Extract ICD code inclusion & exclusion criteria from tables in Gonzalez-Izquierdo et al. 2010, and Schnitzer et al 2011, and compare them

Jan Savinc

04 June, 2019

Contents

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

```
## -- Attaching packages ------------------------------------------------------------------------------
## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.5.2
```

```r
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.5.2
```

Intro

TODO: write a brief introduction based on the previous introduction of schnitzer 2011 below

Rationale for loading complete listing of ICD codes

The reason for importing the complete listing of ICD codes is to assist in extracting the ICD codes used as inclusion/exclusion criteria in previous literature, specifically where ranges of codes were defined, such as "X85 - Y03, Y08 - Y09"; not all codes in the X85 - X99 range are valid codes, for example, and by importing the ocmplete listing, we can generate a range of valid codes. These will be used as predictors in the analysis in the CHASe project.

Loading ICD codes

The UK uses unmodified WHO-published ICD codes (unlike the US where they use the - CM modification, or Australia, where they use -AM). ICD-9 was used until 1996, and ICD-10 since then.

UK codes

These were previously imported from the UK Biobank data, and the processed data is loaded:

```r
icd9_codes_UK <- read_csv("./processed_ICD_codes/master_icd9_code_list_UK(WHO).csv")
```

```
## Parsed with column specification:
## cols(
##   code = col_character(),
##   meaning = col_character(),
##   node_id = col_double(),
##   parent_id = col_double(),
##   selectable = col_character(),
##   code_decimal = col_character()
## )
```

```
icd10_codes_UK <- read_csv("./processed_ICD_codes/master_icd10_code_list_UK(WHO).csv")
```

```
## Parsed with column specification:
## cols(
##   code = col_character(),
##   meaning = col_character(),
##   node_id = col_double(),
##   parent_id = col_double(),
##   selectable = col_character(),
##   code_decimal = col_character()
## )
```

USA (CM) codes

These are included in the *icd* package for R.

```
library(icd)
```

```
## Warning: package 'icd' was built under R version 3.5.2
```

```
## Loading required package: icd.data
```

```
## Warning: package 'icd.data' was built under R version 3.5.2
```

General Equivalence Mappings for the CM versions of ICD

The latest GEM files were downloaded from TODO: source

```
gem_icd9cm <-
  read_fwf("./raw/2018_I9gem.txt",
           fwf_cols(
             source = c(1, 5),
             target = c(7, 13),
             approximate=c(15,15),
             no_map=c(16,16),
             combination=c(17,17),
             scenario=c(18,18),
             choice_list=c(19,19)
             )
           )
```

```
## Parsed with column specification:
## cols(
##   source = col_character(),
##   target = col_character(),
##   approximate = col_double(),
##   no_map = col_double(),
##   combination = col_double(),
##   scenario = col_double(),
##   choice_list = col_double()
## )
```

```
gem_icd10cm <-
  read_fwf("./raw/2018_I10gem.txt",
           fwf_cols(
             source = c(1, 7),
             target = c(9, 13),
             approximate=c(15,15),
             no_map=c(16,16),
             combination=c(17,17),
             scenario=c(18,18),
             choice_list=c(19,19)
             )
           )
```

```
## Parsed with column specification:
## cols(
##   source = col_character(),
##   target = col_character(),
##   approximate = col_double(),
##   no_map = col_double(),
##   combination = col_double(),
##   scenario = col_double(),
##   choice_list = col_double()
## )
```

```
## for our purposes the distinction between forward & backward maps isn't relevant, so we can combine th
gem_combined <-
  bind_rows(
    gem_icd9cm %>% rename(icd9cm=source, icd10cm=target) %>% mutate(direction = "f"),
    gem_icd10cm %>% rename(icd9cm=target, icd10cm=source) %>% mutate(direction = "b")
  )
```

Parsing code ranges

Sources often supply code definitions as ranges, e.g. *Y08-Y09*. Because of the complicated rules of how codes are defined, ranges need to be parsed into sequences, that is, *Y08-Y09* needs to be turned into the sequence *Y08,Y09*. This is not trivial, so below is an algorithm for producing these. Note that this is omitted in the report, but is included in the .Rmd source file!

Helper functions

Below are some helper functions to use in looking at the codes

```
## remove dot
remove_dot <- function(x) gsub(x, pattern="\\.", replacement="")

## helper function to expand range codes eg "960-978" into numeric range...
## and collapse to string eg "960,961,962"
expand_range_to_csv_simple_case <- function(this_range) {
  this_range %>%
   str_split(., pattern="-|-") %>%  # the Schnitzer tables use mid-length dash
   unlist %>%
   trimws %>%
   as.numeric %>%
   {seq(from=min(.), to=max(.), by=1)} %>%
   paste(., collapse=",")
}

## helper functions for generating forward & backward mapped codes from the CM GEM lists
compile_mapping_cm <- function(code) {
  lapply(code, function(x) {
      x <- remove_dot(x)
      gem_combined %>%
      filter(str_starts(icd9cm, pattern=x)) %>%
      mutate(prefix_code=x)
  }) %>% bind_rows
}

## shortcuts for finding codes with a prefix
find_prefix_in_master_list <- function(this_code, master_list) {
  master_list %>%
    filter(str_starts(code, remove_dot(this_code)))
}
find_prefix_in_icd9 <- function(this_code) {
  find_prefix_in_master_list(this_code, master_list = icd9_codes_UK)
}
find_prefix_in_icd10 <- function(this_code) {
  find_prefix_in_master_list(this_code, master_list = icd10_codes_UK)
}
## shortcuts for finding codes by keyword
find_keyword_in_master_list <- function(key_word, master_list) {
  master_list %>%
    filter(str_detect(meaning, regex(pattern = key_word, ignore_case = TRUE)))
}
find_keyword_in_icd9 <- function(key_word) {
  find_keyword_in_master_list(key_word, master_list = icd9_codes_UK)
}
find_keyword_in_icd10 <- function(key_word) {
  find_keyword_in_master_list(key_word, master_list = icd10_codes_UK)
}
## shortcuts for looking up child codes in ICD-9-CM and ICD-10-CM
find_children_in_icd_cm <- function(parent_code) {
  tibble(
    code = icd::children(as.character(parent_code))
  ) %>%
    mutate(
      description = sapply(code, icd::explain_code, USE.NAMES = FALSE)
    )
}
```

ICD-9-CM to ICD-10-CM cross-mapping

Cross-mapping was done manually in consultation with the master lists of ICD-9 and ICD-10 codes, and the online version of the https://icd.who.int/browse10/2016/en. To interpret the original intent of ICD-9-CM codes specified by Schnitzer et al. 2011 and Schnitzer 2004, we also consulted the -CM lists of codes as provided by the *icd* package, specifically using the function *icd::explain_code()*, and the online listing of codes at the (following address)[http://icd9.chrisendres.com/index.php].

Finally, the ICD-9-CM to ICD-10-CM (and reverse) General Equivalence Mapping was also consulted - the helper function *compile_mapping_cm()* was used for this.

The helper functions defined earlier were used to find codes in the master lists by prefix, *find_prefix_in_icd9()*, or by keyword *find_keyword_in_icd9()*.

Some mappings were simple: e.g. gonococcal infection in ICD-9, *098*, has an equivalent ICD-10 code, *A54*.

The mappings were more complicated with edge cases, especially where codes fell into Not otherwise specified (NOS) or Other categories, because some conditions that ICD-9 categorised as Other were categorised as a separate code in ICD-10. For example: E988 in ICD-9 (Injury by other/unspecified means undeteremined whether accidentally or purposely inflicted), subdivided into:

meaning
E988 Injury by other/unspec. means undet. acc./purposely inflicted
E9880 Injury by jumping/lying before moving object undet. acc./purpose inflicte
E9881 Injury by burns, fire undet. acc./purposely inflicted
E9882 Injury by scald, undetermined whether accidentally or purposely inflicted
E9883 Injury by extremes of cold undet. acc./purposely inflicted
E9884 Injury by electrocution undet. acc./purposely inflicted
E9885 Injury by crashing of motor vehicle undet. acc./purposely inflicted
E9886 Injury by crashing of aircraft undet. acc./purposely inflicted
E9887 Injury by caustic subst. ex. poisoning undet. acc./purposely inflicted
E9888 Injury by other spec. means undet. acc./purposely inflicted
E9889 Injury by unspecified means undet. acc./purposely inflicted

. Some of the 4th digit categories in ICD-9 exist as separate 3-character codes in ICD-10:

meaning
Y19 Poisoning by and exposure to other and unspecified chemicals and noxious substances, undetermined intent
Y26 Exposure to smoke, fire and flames, undetermined intent
Y27 Contact with steam, hot vapours and hot objects, undetermined intent
Y31 Falling, lying or running before or into moving object, undetermined intent
Y32 Crashing of motor vehicle, undetermined intent
Y33 Other specified events, undetermined intent
Y34 Unspecified event, undetermined intent

.

