An Introduction to the Use of R for Clinical Research

Dimitris Rizopoulos

Department of Biostatistics, Erasmus Medical Center

d.rizopoulos@erasmusmc.nl

PSDM Event: Open Source Software in Clinical Research

June 19th, 2012

Outline

- What is R and how to obtain in it

- Features of R

- Using R

- R and clinical research

What is R?

- R is a free software environment for statistical computing and graphics.

  ▷ it was initiated in 1992 by Ross Ihaka and Robert Gentleman at University of Auckland, New Zealand

  ▷ in 1997 the R Core Team was established with renowned members of the statistical computing community

  ▷ nowadays, the R Core Team has grown and consists of about 20 members, experts in computing

- Free Software

  ▷ the source code is available

  ▷ users are allowed to modify and redistribute the code

How to Install R?

- Download R from the CRAN web site

  ▷ `http://cran.r-project.org`

  ▷ choose your platform, e.g., Windows, Linux

  ▷ e.g., for Windows: `Windows` → `base` → `Download R 2.15.0 for Windows`

  ▷ Install . . .

- Download R packages from the CRAN web site ⇒ within R

  ▷ Packages

  ▷ Install package(s) . . .

  ▷ make your choice(s)

  ▷ load the package using `library()` (note: install does not mean load)

Features of R

- Why R

  ▷ because is free

  ▷ it compiles and runs on a wide variety of UNIX platforms as well as Windows and MacOS

  ▷ R has extensive and powerful graphics & data manipulation capabilities

  ▷ it can easily interface with low-level programming languages, e.g., C/C++ or Fortran

  ▷ it can be easily extended via R packages

Features of R (cont'd)

- Disadvantages of R

  ▷ steep learning curve (some might say)

  ▷ output is not so nice looking (but there are some alternatives)
    * Sweave, odfWeave

  ▷ exporting output is more difficult

  ▷ cannot easily handle very very big data sets (depends on the installed RAM)
    * use 64bit OSs

  ▷ a lot of things are available but it is sometimes hard to find your way

  ▷ the quality of the available packages is greatly varying

Examples using R

- R is a command-based functional language

  ▷ write and execute commands

  ▷ use and define functions

- You may write the commands in the R console (Windows) or in a shell (Linux)

- Strongly advisable to use a suitable text editor – Some available options:

  ▷ Tinn-R (for Windows; `http://sciviews.org/Tinn-R/`)

  ▷ Rstudio (all major platforms; `http://www.rstudio.org/`)

  ▷ for more check `http://www.sciviews.org/_rgui/projects/Editors.html`

Examples using R (cont'd)

- R has very flexible and compact syntax

  ▷ Example: Calculate the coefficient of variation (sample std. dev. / sample mean), for blood pressure separately for males and females, in the age groups (20, 40) and (40, 60), and conditionally of being obese (BMI > 30) or not

```
with(BPdata,
    tapply(BP, list(cut(age, c(20, 40, 60)), sex, weight / height^2 > 30),
        function (x) sd(x)/mean(x))
```

Examples using R (cont'd)

▷ Example: Fit a linear model for blood pressure levels taking as explanatory variables the linear and quadratic effects of age the main effect of gender and their interaction

```
fm <- lm(BP ~ poly(age, 2) * sex, data = BPdata)

summary(fm) # parameter estimates, standard errors, etc.

plot(fm) # all basic residuals plots

fitted(fm) # extract fitted values

predict(fm, newdata) # make predictions for new patients
```

