Erik Bulow

- Swedish
- PhD student in medicine (orthopedics)
- Statistician
- Use micro data from official sources (Statistics Sweden)

Categorization

- Deterministic (defined mapping between codes and categories)
- Not statistical classification

TOC

- What’s the problem?
- How to solve it?
- The coder package.
- Example with comorbidity.
- Details about the package.

My problem

- Patients from one register
- Previous medical conditions (comorbidity) from another (much bigger) register.

How can we do this?

- SAS macro exist (doh …!)
- Possible with some basic but messy R-script (18 hours to run)
- R-package `comorbidities.icd10` (slow and complicated, not on CRAN)
- R-package `icd` (lacks some of the desired features, was initially slow and only contains specific coding schemes)
- Let’s make another package!

**coder**

- Separate functionality from classification schemes
- Classification using regular expression
- Some use of matrix algebra when possible
- Rely on `data.table`
- Avoid other dependencies (for internal IT system)
- Keep it simple
- Pipe-friendly (or use just one function)

Centered around three objects

1. Item data (the patients)
2. Code data for the same items (the patient register)
3. Classification scheme (based on reg-ex and with weighting schemes and additional conditions).

Codes

- Item data with many units (individuals)
- Code data with many instances (hospital visits) per item
  - Some time frame of relevance (one year before surgery or 30 days after)
- Many codes per code instance
  - Some of them relevant
  - Some of them with additional characteristics (main vs additional diagnostic codes)
- Relevant codes grouped by category
  - One code can belong to more than one category
- Weighting schemes for combined indices (total comorbidity burden based on individual comorbidities)

Example: Charlson Comorbidity

- Group of patients
- Link to National Patient Register for comorbidity data during the year before surgery
- Identify comorbidities based on the Charlson comorbidity classification scheme
- Calculate comorbity index due to one of the proposed standards
The coder package

Patients

Here by name, otherwise by personal identity numbers (PIN) for (almost) perfect linkage.

head(ex_people)
##     name    surgery
## 1 Beaver, Tristin 2016-09-11
## 2  Maestas, Lilibeth 2016-10-23
## 3       Jung, Derek 2017-02-20
## 4     Hayes, Kylihah 2016-12-31
## 5     el-Riaz, Aadam 2016-04-19
## 6  Sanchez, Dominique 2017-02-25

Medical records

head(ex_pardata)
##        id variable   code  code_date  hdia
## 1:  el-Haider, Ruwaid     hdia  T840F 2012-01-04  TRUE
## 2:      Roacho, Marla    bdia2   Z510 1993-01-23 FALSE
## 3:   el-Hariri,  Huda     hdia   A469 2008-10-07 FALSE
## 4: el-Parsa, Ghazaala     hdia  M085 2010-10-28 FALSE
## 5:  Martinez, Jonathan     hdia 989,98 2008-02-12 TRUE
## 6:  Martinez, Crystal     hdia H401C 2006-09-05 TRUE

Charlson comorbidity

charlson_icd10[c(5, 19), 1:5]
##        group                  regex charlson deyo_ramano dhoore
## 5  dementia  (^F0([0-3]|51)|G3(0|11))        1           1      1
## 19 AIDS/HIV  (^B2[0124])        6           1      1

Visualize classification scheme

visualize(charlson_icd10, c("congestive heart failure"))

Summarise classification scheme

suppressPackageStartupMessages(library(tidyverse))
summary(charlson_icd10) %>%
  mutate(codes = substr(codes, 1, 30))
##                               group   n                          codes
## 1                          AIDS/HIV  22 B200, B201, B202, B203, B204,
## 2 cerebrovascular disease  82 G450, G451, G452, G453, G454,
## 3  chronic pulmonary disease 57 I278, I279, J409, J410, J411,
## 4                  congestive heart failure  8 I099, I110, I130, I132, I500,
## 5              dementia  23 F000, F001, F002, F009, F010,
## 6           diabetes complication 71 E102, E102A, E102B, E102C, E10
## 7 without complication  55 E100, E100A, E100B, E100C, E10
## 8            hemiplegia or paraplegia 22 G841, G841, G841, G841A, G841B
## 9 malignancy 525 C000, C001, C002, C003, C004,
## 10 metastatic solid tumor  29 C770, C771, C772, C773, C774,
## 11                  mild liver disease  83 I120, I121, I122, I123, I124,
## 12 moderate or severe liver disease 11 I850, I859, I864, I862, I784,
## 13                  myocardial infarction  15 I210, I211, I212, I213, I214,
## 14             peptic ulcer disease  36 K250, K251, K252, K253, K254,
## 15 peripheral vascular disease  38 I700, I701, I702, I702A, I702C
## 16           renal disease  27 I120, I131, I133, I135, I137,
## 17         rheumatic disease  63 M050, M051, M052, M053, M058,

Add Charlson comorbidity

ex <-
ex_people %>%
categorize(
  ex_icd10,  # ... data from NPR
  "charlson_icd10",  # based on Charlson comorbidity from ICD-10
  id = "name",  # Identify id variable from SHAR
date = "surgery",  # Identify date variable to relate to
days = c(-365, -1),  # Only include comorbidity during this period
ind = "quan_original"  # Calculate comorbidity index with specified weights
)

Result

ex[, c(1, 7, 16, 21:22)]
##     name dementia malignancy AIDS/HIV quan_original
## 1 Abzari, Joseph       FALSE    TRUE       FALSE             2
## 2 Alexis, Dayveon       FALSE       FALSE       FALSE             0
More control

Function categorize takes care of everything. Alternative workflow:

```r
ex_people %>%
codify(ex_icd10, "name", "surgery", days = c(-365, -1)) %>%
classify(charlson_icd10) %>%
index("quan_original")
```
Schubert, Patrick                  Smith, John
  0                             0
Smith, Ruasha              Spoerl, John
  0                             2
Stephan, Scott             Suriwong, Sarah
  2                             1
Thompson, Stephen          Thwaites, Travis
  0                             0
Todacheene, Erminio        Tomlinson, Amanda
  1                             0
Tuilele-Britton, Janet        Valencia, Johnny
  2                             0
Vallie, Mitchelle           Wigo, Shi Hyung
  1                             2
Wilkerson, Teesa            Wood, Elijah
  0                             0

Classification schemes

- Comorbidity
  - Charlson (ICD-10)
  - Elixhaüser (ICD-10)
  - comorbidity-polypharmacy score (ICD-10)
  - RxRiskV (based on ATC codes) (ICD-10)
- Adverse events
  - After hip arthroplasty (ICD-10 and KVÅ)
  - After knee arthroplasty (ICD-10)
- Example scheme for car brands
- S3 class mechanism to make tailored classification schemes

Where to find it?

- Github: [www.github.com/eribul/coder](http://www.github.com/eribul/coder)
- Documented by pkgdown: [eribul.github.io/coder](http://eribul.github.io/coder)
  - Some vignettes exist and more are planned
- Plan to increase test coverage (now 65%)
- Plan to release on CRAN

So … how fast is it?

- It's quite fast!
- What took 18 hours before now takes 30 seconds!
- The newly updated icd package is even faster though, but is not as generic and.

Thanks!

Questions?