Failed attempt at fully algorithmic mapping

The plan was initially to compile a full GEM mapping from ICD-9-CM to ICD-10-CM, and then to attempt mapping from the -CM variant codes to ICD-9 and ICD-10. This turned out to be too arduous a task: the fully automated GEM mapping produces several false leads, and omits some meaningful mappings that are easily spotted by a human, but difficult to write rules for.

Importing ICD inclusion/exclusion criteria from past literature

Gonzalez-Izquierdo et al. 2010

Arturo Gonzalez-Izquierdo kindly provided the list of ICD codes compiled by his team for the 2010 paper, which is in the file **\raw__ICD__codes\CODES - concern groups 2011 v6.xlsx**, specifically in the sheet *Victimization related*.

Processing Gonzalez-Izquierdo et al. 2010 codes

The spreadsheet is nicely formatted for reading, and we can remove several rows and columns to make importing the data easier. Columns 2,3,4 correspond to code description, ICD-10 code, and ICD-9 code, respectively. The data is in rows 2 to 40, which in Excel corresponds to the cell range **B2:B40**.

```
agi_2010_file <- "./raw/CODES - concern groups 2011 v6.xlsx"
agi_2010_raw <- read_excel(agi_2010_file, sheet = "Victimization related", trim_ws = TRUE, range = "B2:

## this produces a "condensed" file, meaning that several codes are expressed as ranges rather than indi
agi_2010_long_condensed <-
  agi_2010_raw %>%
  rename(description_icd10=Description, icd10code =`ICD-10 Codes`, icd9code=`ICD-9 CODES`) %>%  # renam
  filter(!is.na(description_icd10)) %>%  # remove blank rows
  mutate(  # extract header to include as broad maltreatment type variable
    maltreatment_type_agi2010 = if_else(
      condition = is.na(icd10code),
      true = description_icd10,
      false = as.character(NA)
    )
  ) %>%
  fill(maltreatment_type_agi2010) %>%  # LOCF on maltreatment type
  filter(!is.na(icd10code) & !is.na(icd9code)) %>%  # remove header row
  mutate(agi_index = 1:nrow(.)) %>%
  gather(-description_icd10,-maltreatment_type_agi2010,-agi_index, key="icd_version", value="code_decima
  mutate(
    icd_version = case_when(
      icd_version == "icd10code" ~ 10,
      icd_version == "icd9code" ~ 9,
      TRUE ~ as.numeric(NA)
    ),
    code_decimal = gsub(code_decimal, pattern="\\s", replacement="")  # remove spaces
  )
```

Some codes have been entered as ranges rather than a list of codes. The ranges need to be expanded.

There will be two steps to this:

1. Split up comma-separated values and put them in separate rows
2. Split up ranges by generating all values in the range

```
agi_2010_long <-
  agi_2010_long_condensed %>%
  separate_rows(code_decimal, sep = ",") %>%  # separate comma-separated values
  mutate(code_decimal = sapply(code_decimal, function(x) {generate_codes_from_range(x) %>% paste(collaps
  separate_rows(code_decimal, sep = ",") #%>%  # split comma-seprated values again
  # filter((icd_version==9 & code %in% icd9_codes_UK$code_decimal) | (icd_version==10 & code %in% icd10_

## now write the file
write.csv(agi_2010_long, file = "./processed_ICD_codes/AGI_et_al_2010_ICD_codes.csv", row.names = FALSE
```

Schnitzer et al. 2004

This paper sets the precedent to the Schnitzer et al. 2011 paper; in this paper they defined a set of ICD-9-CM definite maltreatment codes, which they would later complement with suggestive codes in the 2011 paper.

The way I extracted the codes was by loading the HTML version of the paper online, and copying **Table 1. ICD-9-CM maltreatment codes** to a spreadsheet, **\raw\Schnitzer_et_al_2004_Table1.csv**

The table has three headings which need to be removed; then the code needs to be separated from the code description - these are in the format **E123: Description**

```
schnitzer_2004_file <- "./raw/Schnitzer_et_al_2004_Table1.csv"
schnitzer_2004_raw <- read.csv(schnitzer_2004_file, header = FALSE, stringsAsFactors = FALSE, strip.whi

schnitzer_2004_icd9cm <-
  schnitzer_2004_raw %>%
  as_tibble() %>%
  rename(code_decimal = V1) %>%
  mutate(
    maltreatment_type_schnitzer2004 =  # extract the header based on phrases they begin with
      if_else(
        condition = str_starts(code_decimal, pattern="Diagnosis codes|External cause of"),
        true = code_decimal,
        false = as.character(NA)
          )
  ) %>%
  fill(maltreatment_type_schnitzer2004) %>%  # LOCF on the header
  filter(str_starts(code_decimal, pattern="Diagnosis codes|External cause of", negate = TRUE)) %>%  # r
  mutate(
    description_icd9 = gsub(code_decimal, pattern="^.*\\: (.*)$", replacement = "\\1"),  # extract desc
    code_decimal = gsub(code_decimal, pattern="^\\s*(.*)\\:.*$", replacement = "\\1"),  # extract code
    inclusion_index = 1:nrow(.)
  )

schnitzer_2004_icd9_10 <-
  schnitzer_2004_icd9cm %>%
  mutate(
    icd9 = case_when(
      code_decimal == "995.50" ~ "995.59",  # child maltreatment unspecified
      code_decimal == "995.51" ~ "placeholder",
      code_decimal == "995.52" ~ "placeholder",
      code_decimal == "995.53" ~ "placeholder",
      code_decimal == "995.54" ~ "placeholder",
      code_decimal == "995.55" ~ "placeholder",
      code_decimal == "995.59" ~ "placeholder",
      code_decimal == "994.2" ~ "placeholder",
      code_decimal == "994.3" ~ "placeholder",
      code_decimal == "E967.0" ~ "placeholder",
      code_decimal == "E967.1" ~ "placeholder",
      code_decimal == "E967.2" ~ "placeholder",
      code_decimal == "E967.3" ~ "placeholder",
      code_decimal == "E967.4" ~ "placeholder",
      code_decimal == "E967.5" ~ "placeholder",
      code_decimal == "E967.6" ~ "placeholder",
      code_decimal == "E967.7" ~ "placeholder",
      code_decimal == "E967.8" ~ "placeholder",
      code_decimal == "E967.9" ~ "placeholder",
      code_decimal == "E968.4" ~ "placeholder",
      code_decimal == "E904.0" ~ "placeholder",
      code_decimal == "E904.1" ~ "placeholder",
      code_decimal == "E904.2" ~ "placeholder",
      code_decimal == "V15.41" ~ "placeholder",
      code_decimal == "V15.42" ~ "placeholder",
      code_decimal == "V15.49" ~ "placeholder",
      code_decimal == "V61.21" ~ "placeholder",
      TRUE ~ code_decimal
    )
  )

## now write the file
# write.csv(schnitzer_2004_processed, file = "./processed_ICD_codes/Schnitzer_et_al_2004_ICD_codes_icd9
```

Schnitzer et al. 2011

Schnitzer et al 2011 provide a list of ICD codes suggestive of child maltreatment, but don't provide the codes in an easily parsed format.

The following is a document of how I extracted the tables and converted them to a usable format.

I tried scraping the web article with no success, so I used a Google Chrome add-on called *table-to-spreadsheet* and saved all tables as .xlsx from the paper to folder *schnitzer2011\raw*.

Note that MS Excel displays a warning about the data being corrupted when you try to open them, but displays the tables just fine - I assume this is the chrome plugin's fault, maybe not following the .xlsx format correctly.

