Example Data Analysis Report
Reproducible Research Using **knitr** with LaTeX

Peter DeWitt

January 27, 2014

1 Introduction

This document serves as an example data analysis report generated using R for the analysis, LaTeX for the markup report writing language, and knitr to bring everything together. The data set used is a fictitious as was generated for example purposes only. The purpose of this document is to provide an example of reproducible research.

Reproducing this report These are the steps required for reproducing this report.

1. Install R and LaTeX on your computer.
2. Open R, install the knitr package if the package is not on your system.

```
install.packages("knitr", repos = "http://cran.rstudio.com")
```

3. Set the working directory in R to the same directory as this file exists in. Run the following commands in R,

```
library(knitr)
knit(input = "basicsLaTeX.Rnw")
```

4. The above R code will generate the file `basicLaTeX.tex`. Use your favorite LaTeX compiler to generate the .pdf, .eps, .ps, ... If you get a copy of the directory with the auxiliary files resulting from compiling the LaTeX, such as .aux, .log, . . . , please delete those files. Only the .Rnw file is truly necessary to reproduce this report.

2 Analysis Methods

Overall survival analysis was done using both Kaplan-Meier estimates and Cox proportional hazard regression models. The analysis was done in R version 3.0.2 (2013-09-25) [R Core Team, 2013] and the survival analysis was done using the `survival` package [Therneau, 2014]. Statistical significance was set at the 0.05 level.

3 Analysis and Results

The data set consisted of 19,039 records. A summary of the data set is presented in Table 1.

We are primarily interested in the differences in survival between patients with different Gleason scores. Figure 1 presents the Kaplan-Meier survival estimates by Gleason score. As expected, the higher the Gleason

Table 1: Data Set summary

	Overall		GS 7		GS 8		GS 9		GS 10	
	n	%	n	%	n	%	n	%	n	%
	19,039		12,986	68.21	3,670	19.28	2,139	11.23	244	1.28
Age (in years)										
[40,50)	3,051	16.03	2,145	16.52	544	14.82	323	15.10	39	15.98
[50,70)	5,945	31.23	4,259	32.80	1,005	27.38	608	28.42	73	29.92
[70,85]	10,043	52.75	6,582	50.69	2,121	57.79	1,208	56.47	132	54.10
Era										
Era 1	8,615	45.25	5,869	45.19	1,659	45.20	970	45.35	117	47.95
Era 2	10,424	54.75	7,117	54.81	2,011	54.80	1,169	54.65	127	52.05
PSA										
[0, 10) ng/ml	11,567	60.75	8,410	64.76	1,997	54.41	1,038	48.53	122	50.00
[10, 20) ng/ml	4,372	22.96	2,845	21.91	927	25.26	531	24.82	69	28.28
[20, Inf) ng/ml	3,100	16.28	1,731	13.33	746	20.33	570	26.65	53	21.72
T Stage										
T Stage 1	9,668	50.78	7,110	54.75	1,699	46.29	770	36.00	89	36.48
T Stage 2	8,189	43.01	5,360	41.28	1,657	45.15	1,065	49.79	107	43.85
T Stage 3/4	1,182	6.21	516	3.97	314	8.56	304	14.21	48	19.67
Observed Deaths										
	2,755		1,611		598		473		73	

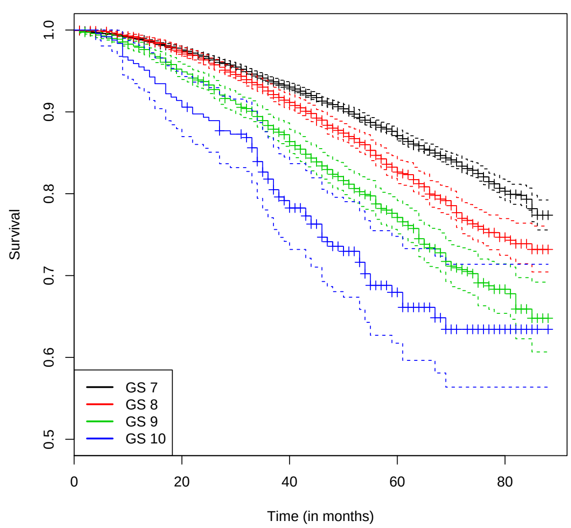

Figure 1: Kaplam-Meier Survival Curves

Table 2: Hazard ratios (HR) along with 95% confidence intervals (LCL, UCL) and p-values for testing if the hazard ratio is statistically different from 1 are presented in this table for both univariable and multivariable regression models for overall survival.

