Obesity Comorbidity Analysis

Chloe Li

11/22/2016

Contents

NOTE

This is an example to show my coding skills in R as well as data processing skills. This contains process before further advanced analysis.

Introduction

This report documents data analytics process for obesity comorbidity analysis on data retrieved from NCBI.

*NOTE: Select **Hide Code** to hide all R codes.*

The objective of tis assignment is to conduct a comorbidity analysis for obesity using NCBI's PubMed database. The data was selected based on a date range of **2000 to 2012**, with a major MeSH descriptor of **"obesity"** and semantic types of **"Disease or Syndrome"**.

```
#preparation

#clear environment if needed
rm(list = ls())

#set working directory
setwd('~/Documents/Dev_dataScienceProjects/BrownUniv/')
```

```r
#install/reuqire libraries
if (!'pacman' %in% installed.packages()){
  install.packages('pacman')
}

pacman::p_load("ggplot2","dplyr", "rentrez","knitr","RCurl","plyr","tidyr","data.table")

#rentrez is the package in R provides an interface to the NCBI's EUtils API
```

Methods

1. Data retrieval from PubMed.
2. Data processing on PubMed data
3. MeSH descriptors XML parsing and processing
4. Mapping MeSH descriptors to PubMed data
5. Data analysis

Data Retrieval

In order to retreive data based on date range and keywords, a correct search query should be formulated. The package of R named "rentrez" allows R users to pull data from NCBI.

PubMed Search

entrez_db_summary() shows summary of a certain database, in this case, **PubMed** database is used.

- Database summary: **PubMed**

```r
#show the database summary information
entrez_db_summary("pubmed")
```

```
##  DbName: pubmed
##  MenuName: PubMed
##  Description: PubMed bibliographic record
##  DbBuild: Build161219-0807m.1
##  Count: 26791660
##  LastUpdate: 2016/12/19 13:08
```

entrez_db_searchable() shows searchable fields under a certain database, users can decide which keywords should be put under which searchable field.

- List of **PubMed**'s *searchable fields* (below only shows **MAJR**)

```r
#Searchable fields for database 'pubmed'
searchField <- entrez_db_searchable("pubmed")

searchField$MAJR
```

```
##  Name: MAJR
##  FullName: MeSH Major Topic
##  Description: MeSH terms of major importance to publication
```

```
##  TermCount: 539668
##  IsDate: N
##  IsNumerical: N
##  SingleToken: Y
##  Hierarchy: Y
##  IsHidden: N
```

- In order to form the PubMed/MEDLINE search for articles between 2000 and 2012 with obesity indicated as the major MeSH descriptor, elements below should be included in the query:
  - ***MAJR*** - *MeSH terms of major importance to publication* is used as searchable fields
  - **2000/01/01:2012/12/31[PDAT]** is date of publication/[PDAT]
  - **pubmed** is the database

Below shows the search result, which contains 59515 records and a web_history object; Here is the summary of search result:

```
query <- "obesity[MAJR] AND 2000/01/01:2012/12/31[PDAT]"

obesity_search <- entrez_search(db="pubmed",
                          query,
                          retmode = "xml",
                          use_history = TRUE,
                          retmax=60000)

kable(summary(obesity_search))
```

	Length	Class	Mode
ids	59517	-none-	character
count	1	-none-	numeric
retmax	1	-none-	numeric
QueryTranslation	1	-none-	character
file	1	XMLInternalDocument	externalptr
web_history	2	web_history	list

- A *web histroy* was returned as well, which NCBI created for users who deals with very large queries. With this *web history*, all records based on the search query were stored on NCBI server waiting for further usage.

```
obesity_search$web_history
```

```
## Web history object (QueryKey = 1, WebEnv = NCID_1_11684...)
```

- **The formulation of the PubMed search is shown below:**

```
obesity_search$QueryTranslation
```

```
## [1] "\"obesity\"[MeSH Major Topic] AND 2000/01/01[PDAT] : 2012/12/31[PDAT]"
```

In order to make sure this search result match with the number of records on NCBI. The same searching criteria were used on NCBI website manually, and the result is shown below:

