CanReg analysis improvements

Morten Johannes ERVIK, Freddie BRAY

May 16, 2011

Contents

1 Introduction

R is a powerful open source freely available software package that could be coupled with CanReg to improve the analytical capabilites.

CanReg5 is an open source tool to input, store, check and analyse cancer registry data. It has modules to do data entry, quality control, consistency checks and basic analysis of the data.

2 What data does registries store in CanReg5

2.1 Data elements

Each case has (at least):

- Sex
- Incidence date
- Birth date (or age at the time of tumour)
- Coded address at the time of the tumour
- Topography, Morphology, Behaviour in ICD-O-3
- Most valid basis of diagnosis
- ICD-10 and ICCC
- Date of last contact
- Vital status
- Source info:
  - (Type of)
  - Number of
  - (Dates)

2.2 Population data

Each registry also have population data sets (denominators).

3 What we *do* with that data now

- Incidence tables (Per 100.000 per cancer group, age group, ASR, CR etc.)
- Number of cases (per cancer group, age group)
- Population pyramids
- Frequencies by year

Otherwise data needs to be exported to be analysed in other software packages.

4 What we *could* do with that data

4.1 Graphs

4.1.1 Bar charts by cancer/sex

- incidence tables
- number of cases

4.1.2 Time trends

- ASRs (world) over time
- Age specific rates over time
- Age specific rates over cohort

4.2 Quality indicators

4.2.1 Validity

- DCO%
- PSU%
- MV%
- Compared with other reg (CI5 IX)
- DCO% over time
  - potentially with graphs

4.2.2 Completeness

- Reference childhood incidence comparison
- Stability of rates over time by cancer/sex
  - potentially with graphs
- Age specific rates by cancer/sex
- Sources
  - number of sources per case
  - number of notifications per case

4.3 Geographic stuff, maps?

- Complicated as this needs to be set up for each registry.

5 What we could do if we linked it to other data

This is more for the future, but might be interesting...

- If linked to mortality data
  - M/I ratios as estimator of completeness

6 Technical aspects

6.1 Getting the data to R - and back

Basically two main ways to do it.

6.1.1 Export the data from CanReg5 to files readable by R and then call R in batch mode

Prefered mehtod - more dynamic and loosely coupled. Easier to potentially reuse R code later.

6.1.2 Use libraries in Java (a Java to R bridge) to call R functions directly

Alternative method we might want to look into.

6.2 Output file formats

Graphics should be output in a format that can be used by many. Possibly also given as a choice to the user. (Ref: http://www.stat.auckland.ac.nz/~paul/R/devices.html)

6.2.1 Portable Network Graphics - PNG

Very convenient for many things, but not scalable. R can export directly to this format.

6.2.2 Enhanced Meta File - EMF

Scalable. Good for Word etc. R can export directly to this format (at least as WMF). Only for Windows?

6.2.3 Scalable Vector Graphics - SVG

Scalable, open standard. Allows for direct editting and manipulation in programs such as Adobe (r) Illustrator or Inkscape. R can export directly to this format.

6.2.4 (Encapsulated) Post Script - (E)PS

Scalable, open standard, ready to print. Good for publications. R can export directly to this format.

6.2.5 Portable Document Format - PDF

Scalable, open standard, ready to print. R can export directly to this format.

6.2.6 Character Separated Values - CSV

It would be good to write the values used in any table to a CSV file as well for further use.

7 What now

We want to prioritize things that can produce figures and/or look better in R then using Java.

7.1 Priorities

1. Age-specific rates for major diagnosis groups - linear and logaritmic (two of the tables "not yet implemented" - see figure 1.)
   (a) Need to specify diagnostic groups. Dynamic? (User definable?)
2. ASR and number of cases in major diagnosis groups in single calendar years of observation (two of the tables "not yet implemented" - see figure 2.)
   (a) Need to specify diagnostic groups. Dynamic? (User definable?)

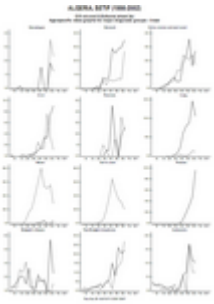

Figure 1: Majour groups - log

Figure 2: Cases per year

3. Quality indicators in major diagnostic groups (the last of the "not yet implemented")
   (a) Need to specify diagnostic groups. Dynamic? (User definable?)
   (b) Compared to another reference registry? In the region/world?
4. Time trends