- Plotting

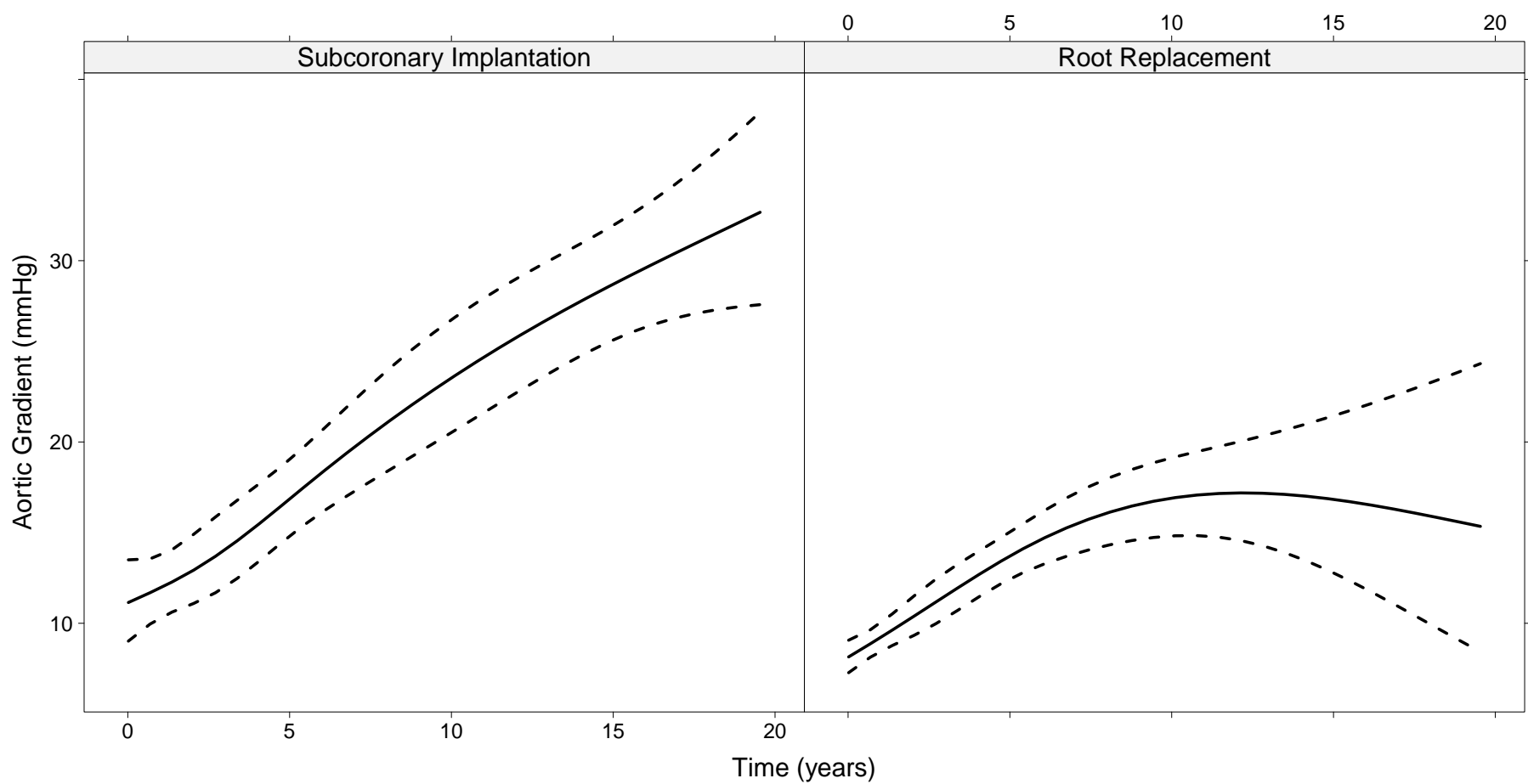 ▷ Predicted values with 95% CIs per operation type

```
xyplot(pred + low + upp ~ time | TypeOp, data = Preds,
    type = "l", col = "black", lty = c(1,2,2),
    ylab = "Aortic Gradient (mmHg)", xlab = "Time (years)")
```

Examples using R (cont'd)

Reporting Results in R : The Standard

- Communicating the results of a statistical analysis

  ▷ perform the analysis using your preferred statistical software

  ▷ results from this analysis constitute the basis for a statistical report

- Usually, this is a two-stage procedure,

  ▷ first do the analysis, and

  ▷ then write the report

- Statistician's hope: I won't have to change the analysis after I have finalized the report

  ▷ unfortunately, this is seldom the case

Reporting Results in R : An Alternative

• Embed the analysis into the report

  ▷ end up with only the report and data files

• Advantages

  ▷ reproducible reports

  ▷ dynamic reports

• Example: The client asks you to redo the whole analysis excluding some patients

  ▷ you just subset your original data and the report gets automatically updated!!

Dynamic Reports with Sweave

- What is Sweave

  ▷ Sweave is a tool that allows to embed the output of R code in LaTeX documents

- How it works

  ▷ your report file will contain both documentation parts (written in LaTeX) and code parts (written in R)

  ▷ the code is evaluated in R

  ▷ the results as plain output or tables and/or figures are embedded into a final `.tex` file

  ▷ you can then run `pdflatex` or `latex` to produce a `pdf` of your report

Dynamic Reports with Sweave (cont'd)

- Requirements:

  - ▷ if you know how to use R and LaTeX

    - * no need to learn something new
    - * Sweave ships directly with R
    - * it is relatively straightforward to start using it

  - ▷ if you do **not** know LaTeX

    - * alternative: `odfWeave` (open document format)

Dynamic Reports with Sweave (cont'd)

- Assuming basic R and LaTeX knowledge

- How does it work

  ▷ write your LaTeX file as usual, but with extension `.Rnw` instead of `.tex`
    e.g., `myfile.Rnw`

  ▷ the file will also contain R code segments *suitably separated* from LaTeX

  ▷ from R execute `Sweave("...\myfile.Rnw")`

  ▷ this will create `myfile.tex`

  ▷ run LaTeX to obtain your report

Dynamic Reports with Sweave (cont'd)

- How do we combine the R and LATEX source code

  ▷ using the Noweb syntax we separate between different segments (chunks) of source code, i.e.,

  * `<< 'options' >>=` denotes the start of an R code chunk
  * `@` denotes the start of a documentation LATEX chunk

- Basic options for code chunks

  ▷ `label`: an optional name for the chunk – useful for locating errors

  ▷ `echo`: if TRUE, the commands are included in the document

  ▷ `fig`: if TRUE, it includes the plot created in the code

Dynamic Reports with Sweave (cont'd)