The following tables were taken from the web article:

```
dirTables <- "./schnitzer2011/"
dir(dirTables)
```

```
## [1] "Table 2. ICD-9 codes suggestivea of maltreatment.xlsx"
## [2] "Table A1. ICD-9-CM codes for illnesses and injuries that might raise suspicion of child maltreat
## [3] "Table A2. Included ages, co-occurring exclusion codes and calculated weights for ICD-9-CM codes
## [4] "Table A3. Exclusions for use with the nutritional deficiency and failure to thrive ICD-9-CM code
## [5] "Table A4. ICD-9-CM codes that may indicate sexual abuse for which no visits were available in th
```

Understanding the criteria

Each table is accompanied by additional notes denoted with superscript letters (^{a}, ^{b}, etc.) which need to be dealt with appropriately and removed from the table contents.

Depending on how we define what codes should count as suggestive of child maltreatment, we would use different tables. Schnitzer et al. 2011 compiled a larger list of suggestive codes through expert consultation and literature review which they then further refined by manually reviewing random samples of 50 cases for each ICD code deemed suggestive, and only retaining those where 66% or more of cases where thought to possibly or probably indicate child maltreatment.

Therefore there are two lists:

- The prior list of suggestive ICD codes (table A1 in Schnitzer et al 2011) - this list is broader, i.e. a more liberal criterion
- The empirically tested list of suggestive ICD codes (table 2 in Schnitzer et al 2011) - this list is narrower, i.e. a more conservative criterion

In addition, there is a further list of possible sexual maltreatment ICD codes that were not present in the dataset reviewed by Schnitzer et al 2011 and so could not be reviewed, yet may still be suggestive.

Processing individual Schnitzer 2011 tables

For ease of use in later data analysis, the tables from the Schnitzer et al 2011 paper need to be converted into a format that can be used as a lookup table for use in analysing ICD 9 codes mapping them onto ICD 10 codes.

The tables will be converted to long format, with one row for each ICD code, and all other information in appropriate columns.

Table 2 (narrower/conservative/empirical list of inclusion criteria)

```
schnitzer_2011_table2_file <- "./schnitzer2011/Table 2. ICD-9 codes suggestivea of maltreatment.xlsx"
schnitzer_2011_table2_raw <- read_excel(schnitzer_2011_table2_file, trim_ws = TRUE)
```

```
## New names:
## * `` -> `..5`
## * `` -> `..6`
## * `` -> `..7`
```

The data is currently in a non-standard format, having been grabbed from the table in the paper.

We can remove the last 4 columns because they report the percentage of cases confirmed as suggesting maltreatment on review. Additionally, we can remove the first row which contains no relevant information. We also rename the columns to a more easily programmed-with format, and remove the leading less-than sign from the age.

The type of maltreatment is currently a heading separating rows, but it should be added as a column instead. For this we use a last observation carried forward (LOCF), from *tidyr* library (*tidyr::fill*). Finally we remove the rows that were used for headers.

Next we deal with with the table notes denoted in the paper by superscript letters (^{a}, ^{b}). From the paper:

> a The suggestive codes are those where more than 66% of records reviewed were classified as probable or possible maltreatment.
>
> b Designates a code with a sample size of less than 5.

Superscript ^{a} was only used in the description of the table, and superscript ^{b} has no bearing on the criteria. This means we can safely remove them from the final table.

```
schnitzer_2011_inclusions_empirical_raw <-
  schnitzer_2011_table2_raw %>%
  select(1:3) %>%  # keeping first 3 columns is same s dropping final 4 columns
  slice(-1) %>%  # remove 1st row
  rename(  # rename the columns to a standard format
    code_decimal = `ICD-9 code`,
    description_icd9cm = `Code description`,
    age_less_than = `Age included (years)`
    ) %>%
  mutate(age_less_than = parse_number(age_less_than)) %>%
  mutate(
    maltreatment_type_schnitzer2011 =  # make a new column for type of maltreatment, using those rows t
      ifelse(
        test = str_detect(code_decimal, "^ICD\\-9"),  # headers begin with string "ICD-9"
        yes = code_decimal,
        no = NA
      )
  ) %>%
  fill(maltreatment_type_schnitzer2011) %>%  # LOCF on maltreatment type
  filter(!str_detect(code_decimal, "^ICD\\-9")) %>%  # remove header rows
  mutate(
    description_icd9cm =  # use reg.ex. to find b in final position and remove it
      gsub(description_icd9cm, pattern = "a$|b$", replacement = "")
  ) %>%
  mutate(inclusion_index = 1:n())  # add an index for later use to each inclusion listed

index_table_schnitzer_2011_inclusions_age <-
  schnitzer_2011_inclusions_empirical_raw %>%
  select(age_less_than, inclusion_index) %>%
  distinct
```

Next we deal with rows where multiple codes are denoted in same row, separated by commas. We'll leave the range-defined codes for poisoning intact for now.

Once we have each inclusion listed and paired with an index number, we'll put them in a separate table for later use in cross-mapping to ICD-10.

```
schnitzer_2011_inclusions_empirical_icd9cm <-
  schnitzer_2011_inclusions_empirical_raw %>%
  separate_rows(code_decimal, sep=", ")
```

Now we separate the range-defined codes as well (poisoning codes *960-979*):

```
schnitzer_2011_inclusions_empirical_icd9cm <-
  schnitzer_2011_inclusions_empirical_icd9cm %>%
  mutate(
    code_decimal =
      sapply(code_decimal, USE.NAMES = FALSE,
             FUN= function(x) if (str_detect(x, pattern="-|-")) expand_range_to_csv_simple_case(x) else
    ) %>%
  separate_rows(code_decimal, sep=",") %>%
  mutate(code = remove_dot(code_decimal))  # add a code with the dot removed

index_table_schnitzer_2011_inclusions_icd9cm <-
```

```
  schnitzer_2011_inclusions_empirical_icd9cm %>%
  select(inclusion_code = code, inclusion_code_decimal=code_decimal, inclusion_index)
```

Converting to ICD-9

ICD-9 and ICD-9-CM are supposed to agree in codes up to 4 characters long. Some codes in one version aren't defined in the other and vice versa. We'll check how many of the inclusion criteria in Schnitzer 2011 are missing in ICD-9, and if truncating the last digit yields an appropriate code:

```
schnitzer_2011_inclusions_empirical_icd9cm %>%
  filter(!code %in% icd9_codes_UK$code) %>%
  mutate(truncated_code = str_sub(code, end = -2)) %>%
  left_join(icd9_codes_UK, by=c("truncated_code"="code")) %>%
  select(code_decimal.x, description_icd9cm, truncated_code, meaning)
```

```
## # A tibble: 4 x 4
##   code_decimal.x description_icd9cm   truncated_code meaning
##   <chr>          <chr>                <chr>          <chr>
## 1 V71.81         Observation for abu~ V718           V718 Observation for ~
## 2 362.81         Retinal hemorrhage   3628           3628 Other retinal di~
## 3 852.2          Traumatic subdural ~ 852            852 Subarachnoid, sub~
## 4 E869.4         Second-hand tobacco~ E869           E869 Accidental poiso~
```

Looking at the above, some truncated codes matched to ICD-9 codes all appear too unspecific as there's no guarantee that they would have been used for the same conditions.

The ICD-9 heading *852 Subarachnoid, subdural and extradural haemorrhage, following injury* seems to be a decent fit for ICD-9-CM *852.2 Traumatic subdural hemorrhage.* Although the latter code is a sub-code of *852* in ICD-9-CM, the 4th digit in the CM system specifies the location of the haemorrhage, whereas in ICD-9 the 4th digit specifies whether an an open intercranial wound was present - this distinction is made in ICD-9-CM at 5th digit level.

Note that there is already another *852* heading entry in the Schnitzer 2011 list, *852.2 Traumatic subarachnoid hemorrhage*; this will also be included by adding the less specific *852* code to the inclusions.

