**Intro**

There are many illnesses and diseases known to man. How do the various stakeholders in the medical science industry classify the same illness? The illness will need to be coded in a standardized manner to aid in fair reimbursements and concise reporting of diseases. The International Classification of Diseases (ICD) provides this uniform coding system. The ICD [*“is the standard diagnostic tool for epidemiology, health management and clinical purposes.”*](http://www.icd-code.org/about). *(There is a more detailed coding system known as the Systematized Nomenclature of Medicine — Clinical Terms (SNOMED-CT) but it will not be covered in this post.)*

The ICD has currently 11 versions. At this point of time, countries and researchers are using either ICD-9 or ICD-10, with those using ICD-9 gradually transiting to ICD-10. ICD-11 has yet to be adopted in clinical practice.

R has a package, icd, which deals with both ICD-9 and ICD-10. The package also includes built in functions to conduct common calculations involving ICD such as Hierarchical Condition Codes and Charlson and Van Walraven score. We will use the icd package to help explain ICD-9 and ICD-10 and do some analysis on an external dataset.

The ICD is a hierarchical based classification. There is a total of 4 levels:

1. chapter
2. sub-chapter
3. major. Each major has a 3\_digital identifier with a character length of three
4. descriptor, long\_desc. Each descriptor has an identifier code with a character length from three to five.

library(tidyverse)

library(icd)

theme\_set(theme\_light())

# Level 1-3

icd9cm\_hierarchy %>% select(chapter, sub\_chapter, major, three\_digit ) %>% head(10)

## chapter sub\_chapter

## 1 Infectious And Parasitic Diseases Intestinal Infectious Diseases

## 2 Infectious And Parasitic Diseases Intestinal Infectious Diseases

## 3 Infectious And Parasitic Diseases Intestinal Infectious Diseases

## 4 Infectious And Parasitic Diseases Intestinal Infectious Diseases

## 5 Infectious And Parasitic Diseases Intestinal Infectious Diseases

## 6 Infectious And Parasitic Diseases Intestinal Infectious Diseases

## 7 Infectious And Parasitic Diseases Intestinal Infectious Diseases

## 8 Infectious And Parasitic Diseases Intestinal Infectious Diseases

## 9 Infectious And Parasitic Diseases Intestinal Infectious Diseases

## 10 Infectious And Parasitic Diseases Intestinal Infectious Diseases

## major three\_digit

## 1 Cholera 001

## 2 Cholera 001

## 3 Cholera 001

## 4 Cholera 001

## 5 Typhoid and paratyphoid fevers 002

## 6 Typhoid and paratyphoid fevers 002

## 7 Typhoid and paratyphoid fevers 002

## 8 Typhoid and paratyphoid fevers 002

## 9 Typhoid and paratyphoid fevers 002

## 10 Typhoid and paratyphoid fevers 002

# Level 3-4

icd9cm\_hierarchy %>% select(major, three\_digit, long\_desc, code) %>% head(10)

## major three\_digit

## 1 Cholera 001

## 2 Cholera 001

## 3 Cholera 001

## 4 Cholera 001

## 5 Typhoid and paratyphoid fevers 002

## 6 Typhoid and paratyphoid fevers 002

## 7 Typhoid and paratyphoid fevers 002

## 8 Typhoid and paratyphoid fevers 002

## 9 Typhoid and paratyphoid fevers 002

## 10 Typhoid and paratyphoid fevers 002

## long\_desc code

## 1 Cholera 001

## 2 Cholera due to vibrio cholerae 0010

## 3 Cholera due to vibrio cholerae el tor 0011

## 4 Cholera, unspecified 0019

## 5 Typhoid and paratyphoid fevers 002

## 6 Typhoid fever 0020

## 7 Paratyphoid fever A 0021

## 8 Paratyphoid fever B 0022

## 9 Paratyphoid fever C 0023

## 10 Paratyphoid fever, unspecified 0029

We can see the subordinate codes of the three\_digit identifier with the function, children.

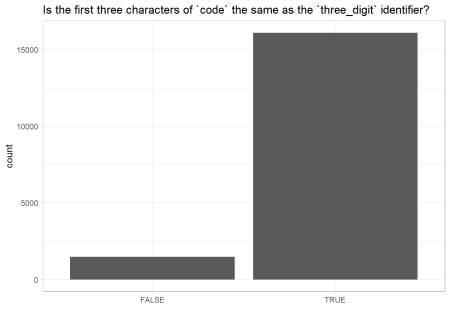
children("001")

## [1] "001" "0010" "0011" "0019"

Beware that in some instances the first three characters of codes are not the same as the three\_digit identifiers.

icd9cm\_hierarchy %>% mutate(first\_3\_char\_of\_code=substr(three\_digit, 1,3),

same=first\_3\_char\_of\_code==three\_digit) %>% ggplot(aes(same)) + geom\_bar()+ labs(x="", title= "Is the first three characters of `code` the same as the `three\_digit` identifier?")