Fetch Data

Based on the search query, all records that matched with the criteria were found and all IDs were returned. Those IDs or the web history could be used to fetch all records from NCBI. However, NCBI only allows users

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<eSearchResult>
   <Count>59515</Count>
   <RetMax>59515</RetMax>
   <RetStart>0</RetStart>
  ▼<IdList>
     <Id>25548090</Id>
     <Id>24694474</Id>
     <Id>24533500</Id>
     <Id>24444198</Id>
     <Id>24315250</Id>
     <Id>24218202</Id>
     <Id>24170641</Id>
     <Id>24170597</Id>
     <Id>24020252</Id>
     <Id>24008025</Id>
     <Id>24002404</Id>
     <Id>24002403</Id>
```


Figure 1:

to pull 10,000 records at once. Therefore, a web history needs to be used to save the search result in the server so that all 59,515 records can be pulled from multiple batches.

- Obtained PubMed/MEDLINE records (in MEDLINE or XML format) for the formulated search using NCBI E-Utilities
- Extracted "pmid","authors","year","articletitle","meshHeadings" from all records and stored values into a data frame.

```
#fetch all records from the formulated search
source('./batchFetch_Fun.R')

#this returns a data frame with extracted information from all records
dt_sum <- batchFetch(obesity_search)
#function code will be shown at the end of this report

#save dt_sum in local so don't need to run every time
#write.csv(dt_sum, "obesity_SumTable.csv")
#dt_sum <- read.csv("obesity_SumTable.csv")
#dt_sum$X <- NULL
```

- Sample of the data frame

```
kable(dt_sum[1:3,])
```

pmid	authors
25548090	So ES|Yoo KS
24694474	Sakurai M|Nakamura K|Miura K|Yoshita K|Takamura T|Nagasawa SY|Morikawa Y|Ishizaki M|Kido T|Naruse Y
24533500	Drenowatz C|Kobel S|Kettner S|Kesztyüs D|Steinacker JM

Data Processing

To prepare the data for further analysis, two variables *authors* and *meshHeadings* should be re formatted since there are multiple values in one record for both variables. *authors* and *meshHeadings* variables should be unlisted so that each row represents an unique record of meshHeadings and authors with pmid, publication year and article title.

```
source('./unlistDT_Fun.R')

#return a data frame of 5 variables with each cell contains only single value
dt_tidy <- unlistDT(dt_sum)
#write.csv(dt_tidy, "tidy_obesitySearch.csv")
```

Mapping

- Goal: create mapping of MeSH descriptors to semantic types using the MeSH Vocabulary file (desc2015.xml) and identify descriptors with type **Disease or Syndrome**
- Link to MeSH Descriptors XML file: _ftp://nlmpubs.nlm.nih.gov/online/mesh/2015/desc2015.xml_

Dictionary

Parse MeSH descriptors XML file:

```
#download XML file - 2015 MeSH descriptor
MeSHdescriptor <- XML::xmlParse("ftp://nlmpubs.nlm.nih.gov/online/mesh/2015/desc2015.xml")
#meshList <- xmlToList(MeSHdescriptor)
```

- In order to map MeSH descriptors to semantic types, **DescriptorName** and **SemanticTypeName** were extracted from the XML file above for the purpose of creating a dictionary, which later was used to link the MeSH descriptor from PubMed search.

```
#extract value from xml
source("./ExtractDict_Fun.R")
dt_MeshDict <- ExtractDict(MeSHdescriptor)
```

Summary of the dictionary before cleaning:

```
kable(dt_MeshDict[1:5,])
```

meshHeadings	SemanticTypeName
Calcimycin|Anti-Bacterial Agents|Calcium Ionophores	Organic Chemical|Antibiotic|Organic Chemic
Temefos|Insecticides	Organophosphorus Compound|Hazardous or
Abattoirs	Manufactured Object
Abbreviations as Topic	Intellectual Product|Intellectual Product
Abdomen|Abdomen|Abdominal Injuries|Abdomen|Radiography, Abdominal	Body Location or Region

The above record contains variables that have multiple values per row/record since one meshHeading might associate with multiple SemanticTypeName. In order to get the data ready for mapping, the data frame needs to be transformed and each cell should be unlisted so that each cell contains only single value.