	Univariable Regressions				**Multivariable Regression**			
	HR	LCL	UCL	p-value	HR	LCL	UCL	p-value
Age								
[40,50)	Reference				Reference			
[50,70)	0.95	0.84	1.09	0.4807	0.96	0.84	1.10	0.5550
[70,85]	1.62	1.44	1.81	<0.0001	1.61	1.43	1.80	<0.0001
Era								
Era 1	Reference				Reference			
Era 2	0.83	0.76	0.90	<0.0001	0.84	0.77	0.92	<0.0001
T.Stage								
T Stage 1	Reference				Reference			
T Stage 2	1.19	1.10	1.29	<0.0001	1.12	1.03	1.21	0.0063
T Stage 3/4	1.54	1.34	1.77	<0.0001	1.24	1.07	1.43	0.0033
PSA								
[0, 10) ng/ml	Reference				Reference			
[10, 20) ng/ml	1.45	1.32	1.58	<0.0001	1.36	1.24	1.48	<0.0001
[20, Inf) ng/ml	1.62	1.47	1.78	<0.0001	1.50	1.36	1.66	<0.0001
Gleason								
GS 7	Reference				Reference			
GS 8	1.34	1.22	1.47	<0.0001	1.23	1.12	1.35	<0.0001
GS 9	1.92	1.73	2.12	<0.0001	1.73	1.55	1.91	<0.0001
GS 10	2.74	2.17	3.46	<0.0001	2.48	1.96	3.14	<0.0001

score, the worse the survival. It should also be noted that even after seven years of tracking patients the median survival time is not estimable. The lowest survival estimate is 63.43%.

Both univariable and multivariable Cox proportional hazard regression models were fitted for overall survival by the age, era of treatment, T stage, PSA, and Gleason score of the patient. Results for all the regression models are presented in Table 2.

The results of a univariable regression model indicated that Patients treated in Era 2 had statistically better survival than patients treated in Era 1, HR = 0.83 (95% CI: 0.76,0.90), and there was no appreciable difference in the hazard ratio found in the multivariable regression model, HR = 0.84 (95% CI: 0.77,0.92). As expected, as patients increase in age, T Stage increase, PSA increase, and Gleason score increases, the hazard also increases.

The hazard ratio between Gleason 8 and Gleason 7, from the multivariable Cox proportional hazard regression model, is HR = 1.23 (95% CI: 1.12,1.35). Further analysis of the pairwise comparisons of the hazards between all four Gleason scores can be provided upon request.

4 Conclusions

The conclusions section for a data analysis report would generally be used to summarize the results presented in Section 3, list any limitations to the study, and generate some discussion topics. Seeing how the purpose of *this* report was to show illustrate the use of `knitr`, the conclusions will focus on reproducible research.

Using `knitr` to write data analysis reports were the written report and the data analysis methods is a version of literate programming. When written well, the report are robust to changes in the data set, but more importantly, every element of the report is commented directly or contextually.

In addition to using **knitr**, a very powerful tool for authoring reports, both as a sole author, or as a collaboration, is to use version control software. I prefer **git**[1], but another viable option is subversion. RStudio has built-in features to working with either. Repository hosting on github.com or bitbucket.org are helpful, but on public servers (private repos are possible, but think about the physical location of the data storage). The git server software can be purchased and set up behind institutional firewalls.

References

[R Core Team, 2013] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Therneau, 2014] Therneau, T. (2014). *survival: Survival Analysis*. R package version 2.37-7.


```r
# for reproducibility, print out the session info for the packages, and
# versions of the packages, used to run the analysis and create this
# document.
print(sessionInfo(), local = FALSE)

## R version 3.0.2 (2013-09-25)
## Platform: x86_64-pc-linux-gnu (64-bit)
##
## attached base packages:
## [1] grid      splines   stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] gdata_2.13.2    Hmisc_3.13-0    Formula_1.1-1   lattice_0.20-24
##  [5] cluster_1.14.4  survival_2.37-7 knitr_1.5       vimcom_0.9-91
##  [9] setwidth_1.0-3  colorout_1.0-0
##
## loaded via a namespace (and not attached):
## [1] evaluate_0.5.1 formatR_0.10   gtools_3.1.1   highr_0.3
## [5] stringr_0.6.2  tools_3.0.2
```

[1] http://git-scm.com/