For the other ICD-9-CM codes without a match in ICD-9, we can attempt to search for keywords:

362.81 Retinal hemorrhage

```
icd9_codes_UK %>% filter(str_detect(tolower(meaning), pattern="retina") & str_detect(tolower(meaning), 
```

```
## # A tibble: 0 x 6
## # ... with 6 variables: code <chr>, meaning <chr>, node_id <dbl>,
## #   parent_id <dbl>, selectable <chr>, code_decimal <chr>
```

V71.81 Observation for abuse/neglect

```
icd9_codes_UK %>% filter(str_detect(tolower(meaning), pattern="observation") | str_detect(tolower(meanin
```

```
## # A tibble: 35 x 6
##    code  meaning                  node_id parent_id selectable code_decimal
##    <chr> <chr>                      <dbl>     <dbl> <chr>      <chr>
##  1 305   305 Nondependent abuse ~    2646        57 N          305
##  2 3050  3050 Nondependent abuse~    2647      2646 Y          305.0
##  3 3051  3051 Nondependent abuse~    2648      2646 Y          305.1
##  4 3052  3052 Nondependent abuse~    2649      2646 Y          305.2
##  5 3053  3053 Nondependent abuse~    2650      2646 Y          305.3
##  6 3054  3054 Nondependent abuse~    2651      2646 Y          305.4
##  7 3055  3055 Nondependent abuse~    2652      2646 Y          305.5
##  8 3056  3056 Nondependent abuse~    2653      2646 Y          305.6
##  9 3057  3057 Nondependent abuse~    2654      2646 Y          305.7
## 10 3058  3058 Nondependent abuse~    2655      2646 Y          305.8
## # ... with 25 more rows
```

E869.4 Second-hand tobacco smoke

```
icd9_codes_UK %>% filter(str_detect(tolower(meaning), pattern="expos") & str_detect(tolower(meaning), pa
```

```
## # A tibble: 0 x 6
## # ... with 6 variables: code <chr>, meaning <chr>, node_id <dbl>,
## #   parent_id <dbl>, selectable <chr>, code_decimal <chr>
```

```
icd9_codes_UK %>% filter(str_detect(tolower(meaning), pattern="second") & str_detect(tolower(meaning), p
```

```
## # A tibble: 0 x 6
## # ... with 6 variables: code <chr>, meaning <chr>, node_id <dbl>,
## #   parent_id <dbl>, selectable <chr>, code_decimal <chr>
```

```
icd9_codes_UK %>% filter(str_detect(tolower(meaning), pattern="second") & str_detect(tolower(meaning), p
```

```
## # A tibble: 0 x 6
## # ... with 6 variables: code <chr>, meaning <chr>, node_id <dbl>,
## #   parent_id <dbl>, selectable <chr>, code_decimal <chr>
```

Second-hand smoke (ICD-9-M E869.4) was not included as a code in ICD-9, but it was in ICD-10 as code Z58.7 Exposure to tobacco smoke.

To conclude, there don't seem to be appropriate ICD-9 codes for the above ICD-9-CM codes.

The following table shows the ICD-9-CM and ICD-9 codes matched up so you can see that the mapping makes sense:

```
schnitzer_2011_inclusions_empirical_icd9cm %>%
  left_join(icd9_codes_UK, by="code") %>%
  select(code=code_decimal.x, icd_9_cm=description_icd9cm, icd_9=meaning) %>%
  kable
```

code	icd_9_cm	icd_9
054.1	Genital herpes	0541 Genital herpes
098	Gonococcal infection	098 Gonococcal infections
614.9	Pelvic inflammatory disease, unspecified	6149 Unspecified inflammatory disease of female pelvic organs and ti
922.4	Contusion of genital organs	9224 Contusion of genital organs
V71.5	Observation after alleged rape	V715 Observation following alleged rape or seduction
V71.81	Observation for abuse/neglect	NA
362.81	Retinal hemorrhage	NA
807.0	Rib fracture	8070 Fracture of rib(s), closed
807.1	Rib fracture	8071 Fracture of rib(s), open
811	Scapula fracture	811 Fracture of scapula
852.2	Traumatic subdural hemorrhage	NA
853.0	Other/unspecified intracranial hemorrhage	8530 Other/unspec. intracranial haem.foll.without open intracranial
863.1	Stomach injury	8631 Injury to stomach, with open wound into cavity
E965	Assault	E965 Assault by firearms and explosives
E966	Assault	E966 Assault by cutting and piercing instrument
E968.2	Assault	E9682 Assault by striking by blunt or thrown object
E968.9	Assault, NOS	E9689 Assault by unspecified means
E988	Undetermined intent, other means	E988 Injury by other/unspec. means undet. acc./purposely inflicted

code	icd_9_cm	icd_9
800	Skull vault fracture	800 Fracture of vault of skull
805	Vertebral fracture	805 Fracture of vertebral column without mention of spinal cord lesio
852.0	Traumatic subarachnoid hemorrhage	8520 Subarach./subdural/extradural haem. foll.inj.without open intr
862	Intrathoracic injury, NEC	862 Injury to other and unspecified intrathoracic organs
863.2	Small intestine injury	8632 Injury to small intestine, without mention of open wound into c
863.3	Small intestine injury	8633 Injury to small intestine, with open wound into cavity
865	Spleen injury	865 Injury to spleen
952	Spinal cord injury	952 Spinal cord lesion without evidence of spinal bone injury
262	Other severe malnutrition	262 Other severe protein-calorie malnutrition
521.0	Dental caries	5210 Dental caries
692.7	Solar radiation dermatitis	6927 Contact dermatitis and other eczema due to due to solar radiat
808	Pelvic fracture	808 Fracture of pelvis
860	Traumatic pneumohemothorax	860 Traumatic pneumothorax and haemothorax
861	Heart or lung injury	861 Injury to heart and lung
863.8	GI injury, NEC	8638 Injury to other/unspec. g.i. sites, without open wound into cav
864	Liver injury	864 Injury to liver
866	Kidney injury	866 Injury to kidney
941	Burn of head	941 Burn of face, head and neck
942	Burn of trunk	942 Burn of trunk
945	Burn of leg	945 Burn of lower limb(s)
946	Burn of multiple sites	946 Burns of multiple specified sites
960	Poisoning by drugs/medicinals	960 Poisoning by antibiotics
961	Poisoning by drugs/medicinals	961 Poisoning by other anti-infectives
962	Poisoning by drugs/medicinals	962 Poisoning by hormones and synthetic substitutes
963	Poisoning by drugs/medicinals	963 Poisoning by primarily systemic agents
964	Poisoning by drugs/medicinals	964 Poisoning by agents primarily affecting blood constituents
965	Poisoning by drugs/medicinals	965 Poisoning by analgesics, antipyretics and antirheumatics
966	Poisoning by drugs/medicinals	966 Poisoning by anticonvulsants and anti-parkinsonism drugs
967	Poisoning by drugs/medicinals	967 Poisoning by sedatives and hypnotics
968	Poisoning by drugs/medicinals	968 Poisoning by other central nervous system depressants
969	Poisoning by drugs/medicinals	969 Poisoning by psychotropic agents
970	Poisoning by drugs/medicinals	970 Poisoning by central nervous system stimulants
971	Poisoning by drugs/medicinals	971 Poisoning by drugs primarily affecting the autonomic nervous sy
972	Poisoning by drugs/medicinals	972 Poisoning by agents primarily affecting the cardiovascular system
973	Poisoning by drugs/medicinals	973 Poisoning by agents primarily affecting the gastrointestinal syste
974	Poisoning by drugs/medicinals	974 Poisoning by water, mineral and uric acid metabolism drugs
975	Poisoning by drugs/medicinals	975 Poisoning by agents prim. act. on smooth/skeletal muscles, respi
976	Poisoning by drugs/medicinals	976 Poisoning by agents prim.aff.skin, mucous memb.; ophth,otorhin,
977	Poisoning by drugs/medicinals	977 Poisoning by other and unspecified drugs and medicaments
978	Poisoning by drugs/medicinals	978 Poisoning by bacterial vaccines
979	Poisoning by drugs/medicinals	979 Poisoning by other vaccines and biological substances
994.1	Drowning, non-fatal submersion	9941 Drowning and nonfatal submersion
E869.4	Second-hand tobacco smoke	NA
E910.2	Swimming accident	E9102 Acc. drowning/submersion sport without diving equ'pt
E910.4	Bathtub (near) drowning	E9104 Acc. drowning/submersion in bathtub
E910.8	Other (near) drowning	E9108 Other spec. acc. drowning/submersion
E910.9	Accidental (near) drowning, NOS	E9109 Unspec. acc. drowning and submersion
E960.0	Unarmed fight, brawl	E9600 Unarmed fight or brawl
E980	Undetermined intent, poisoning	E980 Injury undet. whether acc. or purposely inflicted (e980-e989)
E985	Undetermined intent, firearm	E985 Injury by firearms, explosives, undet. acc./purposely inflicted
V60	Household circumstances	V60 Housing, household and economic circumstances

All the above shown algorithmically matched codes appear to match in meaning.