Let’s examine which codes are these. Looks like codes beginning with “E” resulted in the mismatch.

icd9cm\_hierarchy %>% mutate(first\_3\_char\_of\_code=substr(three\_digit, 1,3),

same=first\_3\_char\_of\_code==three\_digit) %>% filter(same=="FALSE") %>% select(code, first\_3\_char\_of\_code, three\_digit) %>% sample\_n(10)

## code first\_3\_char\_of\_code three\_digit

## 90 E0129 E01 E012

## 380 E828 E82 E828

## 6 E0009 E00 E000

## 1360 E9830 E98 E983

## 1011 E9284 E92 E928

## 1085 E9353 E93 E935

## 1143 E9422 E94 E942

## 1117 E9389 E93 E938

## 1024 E9298 E92 E929

## 456 E8359 E83 E835

**Difference between ICD-9 and ICD-10**

**Breath and depth**

Now that we understand the structure of ICD. Let’s understand the difference between ICD-9 and ICD-10. ICD-10 has more chapters and more permutations and combinations of subordinate members than ICD-9. Thus, ICD-10 is a longer dataset than ICD-9.

cbind(ICD9=nrow(icd9cm\_hierarchy), ICD10=nrow(icd10cm2019)) %>% as\_tibble()

## # A tibble: 1 x 2

## ICD9 ICD10

##

## 1 17561 94444

**Coding**

Majority of ICD-9 uses numeric values for the first character for the three\_digit identifier (and therefore also for its code).

substr( icd9cm\_hierarchy$three\_digit, 1,1) %>% unique()

## [1] "0" "1" "2" "3" "4" "5" "6" "7" "8" "9" "V" "E"

Whereas ICD-10 uses all alphabets for the first character.

substr( icd10cm2019$three\_digit, 1,1) %>% unique() #https://stackoverflow.com/questions/33199203/r-how-to-display-the-first-n-characters-from-a-string-of-words

## [1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" "M" "N" "O" "P" "Q"

## [18] "R" "S" "T" "V" "W" "X" "Y" "Z"

I will be referring to ICD-9 for the rest of the post.

**code format**

code can be expressed in two ways:

1. Short format which has been used in all the above examples. It has a character length from three to five. The first three characters of code are the same as the 3\_digitalidentifier on most occasions. The mismatch occurs when the code begins with the letter “E”.
2. Decimal format. A handful of healthcare databases and research datasets adopt this format. code in this format have three characters on the left side of the decimal point which are the same as the three\_digit identifier. At the most two characters on the right side of the decimal point (e.g. “250.33”). However, due to formatting of electronic medical records or exporting the code to Excel, the code may be truncated. For instance, zeros before a non- zero numeric character will be dropped off (e.g. “004.11” -> “4.11” ). Zeros after a non-zero numeric character on the right side of the decimal point also will be dropped off (e.g. “250.50”-> “250.5”).

**Inspecting for data entry errors**

Data entry is susceptible to errors considering the code format and the magnitude of permutations and combinations of code. The icd package has two functions to identify data entry errors.

**Validation of code appearance**

is\_valid will help to determine if the code looks correct

is\_valid("123.456") #max of 2 char of R side of decimal point

## [1] FALSE

is\_valid("045l") #l is an invalid character

## [1] FALSE

is\_valid("099.17", short\_code = T) #expecting `code` to be short format and not decimal format

## [1] FALSE

is\_valid("099.17", short\_code = F) #plausible `code` in decimal format

## [1] TRUE

**Legitimate definition behind code**

codes which appear valid may not be not have any underpinning meaning. is\_defined helps to determine if the code can be defined.

as.icd9cm("089") %>% #as.icd9cm informs is\_defined which ICD version you are referring to

is\_defined()

## [1] FALSE

[The code 088 and 090 exists but 089 does not exist.](http://www.icd9data.com/2015/Volume1/001-139/default.htm)

**Application**

After completing a crash course on the concepts of ICD, let’s see how the package can help us with our data wrangling. We will be using a [dataset on hospital admission of individuals with diabetes](https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008).

diabetic<- read\_csv("diabetic\_data.csv") %>% select(primary=diag\_1, secondary=diag\_2)%>% #only using primary and secondary diagnosis for this exercise

gather(primary, secondary, key = "diagnosis", value= "code") #longer tidy format

**Exploring and cleaning the data**

**What format are the codes in ?**

The codes are formatted in the decimal form.

diabetic %>% select(diagnosis) %>% str\_detect(".")