```
#unlist variables (tidy)
source("./MeshDict_Fun.R")
tidy_Dict <- MeshDict(dt_MeshDict)

#reorder
tidy_Dict <- dplyr::arrange(tidy_Dict, SemanticTypeName)
```

Sample of the MeSH-SemanticType dictionary

```
kable(tidy_Dict[1:10,])
```

meshHeadings	SemanticTypeName
Airway Remodeling	Acquired Abnormality
Amputation Stumps	Acquired Abnormality
Amyloidosis	Acquired Abnormality
Aneurysm, Dissecting	Acquired Abnormality
Ankylosis	Acquired Abnormality

meshHeadings	SemanticTypeName
Anti-Ulcer Agents	Acquired Abnormality
Aphakia, Postcataract	Acquired Abnormality
Arachnoid Cysts	Acquired Abnormality
Arcus Senilis	Acquired Abnormality
Biofilms	Acquired Abnormality

Mapping

```
#save the dataframe in case
dt_PubMed <- dt_tidy

#merge/VLOOKUP to map the semantictype to dt_PubMed
dt_Mapped <- merge(dt_PubMed, tidy_Dict, by="meshHeadings",all.x=TRUE)

#rearrange columns and rows
dt_Mapped <- dplyr::select(dt_Mapped, pmid,meshHeadings,SemanticTypeName,articletitle,authors,year)
dt_Mapped <- dplyr::arrange(dt_Mapped, SemanticTypeName)
```

Sample record of mapped PubMed search:

```
kable(dt_Mapped[1:5,])
```

pmid	meshHeadings	SemanticTypeName	articletitle
21453798	Airway Remodeling	Acquired Abnormality	Impact of obesity on airway and lung parenchyma remodeling in
18313627	Amyloidosis	Acquired Abnormality	Heart failure with preserved ejection fraction: hypertension, diab
17608272	Amyloidosis	Acquired Abnormality	Biliopancreatic diversion in a renal transplant patient.
17608272	Amyloidosis	Acquired Abnormality	Biliopancreatic diversion in a renal transplant patient.
19825338	Amyloidosis	Acquired Abnormality	Renal AA amyloidosis secondary to morbid obesity?

Data Processing

- Remove meshHeadings == "Obesity"
- Select only SemanticTypeName == 'Disease or Syndrome'
- Data is ready for analysis:

```
#remove meshHeadings that are 'Obesity'
dt_filter <- dplyr::filter(dt_Mapped, meshHeadings != 'Obesity')
#select only 'Disease or Syndrome' as SemanticTypeName
dt_filter <- dplyr::filter(dt_filter, SemanticTypeName == 'Disease or Syndrome')
#check if neccessary
#str(dt_filter)
kable(dt_filter[1:5,])
```

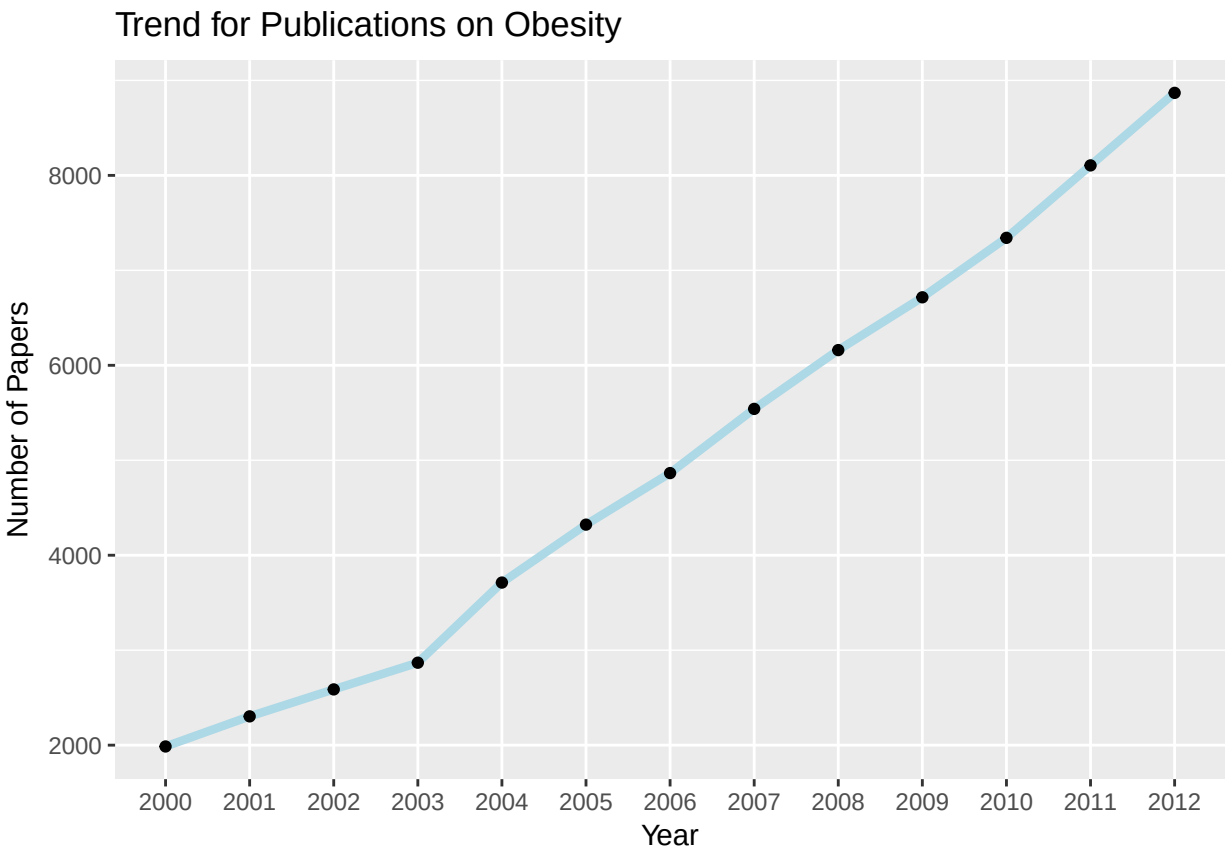
pmid	meshHeadings	SemanticTypeName	articletitle
21829000	22q11 Deletion Syndrome	Disease or Syndrome	Auxological evaluation in patients with a 22q1
20098742	3-Oxo-5-alpha-Steroid 4-Dehydrogenase	Disease or Syndrome	Effects of proportions of dietary macronutrien
10650936	3-Oxo-5-alpha-Steroid 4-Dehydrogenase	Disease or Syndrome	Understanding the role of glucocorticoids in o
15550507	3-Oxo-5-alpha-Steroid 4-Dehydrogenase	Disease or Syndrome	Reduced adipose glucocorticoid reactivation a
19308646	Abdominal Abscess	Disease or Syndrome	Late perforation and abscess formation at the

Analysis

Goal: Use a statistical or modeling approach to identify and to rank comorbidities for obesity based on publications that share MeSH descriptors for multiple disease or syndrome.

Trend

- Let's take a look at the trend on publications about obesity from 2000-2012:



As the line chart above shown, the number of publications with **Obesity** as major MeSH topic was increasing over the period of 2000 to 2012. Over time, the trend held consistent increase rate but from 2003 to 2004, it seemed that there was a large increase on the number of pulications on obesity.

Ranking

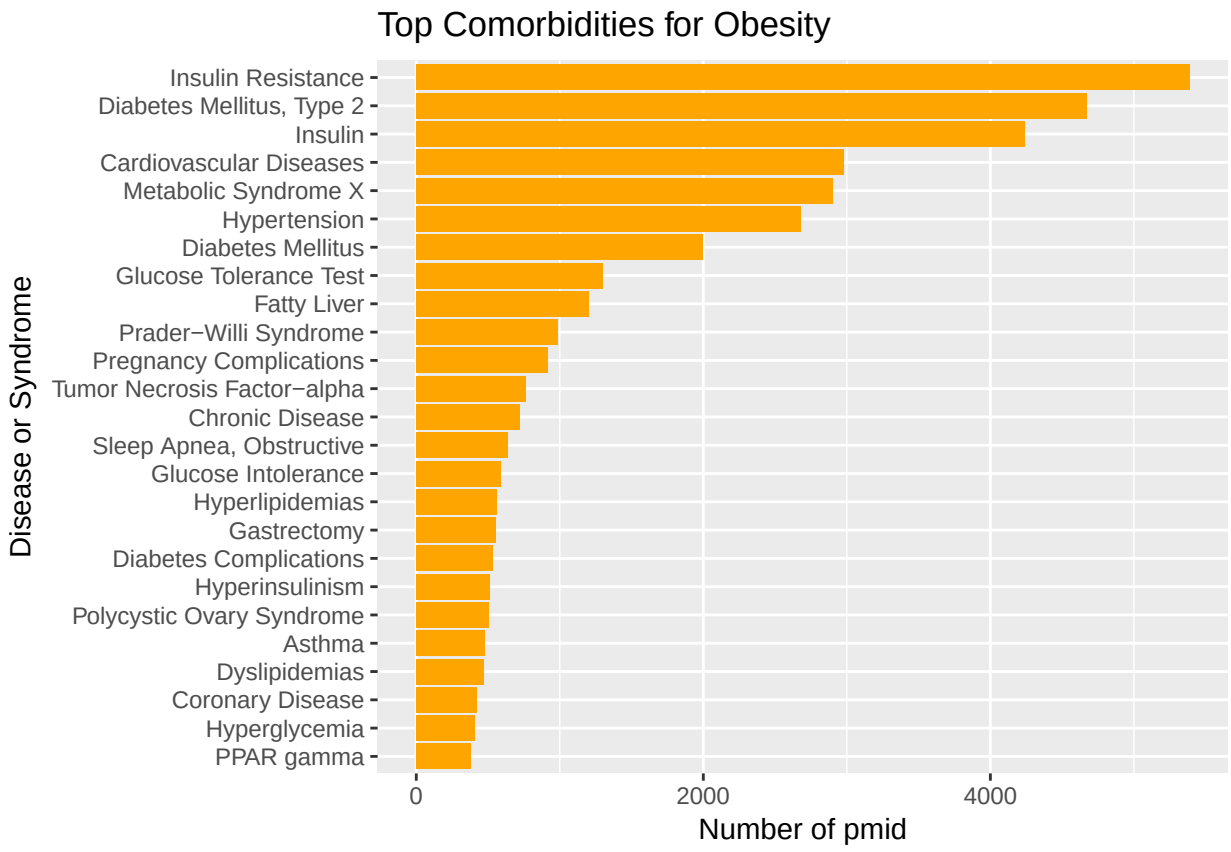
- Here is a sample record of unique **pmid** count on **meshHeadings**:

```
#group by meshHeadings and count number records for each MeSH
meSH_count <- setDT(dt_filter)[, .(count = uniqueN(pmid)), by = meshHeadings]
meSH_count <- as.data.frame(meSH_count)
#rarrange
meSH_count <- dplyr::arrange(meSH_count, -count)
meSH_count <- dplyr::select(meSH_count, meshHeadings,Count = count)

kable(meSH_count[1:10,])
```

meshHeadings	Count
Insulin Resistance	5387
Diabetes Mellitus, Type 2	4669
Insulin	4238
Life Style	3048
Cardiovascular Diseases	2976
Feeding Behavior	2939
Metabolic Syndrome X	2900
Hypertension	2676
Diabetes Mellitus	1995
Health Behavior	1549

- Histogram of top 25 disease or syndrome:

```
#subset the top 25 records
Top25 <- meSH_count[1:29,]
Top25 <- dplyr::filter(Top25, meshHeadings != "Life Style"& meshHeadings != "Health Behavior" & meshHeac


ggplot(Top25) +
    geom_bar(aes(x=reorder(meshHeadings, Count),y=Count),stat="identity",fill = "orange") +
                                ggtitle("Top Comorbidities for Obesity") +
                                xlab("Disease or Syndrome") +
                                ylab("Number of pmid") +
                                coord_flip()#+
```



Conclusions

The most common complication of obesity is **Insulin Resistance** syndrome while the second most common complication is **Diabetes Mellitus, Type 2**. Among top most comorbidities of obesity, most of diseases or syndromes fall under the category of metabolic syndrome. Moreover, according to the trend of number of publications on obesity, it was certain that many researches and discussions were on obesity.