- Basic options for code chunks

  ▷ `eval`: if `TRUE`, the R code is evaluated

  ▷ `results`:
    * if `hide`, all output is completely suppressed
    * if `tex`, the output is taken to be already proper LaTeX markup and included as is
    * if `verbatim`, the output of R commands is included in a verbatim-like R output environment

  ▷ ... (check `?RweaveLatex`)

Dynamic Reports with Sweave (cont'd)

- More info/material for Sweave available online:

  ▷ `http://www.stat.uni-muenchen.de/~leisch/Sweave/Sweave-manual.pdf`

  ▷ `http://www.stat.uni-muenchen.de/~leisch/Sweave/FAQ.html`

  ▷ `http://www.stat.umn.edu/~charlie/Sweave/`

  ▷ `http://www.biostat.jhsph.edu/~rpeng/ENAR2009/lecture-slides.pdf`

  ▷ `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/SweaveLatex/fhsweave.pdf`

  ▷ many more ⇒ Google it

Getting Help in R

- Within R

  ▷ `help.search("topic")` or `??"topic"` (depends on the installed packages)

  ▷ `RSiteSearch("topic")` (requires internet connection)

  ▷ `help()` or `?` invoke the on-line help file for the specified function

  ▷ checking the FAQ

- On the internet

  ▷ R-help (`https://stat.ethz.ch/mailman/listinfo/r-help` – mailing list)

  ▷ R-seek (`http://www.rseek.org` – Google-like searched engine)

  ▷ R-wiki (`http://rwiki.sciviews.org/doku.php`)

Getting Help in R (cont'd)

- On the internet

  ▷ CRAN Task Views (http://cran.r-project.org/web/views/ –
  categorization of packages)

  ▷ Crantastic (http://crantastic.org/ – categorization of packages + reviews)

  ▷ Equalis (http://www.equalis.com/forums/ – R forum)

  ▷ R4stats (http://www.r4stats.com/) – examples of basic R programs

Getting Help in R (cont'd)

- Intro with applications in statistics

  ▷ Dalgaard, P. (2008) *Introductory Statistics with R, 2nd Ed*. New York: Springer-Verlag. (moderate)

  ▷ Venables, W. and Ripley, B. (2002) *Modern Applied Statistics with S*. New York: Springer-Verlag. (advanced)

- Programming

  ▷ Venables, W. and Ripley, B. (2000) *S Programming*. New York: Springer-Verlag.

  ▷ Chambers, J. (2008) *Software for Data Analysis Programming with R*. New York: Springer-Verlag.

Getting Help in R (cont'd)

- Clinical research

  ▷ Peace, K. and Chen, D.-G. (2010) *Clinical Trial Data Analysis Using R*. Boca Raton: Chapman and Hall/CRC.

- More books that use R (or S) can be found at:
  `http://www.r-project.org/doc/bib/R-books.html`, or
  `http://www.r-project.org/doc/bib/R-jabref.html`

R For Clinical Research

- The **Is R Validated?** saga: There is the PERCEPTION that a certain three-lettered statistical analysis system is the "Gold Standard" and, worse, is perhaps the only one accepted by the FDA

  ▷ This is **not TRUE!**

- A key aspect of the CT regulatory framework is 21 CFR 11 with respect to digital signatures, audit trails, etc.

- Questions regarding the applicability of 21 CFR 11 to "stand-alone" statistical applications as opposed to databases that acquire, store and manage source electronic records

R For Clinical Research (cont'd)

- Most decision makers want to see documentation of compliance with applicable aspects of the regulations

- Efforts to create a guidance document for R began in earnest at useR! 2006 conference in Vienna

- "Working Group" began drafting a document with the goal of addressing key issues as they specifically pertain to R

  ▷ Marc Schwartz (Vice President, Biostatistics, MedNet Study Solutions)

  ▷ Frank Harrell (Chair at Dept. Biostatistics, Vanderbilt University School of Medicine)

  ▷ Tony Rossini (Group Head, Novartis Pharma AG)

R For Clinical Research

- Leverage existing information on development, version control, testing, maintenance, bug reporting/resolution, stable release cycles, updates, documentation, end user support, etc.

  ▷ received constructive criticism from multiple parties

- Document submitted to The R Foundation for approval on June 15, 2007

- Notified of approval by The R Foundation on July 27, 2007

- Available at: `http://www.r-project.org/doc/R-FDA.pdf`

R For Clinical Research

- Covers explicitly listed packages from **"Base R"** and the **"Recommended Packages"**

- Does **NOT** cover other CRAN and non-CRAN R packages

- Qualification and Validation

- Specifically addresses 21 CFR 11.10 (a-i) and 11.30 functional requirements

R For Clinical Research

- Changing from the standard to R is possible for clinical research

- **However,** it will require time ($\Rightarrow$ learning curve)

- Time is money! The relevant question is:

  > **How much money compared to the money payed annually for licences?**

Thank you for your attention!