Now we can construct a list of ICD-9-CM codes maped to ICD-9 codes.

```
schnitzer_2011_inclusions_empirical_icd9cm_and_icd9 <-
  schnitzer_2011_inclusions_empirical_icd9cm %>%
  left_join(icd9_codes_UK %>% select(code, description_icd9=meaning), by=c("code")) %>%
  mutate(
    code_icd9 = case_when(
      str_starts(code, pattern="852") ~ "852",
      !is.na(description_icd9) ~ code,
      TRUE ~ as.character(NA)
      ),
    description_icd9 = ifelse(code_icd9=="852", icd9_codes_UK$meaning[icd9_codes_UK$code=="852"], descr
  )

schnitzer_2011_inclusions_icd9 <-
  schnitzer_2011_inclusions_empirical_icd9cm_and_icd9 %>%
  filter(!is.na(code_icd9)) %>%
  select(code = code_icd9, description_icd9, inclusion_index, maltreatment_type_schnitzer2011) %>%
  mutate(code_decimal = sapply(code, function(x) icd::short_to_decimal(x), USE.NAMES = FALSE)) %>%  # a
  left_join(schnitzer_2011_inclusions_empirical_icd9cm %>% select(inclusion_index,age_less_than) %>% di

index_table_schnitzer_2011_inclusions_icd9 <-
  schnitzer_2011_inclusions_icd9 %>%
  select(inclusion_index, inclusion_code=code, inclusion_code_decimal=code_decimal)

## write interim results to csv file
# write.csv(schnitzer_2011_inclusions_empirical_icd9, file = "./processed_ICD_codes/Schnitzer_et_al_201
```

Converting to ICD-10

```
# index_table_schnitzer_2011_inclusions_icd9 %>% left_join(icd9_codes_UK %>% select(code, meaning), by=

schnitzer_2011_inclusions_icd10 <-
  index_table_schnitzer_2011_inclusions_icd9cm %>%
  mutate(
    icd10 = case_when(
      inclusion_code == "0541" ~ "A60",
      inclusion_code == "098" ~ "A54",
      inclusion_code == "6149" ~ "N739",
      inclusion_code == "9224" ~ "S302",
      inclusion_code == "V715" ~ "Z044",
      inclusion_code == "V7181" ~ as.character(NA),  # Observation and evaluation for Abuse and neglect
      inclusion_code == "36281" ~ "H356",
      inclusion_code == "8070" ~ "S2230",
      inclusion_code == "8071" ~ "S2231",
      inclusion_code == "8522" ~ "S065",
      inclusion_code == "8530" ~ "S068",
      inclusion_code == "8631" ~ "S3631",
      inclusion_code == "E965" ~ "X93-X96",
      inclusion_code == "E966" ~ "X99",
      inclusion_code == "E9682" ~ "Y00",
      inclusion_code == "E9689" ~ "Y09",
      inclusion_code == "E988" ~ "Y19,Y26,Y27,Y31-Y34",  # several items categorised at 4th digit level
      inclusion_code == "800" ~ "S02.0",
      inclusion_code == "805" ~ "S12.0,S12.1,S12.2,S12.7,S12.9,S22.0,S22.1,S32.0,S32.7", # ICD-9 doesn'
      inclusion_code == "811" ~ "S42.1",
      inclusion_code == "8520" ~ "S066",
      inclusion_code == "862" ~ "S277-S279",  # ICD-9 862 excludes pneumo/hemothorax, whereas ICD-10 co
      inclusion_code == "8632" ~ "S3640",
      inclusion_code == "8633" ~ "S3641",
      inclusion_code == "865" ~ "S360",
      inclusion_code == "952" ~ "S140,S141,S240,S241,S340,S341,T060,T061,T093", # ICD-9 speficies locat
      inclusion_code == "262" ~ "E43",
      inclusion_code == "5210" ~ "K02",
      inclusion_code == "6927" ~ "L578",
      inclusion_code == "808" ~ "S321-S328,T021",  # in ICD-9 pelvis fractures are a single code; in IC
      inclusion_code == "860" ~ "S270-S272",  # see note about intrathoracic injuries above
      inclusion_code == "861" ~ "S26, S273-S276",  # ICD-9 specified heart & lung in single code; ICD-1
      inclusion_code == "8638" ~ "S362,S368,S369",  # ICD-9 separately specifies stomach, small intenst
      inclusion_code == "864" ~ "S361",
      inclusion_code == "866" ~ "S370",
      inclusion_code == "941" ~ "T20",
      inclusion_code == "942" ~ "T21",
      inclusion_code == "945" ~ "T24,T25",  # 945 means burns of lower limbs, and doesn't specify if fe
      inclusion_code == "946" ~ "T29",
      inclusion_code %in% as.character(960:979) ~ "T36-T50",  # poisoning
      inclusion_code == "9941" ~ "T751",
      inclusion_code == "E8694" ~ "Z58.7",  # exposure to tobacco smoke
      inclusion_code == "E9102" ~ "W67-W70",
      inclusion_code == "E9104" ~ "W65,W66",
      inclusion_code == "E9108" ~ "W73",
      inclusion_code == "E9109" ~ "W74",
      inclusion_code == "E9600" ~ "Y04",
      inclusion_code == "E985" ~ "Y22-Y25",
      inclusion_code == "E980" ~ "Y10-Y19",
      inclusion_code == "V60" ~ "Z59",
      TRUE ~ as.character(NA)
    )
  )

index_table_schnitzer_2011_inclusions_icd10 <-
  schnitzer_2011_inclusions_icd10 %>%
  select(inclusion_index, inclusion_code=icd10) %>%
  separate_rows(inclusion_code, sep=",") %>%
  filter(!is.na(inclusion_code)) %>%  # remove missing mapping
  mutate(
    inclusion_code = remove_dot(inclusion_code),
    inclusion_code = sapply(inclusion_code, function(x) generate_codes_from_range(x) %>% paste(collapse
    ) %>%
  separate_rows(inclusion_code, sep=",") %>%
  mutate(inclusion_code_decimal = sapply(inclusion_code, icd::short_to_decimal, USE.NAMES = FALSE)) %>%
  distinct

## Now we write the inclusion tables to files
```

```
write_csv(index_table_schnitzer_2011_inclusions_age, path = "./processed_ICD_codes/schnitzer_et_al_2011_
write_csv(index_table_schnitzer_2011_inclusions_icd9, path = "./processed_ICD_codes/schnitzer_et_al_201
write_csv(index_table_schnitzer_2011_inclusions_icd10, path = "./processed_ICD_codes/schnitzer_et_al_20
```

Exclusion criteria

Like with inclusion criteria, we will compile a table of exclusion criteria. These are extracted from *Tables A2* & *A3*, and are likewise in a one code per row format, except there are two columns of codes - one for the inclusion criterion and another for the exclusion criterion.

Table A2 lists the same criteria we've extracted from *Table 2* above, but with added exclusions. There is an additional note to exclude some codes where the 4th digit is .6 or.7, denoted by superscript *b*; superscript *a* refers to cases whgere N<5 cases were reported in the paper and can be safely removed. Additionally, there is an entry in Table A2 which lists an additional requirement for code **E869.4 Second-hand tobacco smoke** to include at least one from a list of codes. This will be included in a separate table of additional requirements.

Table A3 lists all exclusion criteria for ICD-9-CM code **262: Other severe malnutrition**, so those will need to be added.

```
schnitzer_2011_table2a_file <- "./schnitzer2011/Table A2. Included ages, co-occurring exclusion codes a
schnitzer_2011_table2a_raw <- read_excel(schnitzer_2011_table2a_file, trim_ws = TRUE)

schnitzer_2011_table3a_file <-
  "./schnitzer2011/Table A3. Exclusions for use with the nutritional deficiency and failure to thrive I
schnitzer_2011_table3a_raw <-
  read_excel(schnitzer_2011_table3a_file, trim_ws = TRUE) %>%
  rename(code=Code, condition=Condition) %>%
  mutate(
    malnutrition_index = 1: nrow(.)
  )
```

Table A2

There is an additional note to exclude some codes where the 4th digit is .6 or.7, denoted by superscript *b*; superscript *a* refers to cases whgere N<5 cases were reported in the paper and can be safely removed. Additionally, there is an entry in Table A2 which lists an additional requirement for code **E869.4 Second-hand tobacco smoke** to include at least one from a list of codes. This will be included in a separate table of additional requirements.

Not all entries in Table A2 also have exclusion codes (or extra requirements), so we remove ones that don't to begin with. We also add the inclusion index generated earlier when we processed inclusion codes - this will allow us a to produce a more compact file.

