Computer-aided estimation of comorbidity from electronic patient medical records

Erik Bülow[1,2], Ola Rolfson[1,2] and Szilárd Nemes[1,2]

[1]Swedish Hip Arthroplasty Register. [2]Department of Orthopedics, Institution of Clinical Sciences, University of Gothenburgm, Sweden.

`classifyr` *is a free and open source R package for fast patient classification of comorbidity, adverse events and more*

Introduction

It is common practice in register studies to use administrative data sets including medical codes (such as ICD or ATC) to classify risk factors and outcomes such as comorbidities and adverse events.

Some software exist to facilitate this process (Table 1) but most alternatives are either expensive, hard to use, or slow for large data sets.

The classifyr R package

We designed a software package called **`classifyr`**. The name reflects:

1. that it can be used for all sorts of classification where units (patients) are grouped into categories such as comorbidity and adverse events, and
2. that it is designed as an add-on package for R (Box 1).

Box 1: R

R is a free and open source, widely used, statistical software and programming language. Its has currently more than 10,000 free add on packages contributed freely by its users (*www.r-pkg.org).*

The package interface is centered around three objects:

1. a data set with unit (patient) data,
2. an additional data set with classification data such as diagnostic codes, and
3. a classification scheme linking individual (diagnostic) codes to (comorbidity/adverse events) categories by regular expressions (Box 2).

The user can either rely on default classification schemes included in the package (Box 3), or specify his or her own schemes following the same structure.

Box 2: Regular expressions

The traditional approach of patient classification based on medical or administrative data is to compare each code individually.

Example: Assume a patient with total hip arthroplasty had the following codes registered at hospital visits during the year preceding surgery:

T863B, L234X, N060, M058L, S451, K132, B901, M244C, D100, P271, E125, L529A, B348, G801B, Z541

If we want to identify any of these codes as "complicated diabetes" according to Elixhauser, we could compare this list of codes to codes identifying complicated diabetes:

E102, E102A, E102B, E102C, E102W, E102X, E103, E103A, E103B, E103C, E103D, E103E, E103F, E103W, E103X, E104, E104B, E104C, E104D, E104E, E104W, E104X, E105, E105A, E105B, E105W, E105X, E106, E106A, E106D, E106E, E106F, E106G, E106W, E107, E108, E112, E112A, E112B, E112C, E112W, E112X, E113, E113A, E113B, E113C, E113D, E113E, E113F, E113W, E113X, E114, E114B, E114C, E114D, E114E, E114W, E114X, E115, E115A, E115B, E115W, E115X, E116, E116A, E116D, E116E, E116F, E116G, E116W, E117, E118, E122, E123, E124, E125, E126, E127, E128, E132, E133, E134, E135, E136, E137, E138, E142, E143, E144, E145, E146, E147, E148

This can be seen as quite straight forward, but is slow for large sets of data. A faster approach is to reformulate the code list in a standardized and compact way known as a regular expression:

```
^(E1[0-4][23-8])
```

This is one way of many to increase computational speed used by the **`classifyr`** package.

Benchmark

To validate and benchmark the use of the package, we used 10,000 data points with:

1. patients data from the Swedish Hip Arthroplasty Register
2. a data set with ICD-10 codes from hospital visits from the Swedish National Patient Register, and
3. a comorbidity classification scheme for Elixhauser.

The goal was to identify comorbidities for patients during one year preceding surgery for total hip arthroplasty.

A first naïve attempt was made without any designated software package but with only base commands in R.

The same task was then performed with the help of R packages (Table 1). We excluded the **`icdcoder`** package since it did not work probably, and the **`medicalrisk`** package since it did not include ICD-10.

Results

The first implementation took almost an hour to run on a modern computer. The use of a designated R-package gave the same output but much faster (Table 2).

Table 2: Comparison of computational speed for classifying 10,000 patient codes by Elixhauser comorbidity.

Package	Time [sek]	Relative
classifyr	0.08	1
icd	6.39	80
comorbidities.icd10	35.81	454

Discussion

`classifyr` was 80 times faster than **`icd,`** and speed is important when calculating comorbidities based on large data sets.

If relying on ICD-9 however, **`icd`** could be preferred since it uses another faster technique for that version. The **`classifyr`** package only includes ICD-10-codes by default, since these are most commonly used in Sweden, although ICD-9 is still used in some other countries.

Conclusions

- The R-package **`classifyr`** is freely available and open source
- It is easily extendible to new classification schemes
- It is optimized for big data
- It is associated with shorter computing times compared to R packages with similar purpose

Box 3: Default classification schemes

- The Charlson comorbidity index based on ICD-10 with index specified by Charlson, Deyo, Romano, D'Hoore, Ghali and Quan (2 versions).
- The Elixhauser comorbidity index based on ICD-10.
- The comorbidity-polypharmacy score (CPS) based on ICD-10.
- The RxRiskV pharmacy based comorbidity index based on ATC.
- A verity of adverse events classifications after hip and knee arthroplasty based on ICD-10 and Swedish KVÅ-codes.

Table 1: Software for classifying patients by comorbidities.

For	Package	Author	URL	Classifications	Medical codes included
R	classifyr	Erik Bülow	www.github.com/eribul/classifyr	See Box 3	ICD-10, ATC, KVÅ and extendable
R	comorbidities.icd10	Max Gordon	www.github.com/gforge/comorbidities.icd10	Charlson, Elixhauser	ICD-9, ICD-10 (US/SWE)
R	icd	Jack Wasey	www.github.com/jackwasey/icd	Charlson, Elixhauser	ICD9, ICD10, ICD-10-CM
R	icdcoder	Wade Cooper	www.github.com/wtcooper/icdcoder	Charlson, Elixhauser	ICD9, ICD-9-CM, ICD10, ICD-10-CM
R	medicalrisk	Patrick McCormick	www.github.com/patrickmdnet/medicalrisk	Charlson, Elixhauser	ICD-9-CM
SAS	Elix. Comorb. Soft.	HCUP	hcup-us.ahrq.gov/tools_software.jsp	Elixhauser	ICD-9, ICD-10

SAHLGRENSKA ACADEMY
Swedish Hip Arthroplasty Register
www.shpr.se
erik.bulow@registercntrum.se

UNIVERSITY OF GOTHENBURG