medicalrisk: Calculating risk and comorbidities from ICD-9-CM codes

Patrick McCormick *patrick.mccormick@alum.mit.edu*

2016-01-23

Introduction

The routines in the **medicalrisk** package (McCormick and Joseph 2015) are designed to help determine comorbidity and medical risk status of a given patient using several popular models published in the peer-reviewed literature.

Administrative healthcare data is frequently the only available source for determining individual risk of mortality when looking at thousands or millions of patient records. Medical chart abstraction just isn't feasible for projects of this scale.

In the United States, the records for every inpatient and outpatient encounter is reviewed by a qualified medical coder who assigns a set of diagnosis and procedural codes based on phrases within the medical record. The coding system currently in use is ICD-9-CM. ICD-9-CM is an adaptation of the venerable ICD-9 standard which was developed in 1978. The U.S. National Center for Health Statistics (NCHS) developed ICD-9-CM, which has been required for Medicare and Medicaid claims since 1979. ICD-9-CM is updated annually.

At some point, perhaps as soon as October 2015, ICD-10-CM codes will need to be used instead. It is likely that "dual coding" of claims in both sets will continue for some time.

In the meantime, there is a wealth of administrative data available within the ICD-9-CM diagnostic and procedural codes stored within US healthcare systems.

Working with ICD-9-CM Data

In order to demonstrate this package, this package includes data on 100 patients from the Vermont Uniform Hospital Discharge Data Set for 2011, Inpatient.

```
library(medicalrisk)
library(plyr)
data(vt_inp_sample)
x <- count(vt_inp_sample, c('id'))
cat("average count of ICD codes per patient is: ", mean(x$freq))


## average count of ICD codes per patient is:  11.52


y <- count(vt_inp_sample, c('icd9cm'))


library(knitr)
kable(head(y[order(-y$freq),], n=5), row.names=F,
    caption='Top 5 most popular ICD-9-CM codes in this dataset')
```

Top 5 most popular ICD-9-CM codes in this dataset	
icd9cm	**freq**
D4019	34
D53081	22
D2724	19
D3051	18
D25000	17

Within this package, ICD-9-CM codes are presented as a string where the first letter is "P" or "D" depending on whether the code is Procedure or Diagnosis. The rest of the code is present as a string of numbers. Periods are omitted. In the list above, the code "D4019" is diagnostic code 401.9 which corresponds to Hypertension.

Comorbidity Maps

The package includes a set of mapping functions that transform a list of ICD-9-CM codes into a comorbidity matrix:

- icd9cm_charlson_deyo
- icd9cm_charlson_romano

- icd9cm_charlson_quan
- icd9cm_elixhauser_ahrq37
- icd9cm_elixhauser_quan
- icd9cm_rcri

"Charlson" refers to the Charlson Comorbidity Index(Charlson et al. 1987).

The names "Deyo"(Deyo, Cherkin, and Ciol 1992), "Romano"(Romano, Roos, and Jollis 1993), and "Quan"(Quan et al. 2005) refer to the primary authors of different methods of determining Charlson comorbidities from ICD-9-CM codes.

"Elixhauser" refers to the Elixhauser comorbidity map, which is a more detailed list than Charlson. "AHRQ37" is an adapation of the AHRQ version 37 software(Agency for Healthcare Research & Quality 2013). "Quan" refers to the same paper by Quan mentioned above.

"RCRI" is the Revised Cardiac Risk Index(Lee et al. 1999) set of categories using a method published by Boersma(Boersma et al. 2005).

For example, the #5 ICD-9-CM code above is D25000, or "250.00", which is for "Diabetes Mellitus Unspecified Type". Here's what happens when that code is passed to a few of the mapping functions listed above:

```
kable(
  icd9cm_charlson_quan(c('D25000')))
```

	mi	chf	perivasc	cvd	dementia	chrnlung	rheum	ulcer	liver	dm	dmcx	para	renal	tumor	modli
D25000	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE

```
kable(
  icd9cm_elixhauser_ahrq37(c('D25000')))
```

	chf	arrhythmia	valve	pulmcirc	perivasc	htn	htncx	para	neuro	chrnlung	dm	dmcx	hypothy	renlfail
D25000	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE

```
kable(
  icd9cm_rcri(c('D25000')))
```

	chf	cvd	dm	ischemia	renlfail
D25000	FALSE	FALSE	TRUE	FALSE	FALSE

For each of these maps the "dm" column becomes TRUE.

The most efficient way to use these maps for a set of patients is to generate a single map for all ICD-9-CM codes in the set and then apply that map to each patient. Here's an example that generates a comorbidity matrix for the first five patients in the Vermont dataset:

```
cases <- vt_inp_sample[vt_inp_sample$id %in% 1:5, c('id','icd9cm')]
cases_with_cm <- merge(cases, icd9cm_charlson_quan(levels(cases$icd9cm)),
  by.x="icd9cm", by.y="row.names", all.x=TRUE)

# generate crude comorbidity summary for each patient
kable(
  ddply(cases_with_cm, .(id),
    function(x) { data.frame(lapply(x[,3:ncol(x)], any)) }),
  row.names=F)
```

id	mi	chf	perivasc	cvd	dementia	chrnlung	rheum	ulcer	liver	dm	dmcx	para	renal	tumor	modliver
1	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

The above process is encapsulated in a single function `generate_comorbidity_df`. This function also includes an optimization from Van Walraven that reduces `dmcx` to `dm` if the specific diabetic complication is separately coded.

```
kable(
  generate_comorbidity_df(cases, icd9mapfn=icd9cm_charlson_quan))
```

id	mi	chf	perivasc	cvd	dementia	chrnlung	rheum	ulcer	liver	dm	dmcx	para	renal	tumor	modliver
1	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

4 FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE FALSE
5 FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE FALSE

This function only considers each ICD-9-CM code once and then merges the resulting comorbidity flags together for each patient. This makes the function quite fast for large data sets.