We also remove superscript ^{a} from codes as it has no bearing on our project.

```
schnitzer_2011_exclusions_empirical_raw <-
  schnitzer_2011_table2a_raw %>%
  select(exclusions=`Co-Occurring exclusion codes`, code_decimal=Code, condition=Condition, -Weight, -`A
  filter(!is.na(exclusions)) %>%  # remove rows where no exclusion code was provided
  # mutate(exclusion_index = 1:nrow(.)) %>%  # add index variable
  left_join(schnitzer_2011_inclusions_empirical_raw %>% select(code_decimal, inclusion_index), by="code_
  mutate(
    condition = sub(condition, pattern="a$", replacement="")  # remove final a in Conditions that inclu
  )
```

Additional inclusion criteria

We'll start by compiling the table of extra inclusion requirements, which can be spotted in table A2 by the phrase *Include only* in the exclusions column:

```
schnitzer_2011_inclusions_empirical_extra_requirements_icd9cm <-
  schnitzer_2011_exclusions_empirical_raw %>%
  filter(str_detect(exclusions, pattern="Include only")) %>%  # the phrase denoting additional code req
  mutate(exclusions = sub(exclusions, pattern="Include only with co-occurring code\\: ", replacement=""
  separate_rows(exclusions, sep="; or |; ") %>%
  select(extra_required_range_icd9cm=exclusions,inclusion_index)

schnitzer_2011_inclusions_empirical_extra_requirements_icd9_icd10_condensed <-
  schnitzer_2011_inclusions_empirical_extra_requirements_icd9cm %>%
  mutate(
    extra_required_range_icd9 = case_when(
      extra_required_range_icd9cm == "480.0-487.8" ~ "480-487",
      extra_required_range_icd9cm == "490.0-491.9" ~ "490,491",
      extra_required_range_icd9cm == "466.0-466.19" ~ "466",
      extra_required_range_icd9cm == "493.0-493.9" ~ "493",
      TRUE ~ as.character(extra_required_range_icd9cm)
    ),
    extra_required_range_icd10 = case_when(
      extra_required_range_icd9cm == "480.0-487.8" ~ "J10-J18",  # pneumonia & influenza (not avian) - 
      extra_required_range_icd9cm == "490.0-491.9" ~ "J40-J42",  # bronchitis not specified acute/chron
      extra_required_range_icd9cm == "466.0-466.19" ~ "J20-J21,J68",  # Acute bronchitis and bronchioli
      extra_required_range_icd9cm == "493.0-493.9" ~ "J44-J46",  # asthma
      extra_required_range_icd9cm == "381.0-381.4" ~ "H65",  # Nonsuppurative otitis media
      TRUE ~ as.character(NA)
    )
  ) %>%
  select(
    inclusion_index,
    icd9=extra_required_range_icd9, icd10=extra_required_range_icd10
  )

schnitzer_2011_inclusions_empirical_extra_requirements <-
  schnitzer_2011_inclusions_empirical_extra_requirements_icd9_icd10_condensed %>%
  gather(-inclusion_index, key="icd_version", value="required_any") %>%
  mutate(
    icd_version = parse_number(icd_version)
  ) %>%
  separate_rows(required_any,sep=",") %>%
  mutate(required_any = sapply(required_any,function(x) generate_codes_from_range(x) %>% paste(collapse
  separate_rows(required_any,sep=",")

index_table_schnitzer_2011_inclusions_extra_requirements_icd10 <-
  schnitzer_2011_inclusions_empirical_extra_requirements %>%
  filter(icd_version==10) %>%
  select(inclusion_index, required_any_code=required_any)
```

```
## Write the file
write_csv(index_table_schnitzer_2011_inclusions_extra_requirements_icd10, path = "./processed_ICD_codes/
```

Exclusions

Now that we've processed the extra requirements, we can remove that row from the remaining data, and the entry for malnutrition which specifies an entire separate table, which will be dealt with later.

```
schnitzer_2011_exclusions_empirical_raw_remaining <-
  schnitzer_2011_exclusions_empirical_raw %>%
  filter(!str_detect(exclusions, pattern="Include only")) %>%  # the phrase denoting additional code re
  filter(!str_detect(exclusions, pattern="See Table")) #%>%  # the phrase denoting the separate malnutr
  # mutate(entry_index = 1:n())  # add entry index for use in emrging later
```

We have 30 rows with a mixture of specified single codes, code ranges, and semi-colon separated codes and code ranges.

Included are code ranges with superscript *b* denoting exclusion codes:

> Unless 4th digit = .6 or.7.

The *b* always follows the second code in a range, i.e.: E815–E819^{b}. Ultimately we will generate the full range of specified codes, and remove cases where 4th digit is .6 or .7 where noted.

Note also that all ranges where superscript *b* applies are listed as an interrupted range, separated by a comma: *E810–E813, E815–E819b*; superscript *b* applies to both of these ranges. This specifies external cause codes for transport accidents, excluding cases where the victim was a pedestrian or cyclist (4th digit .6 or .7). E814 is entirely for pedestrian victims and so was excluded in the range.

This means there are now these cases: * Single code entries * Semicolon-separated entries + Single codes + Code ranges + Interrupted code ranges, separated by comma

The elegant solution is to split comma-separated values first, then deal with range-defined codes by transforming the range into a long list of comma-separated codes, and then split comma-seprated values a second time.

Instead of compiling these line-by-line, we can group the exclusion codes, process them individually, and then re-merge them with the target codes. We'll split the codes with semicolons into separate rows, and it should be obvious afterwards that there's only a small number of groups of exclusion codes to deal with that repeat throughout the table.

Cross-mapping exclusions to ICD-9 and ICD-10

Next we deal with the exclusions. We'll put semicolon separated values in separate rows to begin with, add an id, and then deal with them individually, mapping them to ICD-9 and ICD-10.

```
schnitzer_2011_exclusions_empirical_separated_semicolon_exclusions <-
  schnitzer_2011_exclusions_empirical_raw_remaining %>%
  separate_rows(exclusions, sep="; ") %>%
  select(inclusion_index, exclusions)

index_table_schnitzer_2011_exclusions_icd9cm_condensed <-
  schnitzer_2011_exclusions_empirical_separated_semicolon_exclusions %>%
  select(exclusions) %>%
  distinct %>%
  arrange(nchar(exclusions)) %>%
  mutate(exclusion_index = 1:n())
```

```
index_table_schnitzer_2011_map_inclusions_and_exclusions <-
  schnitzer_2011_exclusions_empirical_separated_semicolon_exclusions %>%
  left_join(index_table_schnitzer_2011_exclusions_icd9cm_condensed, by="exclusions") %>%
  select(matches("index"))
```

As mentioned above, there are only a small(ish) number of separate exclusions listed:

exclusions	exclusion_index
767	1
765	2
771.2	3
098.4	4
771.6	5
756.51	6
E960.1	7
E968.4	8
286–287	9
E800–E819	10
E890–E897	11
E870–E876	12
733.10–733.19	13
E810–E813, E815–E819b	14

We also now have a lookup table linking the code entries for target codes in the exclusions table, and the exclusions:

inclusion_index	exclusion_index
1	3
2	4
2	5
4	10
4	9
7	14
7	9
8	14
8	6
8	1
8	13
8	2
9	14
9	6
9	1
9	13
9	2
10	14
10	9
11	14
11	9
12	14
13	7
13	8
14	7

inclusion_index	exclusion_index
14	8
16	14
16	6
16	1
16	13
16	2
17	14
17	6
17	1
17	13
17	2
18	14
18	9
19	14
20	14
21	14
22	10
26	14
26	6
26	1
26	13
26	2
27	14
28	14
29	14
30	14
31	14
32	11
33	11
34	11
35	11
36	12
43	7
43	8