## [1] TRUE

**Are there NA values?**

There are no NA values.

diabetic %>% map\_dbl(~sum(is.na(.x)))

## diagnosis code

## 0 0

However, by physically viewing the dataset, there are observations recorded as “?”. “?” suggests unknown or missing values. We’ll coerce “?” values into NA

diabetic<-diabetic %>% mutate(code=ifelse(code=="?", NA, code))

**Providing the disease name**

The codes allow encoding of diseases to be more convenient but render it less comprehensible. We will extract the name of the diseases from major, the disease types from sub-chapter and the disease class from chapter.

**Converting into short format**

The ICD dictionary code is in the short form while the code in the dataset is in the decimal form. I will need to convert the format of code in the dataset from the decimal form to the short type.

diabetic<-diabetic %>% mutate(code= decimal\_to\_short(code))

**Extracting the names**

# shorten `chapter` name to range of `three\_digit` identifier

icd9cm\_hierarchy$chapter<-fct\_recode( icd9cm\_hierarchy$chapter,

`001-139`="Infectious And Parasitic Diseases",

`140-239`= "Neoplasms",

`240-279`= "Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders",

`280-289`= "Diseases Of The Blood And Blood-Forming Organs",

`290-319`= "Mental Disorders",

`320-389 `= "Diseases Of The Nervous System And Sense Organs",

`390-459`= "Diseases Of The Circulatory System",

`460-519`= "Diseases Of The Respiratory System",

`520-579`="Diseases Of The Digestive System",

`580-629`="Diseases Of The Genitourinary System",

`630-679`= "Complications Of Pregnancy, Childbirth, And The Puerperium",

`680-709`="Diseases Of The Skin And Subcutaneous Tissue",

`710-739`= "Diseases Of The Musculoskeletal System And Connective Tissue",

`740-759`="Congenital Anomalies",

`760-779`="Certain Conditions Originating In The Perinatal Period",

`780-799`= "Symptoms, Signs, And Ill-Defined Conditions",

`800-999`="Injury And Poisoning",

`V01-V91`="Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services",

`E000-E999`="Supplementary Classification Of External Causes Of Injury And Poisoning")

# merge dataset with ICD dictionary to extract disease names, types, classes

diabetic\_names<-left\_join(diabetic, icd9cm\_hierarchy,

by=c("code"="code")) %>% #making the arg explicit

select(diagnosis, disease\_name=major, disease\_type=sub\_chapter, disease\_class=chapter)

head(diabetic\_names,10)

## # A tibble: 10 x 4

## diagnosis disease\_name disease\_type disease\_class

##

## 1 primary Diabetes mellitus Diseases Of Other E~ 240-279

## 2 primary Disorders of fluid, electr~ Other Metabolic And~ 240-279

## 3 primary Other current conditions i~ Complications Mainl~ 630-679

## 4 primary Intestinal infections due ~ Intestinal Infectio~ 001-139

## 5 primary Secondary malignant neopla~ Malignant Neoplasm ~ 140-239

## 6 primary Other forms of chronic isc~ Ischemic Heart Dise~ 390-459

## 7 primary Other forms of chronic isc~ Ischemic Heart Dise~ 390-459

## 8 primary Heart failure Other Forms Of Hear~ 390-459

## 9 primary Other rheumatic heart dise~ Chronic Rheumatic H~ 390-459

## 10 primary Occlusion of cerebral arte~ Cerebrovascular Dis~ 390-459

**Summary of Diagnosis**

**Disease names**

The most common disease name for primary diagnosis is diabetes. Not surprised given that the dataset is about individuals with diabetes. The most common class of disease is cardio- vascular (390-459) which relates to the heart and the blood circulatory system

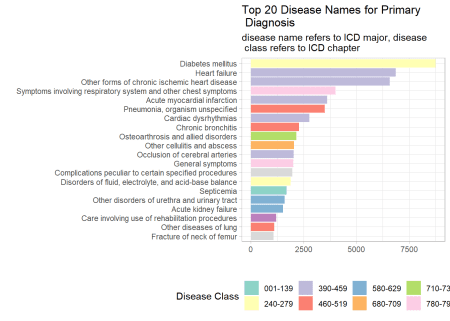
#top 20 primary diagnosis

diabetic\_names %>% filter(diagnosis=="primary") %>% count( disease\_name, disease\_class,sort = T) %>% top\_n(20) %>%

mutate(disease\_name=fct\_reorder(disease\_name,n)) %>% ggplot(aes(disease\_name, n, fill=disease\_class))+ geom\_col() + coord\_flip() +

theme(legend.position="bottom") + guides(fill=guide\_legend(title= "Disease Class", ncol = 5)) + # legend based on aes fill, split into 4 col as legend broken off page. change legend title

labs(x="", y="", title = "Top 20 Disease Names for Primary \n Diagnosis", subtitle = "disease name refers to ICD major, disease \n class refers to ICD chapter ") + scale\_fill\_brewer(palette = "Set3")



