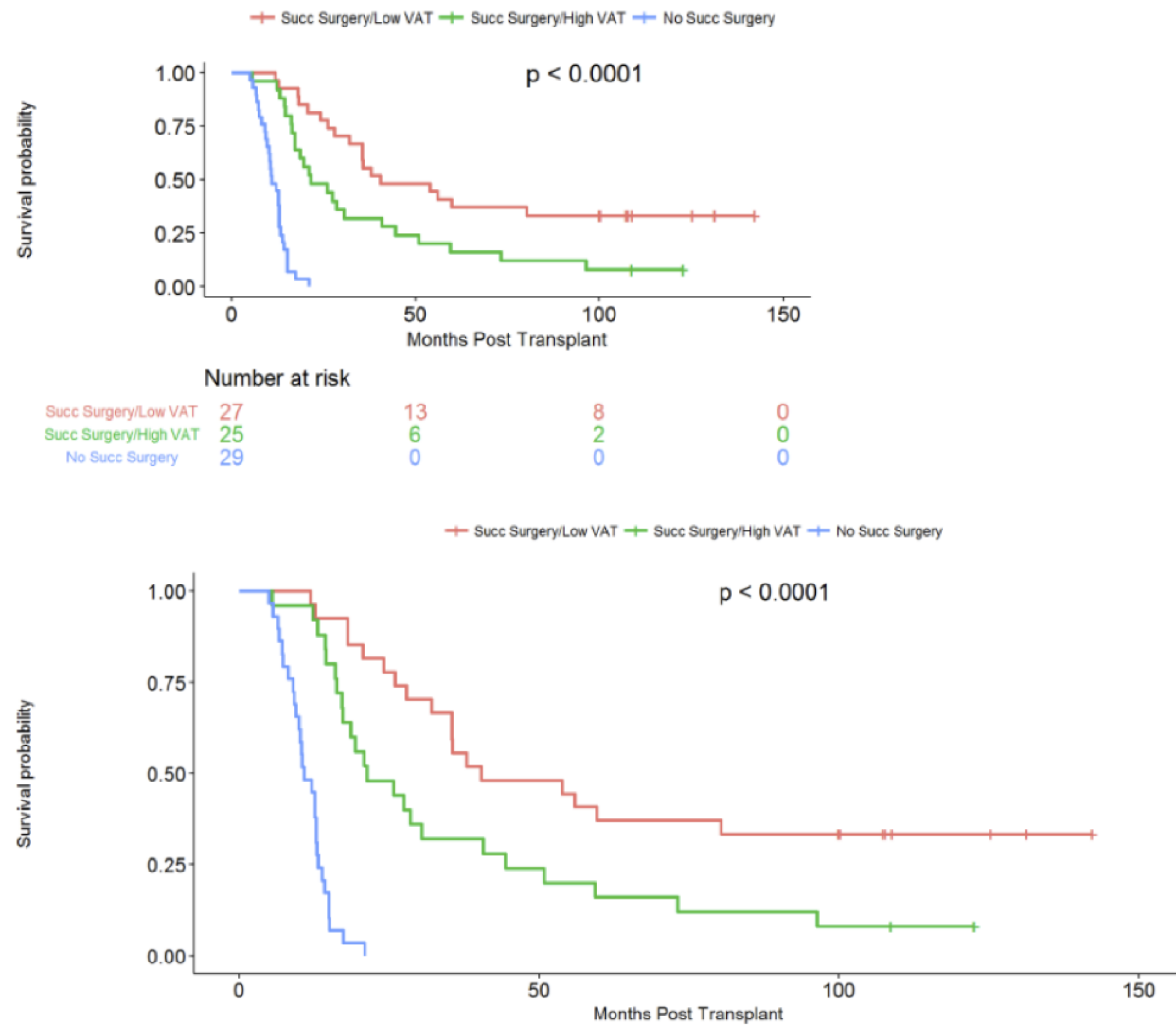
The Plot

Figure 1



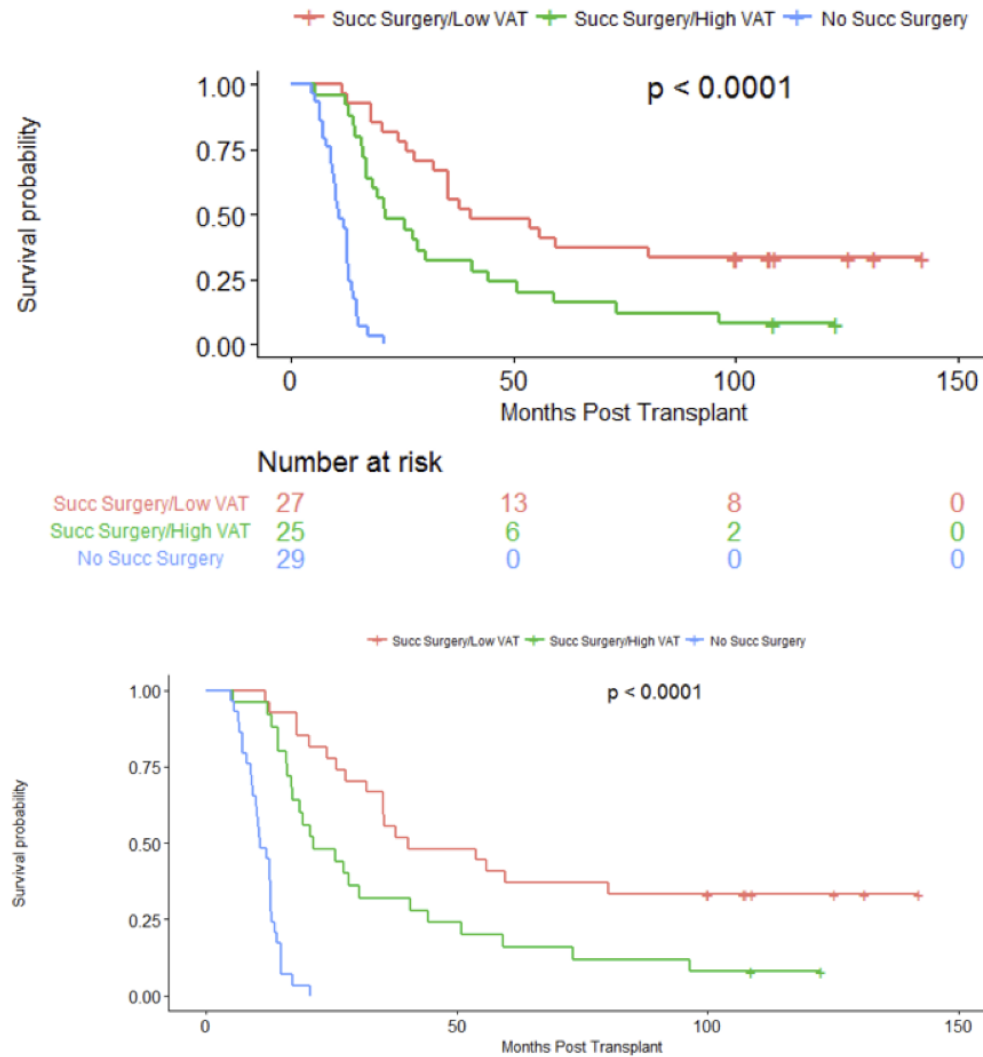
Width 6.5in or 10in in HTML

Figure 1



Width 6.5in or 10in in Word

Figure 1



Adding References

Specify a .bib file in the YAML

```
bibliography: becky.bib  
csl: nature-genetics.csl
```

Cite by reference keys in the markdown:

For additional survival models, Cox regression
[@cox72] was implemented

A .bib File

```
@article{cox72,  
  author = "David Roxbee Cox",  
  title = "Regression Models and Life-Tables",  
  journal = "Journal of the Royal Statistical Society",  
  year = "1972",  
  volume = "34",  
  number = "2",  
  pages = "187-220",  
  note = ""  
}
```

```
@article{kaplanMeier58,
```

What if we don't know the format?

knitcitation to the Rescue

In one file,

```
library(knitcitations)
cleanbib()

## get DOI for Kaplan and Meier
citep("10.2307/2281868")

## get DOI for Mehta and Patel
citep("10.1080/01621459.1983.10477989")

...

write.bibtex(file = "autorefs.bib")
```

The Pervasiveness of DOIs

JASA seems to have assigned DOIs to all of its papers by this point. JRSSB seems to have them going back to the late 90s (about 1997)

autorefs.bib:

```
@Article{Kaplan_1958,  
  doi = {10.1080/01621459.1958.10501452},  
  url = {https://doi.org/10.1080%2F01621459.1958.1},  
  year = {1958},  
  month = {jun},
```

But How Are References Formatted?

This is actually specified by a “citation style language” (.csl) file specified in the YAML above (see <http://docs.citationstyles.org/en/stable/primer.html> for more).

Such files exist for a few thousand journals, and can be examined at <https://www.zotero.org/styles>

Mousing over entries will show popups illustrating
(a) how citations will appear in the text, and
(b) how they will appear in the bibliography

The Output

References

1. Mehta, C. & Patel, N. R. A network algorithm for performing fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association* 78, 427–434 (1983).
 2. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53, 457–481 (1958).
 3. Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* 34, 187–220 (1972).
 4. Lausen, B., Sauerbrei, W. & Schumacher, M. Classification and regression trees (cart) used for the exploration of prognostic factors measured on different scales. in *Computational statistics* (eds. Dirschedl, P. & Ostermann, R.) 1483–1496 (Physica Verlag, 1994).
-

This Took a Lot of Time

We can't spend this much time for every analysis.

But by setting up a template which can be edited to produce the desired type of output, we increase the odds the approach will be used.

We can increase these odds yet further by bundling a set of templates into a package, and including vignettes describing these modifications.

Working on it now...

(SAS and STATA templates are being explored too)

Where Would the Reports be Stored?

In an internal instantiation of GitLab

This can be searched for PI, analyst, time, etc

Should allow for porting of analyses in repository form to GitHub at time of publication (subject to de-identifying the data)