Now that each of the unique groupings of exclusions have had an id assigned and we've also compiled a mapping of exclusion ids to inclusion ids, we can also convert the exclusion codes to ICD-9 and ICD-10.

```
index_table_schnitzer_2011_exclusions_icd9_10_condensed <-
  index_table_schnitzer_2011_exclusions_icd9cm_condensed %>%
  mutate(
    icd_9 = case_when(
      exclusions == "771.2" ~ "771.22",  # congenital herpes simplex - in ICD-9-CM it's grouped with "O
      exclusions == "756.51" ~ "756.50",  # osteogenesis imperfecta - different 5th digits
      exclusions == "733.10-733.19" ~ "733.1",  # pathological fracture - covers entire 4th digit range
      TRUE ~ exclusions
    ),
    icd_10 = case_when(
      exclusions == "767" ~ "P10-P15,P52.4,P52.6,P52.8,P52.9",  # birth trauma, plus several non-trauma
      exclusions == "765" ~ "P05,P07",  # Disorders relating to short gestation and low birthweight
      exclusions == "771.2" ~ "P35.2",  # congenital herpes
      exclusions == "098.4" ~ "A54.3",  # Gonococcal infection of eye
      exclusions == "771.6" ~ "P39.1",  # Neonatal conjunctivitis and dacryocystitis
      exclusions == "756.51" ~ "Q78.0",  # Osteogenesis imperfecta
      exclusions == "E960.1" ~ "Y05,T74.2",  # rape (I included attempted rape Y05 - this would flag as
      exclusions == "E968.4" ~ "Y06",  # ICD-9 criminal neglect, ICD-10 Neglect and abandonment
      exclusions == "286-287" ~ "D65-D69",  # Coagulation defects, purpura and other haemorrhagic condi
      exclusions == "E800-E819" ~ "V01-V99",  # ICD-9-CM railway & motor traffic accidents: ICD-10 tran
      exclusions == "E890-E897" ~ "X00-X09",  # ICD-9-CM ACCIDENTS CAUSED BY FIRE AND FLAMES (but witho
      exclusions == "E870-E876" ~ "Y60-Y69",  # ICD-9-CM MISADVENTURES TO PATIENTS DURING SURGICAL AND
      exclusions == "733.10-733.19" ~ "M48.5,M80,M84.4,M90.7",  # ICD-9 pathological fracture: scattere
      exclusions == "E810-E813, E815-E819b" ~ "V20-V99",  # ICD-9 MOTOR VEHICLE TRAFFIC ACCIDENTS (E810
      TRUE ~ as.character(NA)
    )
  ) %>%
  select(-exclusions)  # remove the original ICD-9-CM codes, no longer needed
```

Separating exclusion codes to individual codes

Most exclusions are listed as ranges of codes, so we'll now separate them next.

The first step is to tag the line with superscript ^{b} so we can deal with it later, then the superscript ^{b} can be removed from the exclusion codes. Note that this only applies to the ICD-9 codes - ICD-10 codes specify pedestrian/cyclist at 3-character level and only appropriate codes have been included already.

Second, we'll separate comma separated codes. Third, we'll expand codes expressed as ranges and convert them into comma-seaparated codes, Fourth, we'll separate the newly generated comma-separated values again

```
## helper function to deal with superscript b cases
## essentially, we are given a 4 character E-code, e.g. E815
## and we need to compile a range of 5th character codes, E815.1 - E815.9, but excluding .6 and .7
generate_5th_digit_codes_superscript_b <- function(code) {
  digits = c(1:9)[-c(6,7)]
  return(paste(code,digits,sep="."))
}

index_table_schnitzer_2011_exclusions_icd9 <-
  index_table_schnitzer_2011_exclusions_icd9_10_condensed %>%
  select(-icd_10) %>%  # remove icd-10 codes
  mutate(
    superscript_b = str_detect(icd_9, pattern="b$"),  # tag superscript b so we can deal with it later
    icd_9 = sub(icd_9, pattern="b$", replacement = "")  # remove final now that we have a tag
  ) %>%
  separate_rows(icd_9, sep = ", ") %>%  # separate comma-separated values
  mutate(icd_9 = sapply(icd_9, FUN=function(x) generate_codes_from_range(x) %>% paste(collapse = ","), 
  separate_rows(icd_9, sep = ",") %>%
  mutate(  # convert superscript b codes to comma separated 5th digit codes
    icd_9 =
      ifelse(
        superscript_b,
          sapply(
            icd_9,
            FUN=function(x) generate_5th_digit_codes_superscript_b(x) %>% paste(collapse = ","),
            USE.NAMES = FALSE
          ),
        icd_9
      )
  ) %>%
  separate_rows(icd_9, sep = ",") %>%  # separate newly generated 5th digit codes
  select(-superscript_b) %>%  # the superscript b tag is no longer needed
  rename(exclusion_code_decimal = icd_9) %>%
  mutate(exclusion_code = remove_dot(exclusion_code_decimal))

index_table_schnitzer_2011_exclusions_icd10 <-
  index_table_schnitzer_2011_exclusions_icd9_10_condensed %>%
  select(-icd_9) %>%  # remove icd-9 codes
  separate_rows(icd_10, sep=",") %>%
  mutate(icd_10 = sapply(icd_10, FUN=function(x) generate_codes_from_range(x) %>% paste(collapse = ",")
  separate_rows(icd_10, sep=",") %>%
  rename(exclusion_code_decimal = icd_10) %>%
  mutate(exclusion_code = remove_dot(exclusion_code_decimal))
```

