**R:**

**Software for Statistical Modelling & Computing**

”R is a free software environment for statistical computing and graphics” (http://www.r-project.org)

”Bioconductor is a software project for the analysis of genomic data” (http://www.bioconductor.org)

* + Currently works as an expansion to R

R consists of a core and packages.Packages contain functions that are not available in the core.

For example, Bioconductor code is distributed as several dozen of packages for R.

* + Software packages
  + Metadata (annotation) packages

**Who is R?**

R is a language and environment for statistical computing and graphics. It is a [GNU project](http://www.gnu.org/) which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S.

**What Data can interact with R?**

R (and S) are accused of being slow, memory-hungry, and able

to handle only small data sets. This is completely true. Fortunately, computers are fast and have lots of memory. Data sets with a few tens of thousands of observations can be handled in 256Mb of memory, and quite large data sets with 1Gb of memory. Workstations with 32Gb or more to handle millions of observations are still expensive (but in a few years Moore’s Law should catch up). Tools for interfacing R with databases allow very large data sets, but this isn’t transparent to the user.

**Why should I use R for my work?**

R has many benefits over other statistical software packages. Its main benefit is that it is open source software. This means that anyone can help develop new packages and features. This allows cutting edge methods to come to R much faster than other software packages. This software can also be downloaded for free by anyone from various sites around the world.

**What is epidemiological1 research with R?**

Data analysis is very important in epidemiological1 research. The capacity ofcomputing facilities has been steadily increasing, moving state of the art

epidemiological studies along the same direction of computer advancement. Currently, there are many commercial statistical software packages widely used byepidemiologists around the world. For developed countries, the cost of software is not a major problem. For developing countries however, the real cost is often too high. Several researchers in developing countries thus eventually rely on a piratedcopy of the software.

1. **Epidemiology** is the science that **studies** the patterns, causes, and effects of health and disease conditions in defined populations. It is the cornerstone of public health, and informs policy decisions and evidence-based practice by identifying risk factors for disease and targets for preventive healthcare. Resource location : [Epidemiology - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/Epidemiology)

https://en.wikipedia.org/wiki/**Epidemiology**

Freely available software packages are limited in number and readiness of use.

**EpiInfo**, for example, is free and useful for data entry and simple data analysis. Advanced data analysts however find it too limited in many aspects. For example, it is not suitable for data manipulation for longitudinal studies. Its regression analysis facilities cannot cope with repeated measures and multi-level modelling. Thegraphing facilities are also limited. A relatively new and freely available software called **R** is promising. Supported by leading statistical experts worldwide, it has almost everything that anepidemiological data analyst needs. However, it is difficult to learn and to use compared with similar statistical packages for epidemiological data analysis such as **Stata**. The purpose of this book is therefore to bridge this gap by making **R** easy to learn for researchers from developing countries and also to promote its use.

My experience in epidemiological studies spans over twenty years with a specialfondness of teaching data analysis. Inspired by the spirit of the open-source software philosophy, I have spent a tremendous effort exploring the potential anduse of **R**. For four years, I have been developing an add-on package for **R** that allows new researchers to use the software with enjoyment.

### The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Among other things it has

* an effective data handling and storage facility,
* a suite of operators for calculations on arrays, in particular matrices,
* a large, coherent, integrated collection of intermediate tools for data analysis,
* graphical facilities for data analysis and display either directly at the computer or on hardcopy, and
* awell developed, simple and effective programming language (called ‘S’) which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed most of the system supplied functions are themselves written in the S language.)

**What R does and does not?**

Does :

* data handling and storage: numeric, textual
* matrix algebra
* hash tables and regular expressions
* high-level data analytic and statistical functions
* classes (“OO”)
* graphics
* programming language: loops, branching, subroutines

does not:

* is not a database, but connects to DBMSs
* has no graphical user interfaces, but connects to Java, TclTk
* language interpreter can be very slow, but allows to call own C/C++ code
* no spreadsheet view of data, but connects to Excel/MsOffice
* no professional / commercial support

**What sorts of things is R *good* at?**

* + there are very many statistical algorithms
  + there are very many machine learning algorithms
  + visualization
  + it is possible to write scripts that can be reused
  + R is a real computer language

**R supports many data technologies?**

* + XML,database integration,SOAP
* R interacts with other languages
  + C; FORTRAN; Perl; Python; Java
* R has good visualization capabilities
* R has a very active development environment
* R is largely platform independent
  + Unix; Windows; OSX

**What Data Analysis and Presentation R can perform?**

* The R distribution contains functionality for large number of statistical procedures.
  + linear and generalized linear models
  + nonlinear regression models
  + time series analysis
  + classical parametric and nonparametric tests
  + clustering
  + smoothing
* R also has a large set of functions which provide a flexible graphical environment for creating various kinds of data presentations.