Given appropriate arguments, the `generate_comorbidity_df` function will use the parallel backend provided by `foreach` to improve performance.

Comorbidity Index

It is common in the medical literature to see a set of comorbidities reduced to an index. When the Charlson Comorbidity Index was first published it had the following weights for each comorbidity:

```
data(charlson_weights_orig)
kable(
  t(charlson_weights_orig))
```

mi	chf	perivasc	cvd	dementia	chrnlung	rheum	ulcer	liver	dm	dmcx	para	renal	tumor	modliver	mets	aids
1	1	1	1	1	1	1	1	1	1	2	2	2	2	3	6	6

However, these weights have not stood the test of time. For example, the prognosis for HIV/AIDS has dramatically improved.
The medicalrisk package offers the revised Charlson weights developed by Schneeweiss(Schneeweiss et al. 2003):

```
data(charlson_weights)
kable(
  t(charlson_weights))
```

mi	chf	perivasc	cvd	dementia	chrnlung	rheum	ulcer	liver	dm	dmcx	para	renal	tumor	modliver	mets	aids
1	2	1	1	3	2	0	0	2	1	2	1	3	2	4	6	4

The `generate_charlson_index_df` function will sum the weights for each patient to generate a final index:

```
kable(
  generate_charlson_index_df(generate_comorbidity_df(cases)), row.names=F)
```

id	index
1	2
2	0
3	0
4	4
5	2

Risk Stratification Index

The Risk Stratification Index uses ICD-9-CM codes to determine four risk estimates:

- 1 Year Mortality
- 30 Day Mortality
- In-Hospital Mortality
- 30 Day Length of Stay

The author of the paper (Sessler) published SPSS code to perform the calculation. The medicalrisk implements the RSi calculation using a method based on the SPSS code.

```
ddply(cases, .(id), function(x) { icd9cm_sessler_rsi(x$icd9cm) } )
```

```
##   id rsi_1yrpod rsi_30dlos rsi_30dpod rsi_inhosp
## 1  1 -2.0186043  0.1560323  -1.699242 -1.8483037
## 2  2 -4.1423990  0.8927947  -3.802495 -3.5425015
## 3  3 -2.6265277  0.8311247  -2.910939 -2.8607594
## 4  4 -0.7984382  0.3357922  -1.551285 -0.2669842
## 5  5  2.5803930 -1.7904270   2.455086  1.7615180
```

Conclusion

The medicalrisk package can be used to generate risk data from ICD-9-CM codes in large datasets. The

above discussion describes basic use of the package. There are some additional helper functions not described above which are included in the per function documentation.

The aim of this package is to include future medical risk estimation procedures as they are published in the literature.

References

Agency for Healthcare Research & Quality. 2013. "HCUP Comorbidity Software. Healthcare Cost and Utilization Project (HCUP). [Version 3.7]." http://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp.

Boersma, Eric, Miklos D Kertai, Olaf Schouten, Jeroen J Bax, Peter Noordzij, Ewout W Steyerberg, Arend F L Schinkel, et al. 2005. "Perioperative cardiovascular mortality in noncardiac surgery: validation of the Lee cardiac risk index." *The American Journal of Medicine* 118 (10): 1134–41. doi:10.1016/j.amjmed.2005.01.064.

Charlson, M E, P Pompei, K L Ales, and C R MacKenzie. 1987. "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation." *Journal of Chronic Diseases* 40 (5): 373–83. http://www.ncbi.nlm.nih.gov/pubmed/3558716.

Deyo, R A, D C Cherkin, and M A Ciol. 1992. "Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases." *Journal of Clinical Epidemiology* 45 (6): 613–9. http://www.ncbi.nlm.nih.gov/pubmed/1607900.

Lee, T H, E R Marcantonio, C M Mangione, E J Thomas, C A Polanczyk, E F Cook, D J Sugarbaker, et al. 1999. "Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery." *Circulation* 100 (10): 1043–9. http://www.ncbi.nlm.nih.gov/pubmed/10477528.

McCormick, Patrick, and Thomas Joseph. 2015. *Medicalrisk: Medical Risk and Comorbidity Tools for ICD-9-CM Data*.

Quan, Hude, Vijaya Sundararajan, Patricia Halfon, Andrew Fong, Bernard Burnand, Jean-Christophe Luthi, L Duncan Saunders, Cynthia A Beck, Thomas E Feasby, and William A Ghali. 2005. "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data." *Medical Care* 43 (11): 1130–9. http://www.ncbi.nlm.nih.gov/pubmed/16224307.

Romano, P S, L L Roos, and J G Jollis. 1993. "Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives." *Journal of Clinical Epidemiology* 46 (10): 1075–9; discussion 1081–90. http://www.ncbi.nlm.nih.gov/pubmed/8410092.

Schneeweiss, Sebastian, Philip S Wang, Jerry Avorn, and Robert J Glynn. 2003. "Improved comorbidity adjustment for predicting mortality in Medicare populations." *Health Services Research* 38 (4): 1103–20. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1360935/.