Table A3

Table A3 is is again a case of parsing out codes that have been denoted as ranges of codes; we do this by separately dealing with single codes and codes that include a hyphen. There were no comma-separated codes listed, so we can go straight into converting to ICD-9 and ICD-10.

```
schnitzer_2011_malnutrition_exclusions_icd9_icd10 <-
  schnitzer_2011_table3a_raw %>%
  mutate(
    icd9_range = case_when(
      code=="009.0" ~ "009.0",  # Infectious colitis, enteritis and gastroenteritis
      code=="010.0-018.9" ~ "010-018",  # covers entire range
      code=="042" ~ "042-044",  # ICD-9 codes 042-044 are for HIV
      code=="070.00-070.9" ~ "070",  # covers entire range
      code=="140.0-208.91" ~ "140-208",  # covers entire range
      code=="243-244.9" ~ "243-244",  # covers entire range
      code=="250.00-250.93" ~ "250",  # covers entire code block
      code=="252.0-252.9" ~ "252",  # covers entire code block
      code=="253.0-253.9" ~ "253",  # covers entire code block
      code=="270.0-275.9" ~ "270-275",  # covers entire code block
      code=="277.00" ~ "277.0",  # cystic fibrosis
      code=="330.0-344.42" ~ "330-344",  # covers entire range
      code=="446.0-446.7" ~ "446",  # covers entire code block
      code=="493.00-493.92" ~ "493",  # covers entire code block
      code=="530.81" ~ "530.8",  # ICD-9 Other disorders of oesophagus = best match for ICD-9-CM Esopha
      code=="555.0-558.9" ~ "555-558",  # covers entire code range
      code=="575.0-576.9" ~ "575-576",  # covers entire code range
      code=="571.0-571.9" ~ "571",  # covers entire code block
      code=="579.0-579.9" ~ "579",  # covers entire code block
      code=="593.9" ~ "585.9",  # Schnitzer says Chronic renal insufficiency, which is actually ICD-9-C
      # TODO: find out if this is typo, or what to do about it!
      code=="710.0-710.9" ~ "710",  # covers entire code block
      code=="714.0-714.9" ~ "714",  # covers entire code block (note how ICD-9-CM is less specific than
      code=="745.0-747.9" ~ "745-747",  # covers entire code block
      code=="749.00-749.25" ~ "749",  # covers entire code block
      code=="758.0-758.9" ~ "758",  # covers entire code block
      code=="760.71" ~ "760.76",  # Fetus/newborn affected by alcohol via placenta/breast milk
      code=="771.0-771.89" ~ "771",  # covers entire code block
      code=="772.10-772.14" ~ "772.1",  # Fetal and neonatal intraventricular haemorrhage
      code=="852.00-853.19" ~ "852-853",  # traumatic intracranial haemorrhages (subarach.,subdural,ext
      code=="984.0-984.9" ~ "984",  # covers entire range
      TRUE ~ code
    ),
    icd10_range = case_when(
      code=="009.0" ~ "A09",  # ICD-9 Infectious colitis, enteritis and gastroenteritis to Other gastro
      code=="010.0-018.9" ~ "A15-A19",  # Tuberculosis
      code=="042" ~ "B20-B24",  # HIV
      code=="070.00-070.9" ~ "B15-B19",  # viral hepatitis
      code=="140.0-208.91" ~ "C00-C97",  # malignancy
      code=="243-244.9" ~ "E00-E03,E89.0",  # congenital & acquired hypothyroidism; E89.0 postprocedura
      code=="250.00-250.93" ~ "E10-E14",  # diabetes mellitus
      code=="252.0-252.9" ~ "E20-E21, E89.2",  # Parathyroid disorders; E89.0 Postprocedural hypothyroi
      code=="253.0-253.9" ~ "E22-E23, E89.3", # E22 Hyperfunction of pituitary gland; E23 Hypofunction
      code=="270.0-275.9" ~ "E70-E90, D89, M10",  # E70-E90 Metabolic disorders; However, there's match
      ## potentially also: TODO: (DECIDE!)
      ## D47 Other neoplasms of uncertain or unknown behaviour of lymphoid, haematopoietic and related
      # compile_mapping_cm(as.character(270:275)) %>% filter(!str_detect(icd10cm,pattern="E[7-9]|D89|D4
      code=="271.3" ~ "E73",  # E73 lactose intolerance
      code=="277.00" ~ "E84",  # E84 Cystic fibrosis
      code=="317-319" ~ "F70-F79",  # mental retardation
      code=="330.0-344.42" ~ "G00-G99, R52",  # neurologic hereditary etc - cocneptually matches G00-G9
      # this range is composed of:
      # HEREDITARY AND DEGENERATIVE DISEASES OF THE CENTRAL NERVOUS SYSTEM (330-337)
      # PAIN (338)
      # OTHER HEADACHE SYNDROMES (339)
      # OTHER DISORDERS OF THE CENTRAL NERVOUS SYSTEM (340-349) except 348-349, which covers Other cond
      code=="431" ~ "I61",  # Intracerebral hemorrhage
      code=="446.0-446.7" ~ "M30-M31",  # Polyarteritis nodosa and allied conditions to ICD-10 M30 Poly
      code=="493.00-493.92" ~ "J45",  # J45 asthma
      code=="530.81" ~ "K21",  # K21 Gastro-oesophageal reflux disease
      code=="555.0-558.9" ~ "K50-K52",  # Inflammatory bowel disease: NONINFECTIOUS ENTERITIS AND COLIT
      code=="575.0-576.9" ~ "K82-K83",  # K82 Other diseases of gallbladder; K83 Other diseases of bili
      code=="571.0-571.9" ~ "K70-K76",  # K70-K76 Diseases of liver (K77 excluded - diseases of liver c
      code=="577.8" ~ "K86.8",  # Pancreatic insufficiency (Other specified diseases of pancreas) - K86
      code=="579.0-579.9" ~ "K90-K91",  # K90 intestinal malabsorption; K91 Postprocedural disorders of
      code=="588.8" ~ "N25.8",  # N25.8 Other disorders resulting from impaired renal tubular function
      code=="593.9" ~ "N18-N19",  # Schnitzer specified "Chronic renal insufficiency", but listed code
      # TODO: confirm chronic renal insufficiency
      code=="599.0" ~ "N39.0",  # N39.0 Urinary tract infection, site not specified
      code=="710.0-710.9" ~ "M32-M36",  # M32-M36 based on matching ICD-9-CM code headings to ICD-10 &
      code=="714.0-714.9" ~ "M05-M06, M08, M12.0",  # based on GEMs, we also include M12.0 Chronic post
      code=="745.0-747.9" ~ "Q20-Q28, P29.3",  # "Cardiac disease, congenital" - Q20-Q28 Congenital mal
      code=="749.00-749.25" ~ "Q35-Q37",  # Q35-Q37 Cleft lip and cleft palate
      code=="750.5" ~ "Q40.0",  # congenital pyloric stenosis
      code=="751.3" ~ "Q43.1",  # Hirschsprung disease
      code=="758.0-758.9" ~ "Q90-Q99",  # Chromosomal abnormalities, not elsewhere classified (Q90-Q99)
      code=="760.71" ~ "Q86.0, P04.3", # Q86.0 Fetal alcohol syndrome (dysmorphic); P04.3 Fetus and new
      code=="767.0" ~ "P10, P11.1, P11.2",  # P10 Intracranial laceration and haemorrhage due to birth
      code=="770.7" ~ "P27.1",  # P27.1 Bronchopulmonary dysplasia originating in the perinatal period
      code=="771.0-771.89" ~ "P35-P39",  # Infections specific to the perinatal period (P35-P39)
      code=="772.10-772.14" ~ "P52",  # P10.2 Intraventricular haemorrhage due to birth injury already
      code=="852.00-853.19" ~ "S063-S068",  # traumatic intracranial haemorrhages
      code=="984.0-984.9" ~ "T56.0",  #  T56.0 toxic effect of lead and its compounds
      TRUE ~ code
    )
  )
```

Merging malnutrition-related exclusion codes to exclusions

We now add the malnutrition-related exclusion codes to the ICD-9 and ICD-10 exclusion codes already generated. In addition, we add a new index representing the malnutrition codes to the mapping of inclusion to exclusion code indexes.

```
malnutrition_code <- "262"
malnutrition_index <- index_table_schnitzer_2011_inclusions_icd9$inclusion_index[index_table_schnitzer_

malnutrition_exclusion_index_new <- max(c(index_table_schnitzer_2011_exclusions_icd9$exclusion_index,in

index_table_schnitzer_2011_malnutrition_exclusions_condensed <-
  schnitzer_2011_malnutrition_exclusions_icd9_icd10 %>%
  select(
    exclusion_range_icd9 = icd9_range,
    exclusion_range_icd10 = icd10_range
    ) %>%
  mutate(exclusion_index = malnutrition_exclusion_index_new)

index_table_schnitzer_2011_malnutrition_exclusions_icd9 <-
  index_table_schnitzer_2011_malnutrition_exclusions_condensed %>%
  select(exclusion_index, exclusion_range_icd9) %>%
  mutate(exclusion_range_icd9 = gsub(exclusion_range_icd9,pattern=" ",replacement = "")) %>%  # remove
  separate_rows(exclusion_range_icd9, sep=",") %>%
  mutate(exclusion_code = sapply(exclusion_range_icd9,function(x) generate_codes_from_range(x) %>% paste
  separate_rows(exclusion_code, sep=",") %>%
  select(exclusion_index, exclusion_code_decimal=exclusion_code) %>%
  mutate(exclusion_code = icd::decimal_to_short(exclusion_code_decimal) %>% as.character)

index_table_schnitzer_2011_malnutrition_exclusions_icd10 <-
  index_table_schnitzer_2011_malnutrition_exclusions_condensed %>%
  select(exclusion_index, exclusion_range_icd10) %>%
  mutate(exclusion_range_icd10 = gsub(exclusion_range_icd10,pattern=" ",replacement = "")) %>%  # remov
  separate_rows(exclusion_range_icd10, sep=",") %>%
  mutate(exclusion_code = sapply(exclusion_range_icd10,function(x) generate_codes_from_range(x) %>% pas
  separate_rows(exclusion_code, sep=",") %>%
  select(exclusion_index, exclusion_code_decimal=exclusion_code) %>%
  mutate(exclusion_code = icd::decimal_to_short(exclusion_code_decimal) %>% as.character)

## add new mapping of maltreatment inclusion & exclusions
index_table_schnitzer_2011_map_inclusions_and_exclusions <-
  index_table_schnitzer_2011_map_inclusions_and_exclusions %>%
  add_case(inclusion_index=malnutrition_index, exclusion_index=malnutrition_exclusion_index_new)

index_table_schnitzer_2011_exclusions_icd9 <-
  bind_rows(
    index_table_schnitzer_2011_exclusions_icd9,
    index_table_schnitzer_2011_malnutrition_exclusions_icd9
  )

index_table_schnitzer_2011_exclusions_icd10 <-
  bind_rows(
    index_table_schnitzer_2011_exclusions_icd10,
    index_table_schnitzer_2011_malnutrition_exclusions_icd10
  )
```

Write completed exclusion indices to files

```
## write index of inclusions mapped to exclusions
write_csv(index_table_schnitzer_2011_map_inclusions_and_exclusions, path = "./processed_ICD_codes/schni

write_csv(index_table_schnitzer_2011_exclusions_icd9, path = "./processed_ICD_codes/schnitzer_et_al_201

write_csv(index_table_schnitzer_2011_exclusions_icd10, path = "./processed_ICD_codes/schnitzer_et_al_20
```

Conclusion

We now have a cross-mapped to ICD-9 and ICD-10 list of Schnitzer 2011 inclusion codes, as well as exclusion codes (including for maltreatment) and a separate list of additional inclusion requirements for the 2nd hand smoke exposure code in ICD-10.