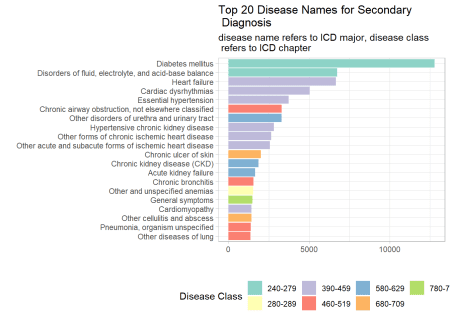
Similarly, the most common disease for secondary diagnosis is diabetes and the most common disease class is cardio-vascular. However, the number of disease class for secondary diagnosis is fewer than primary diagnosis.

diabetic\_names %>% filter(diagnosis=="secondary") %>% count( disease\_name, disease\_class,sort = T) %>% top\_n(20) %>%

mutate(disease\_name=fct\_reorder(disease\_name,n)) %>% ggplot(aes(disease\_name, n, fill=disease\_class))+ geom\_col() + coord\_flip() +

theme(legend.position="bottom") + guides(fill=guide\_legend(title= "Disease Class", ncol = 5))+

labs(x="", y="", title = "Top 20 Disease Names for Secondary \n Diagnosis ", subtitle = "disease name refers to ICD major, disease class \n refers to ICD chapter")+ scale\_fill\_brewer(palette = "Set3")



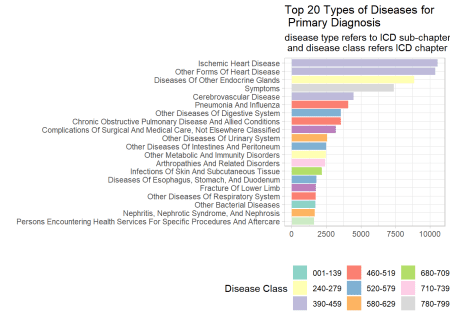
**Disease types**

The disease type for diabetes is “Diseases of Other Endocrine Glands” and knowing that diabetes is the most common disease name for primary diagnosis, let’s see if “Diseases of Other Endocrine Glands” will also be the most common disease type.

diabetic\_names %>% filter(diagnosis=="primary") %>% count( disease\_type, disease\_class,sort = T) %>% top\_n(20) %>%

mutate(disease\_type=fct\_reorder(disease\_type,n)) %>% ggplot(aes(disease\_type, n, fill=disease\_class))+ geom\_col() + coord\_flip() +

theme(legend.position="bottom") + guides(fill=guide\_legend(title= "Disease Class", ncol = 4)) + labs(x="", y="", title = "Top 20 Types of Diseases for \n Primary Diagnosis", subtitle = "disease type refers to ICD sub-chapter \n and disease class refers ICD chapter") + scale\_fill\_brewer(palette = "Set3")



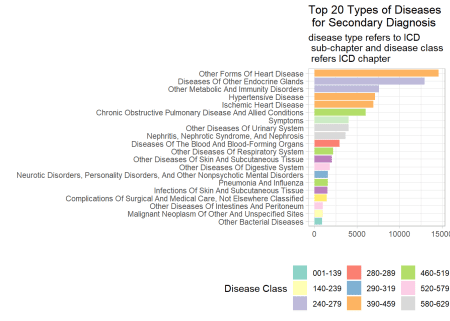
When we collapsed disease names for primary diagnosis to their superordinate, disease types, the most common disease type is “Ischemic Heart Diseases”. Though, “Diseases of Other Endocrine Glands” is the third most common disease type.

Let’s see if this is the same for secondary diagnosis.

diabetic\_names %>% filter(diagnosis=="secondary") %>% count( disease\_type, disease\_class,sort = T) %>% top\_n(20) %>%

mutate(disease\_type=fct\_reorder(disease\_type,n)) %>% ggplot(aes(disease\_type, n, fill=disease\_class))+ geom\_col() + coord\_flip() +

theme(legend.position="bottom") + guides(fill=guide\_legend(title= "Disease Class", ncol = 5)) + labs(x="", y="", title = "Top 20 Types of Diseases \n for Secondary Diagnosis", subtitle = "disease type refers to ICD \n sub-chapter and disease class \n refers ICD chapter") + scale\_fill\_brewer(palette = "Set3")



“Diseases of Other Endocrine Glands” is still not the most common disease type though it moved up a spot. “Ischemic Heart Diseases” is now the 5th most common disease type.

**To sum up**

In this post, we learned about the International Classification of Diseases which is an invaluable reference for various stakeholders in healthcare to have a uniform code for illnesses. The icd package was introduced to aid in the processing of datasets with ICD codes.