Extract ICD code inclusion & exclusion criteria from tables in Schnitzer et al 2011

Jan Savinc

15 January 2019

Intro

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 fine.

The following tables were taken from the web article:

```
dirTables <- "./schnitzer2011/raw/"
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 maltrea
## [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 cod
## [5] "Table A4. ICD-9-CM codes that may indicate sexual abuse for which no visits were available in t
```

Understanding the criteria

Each table is accompanied by additional notes denoted with superscript letters ($^\text{a}$, $^\text{b}$, 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 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 and possibly to map 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.

All processed tables will be exported to *schnitzer2011\output*

Table 2 (narrower/conservative list of codes)

```
table2Path <- "./schnitzer2011/raw/Table 2. ICD-9 codes suggestivea of maltreatment.xlsx"
table2Raw <- read_excel(table2Path, trim_ws = TRUE)

dirOutput <- "./schnitzer2011/output"
```

The data is currently in a non-standard format.

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.

```
table2Processed <-
  table2Raw %>%
  select(1:3) %>%  # keeping first 3 columns is same s dropping final 4 columns
  slice(-1)  # remove 1st row
```

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. Finally we remove the rows that were used for headers.

```
table2Processed <-
  table2Processed %>%
  mutate(
    MaltreatmentType =  # make a new column for type of maltreatment, using those rows that were used a
      ifelse(
        test = str_detect(`ICD-9 code`, "ICD-9"),  # headers begin with string "ICD-9"
        yes = `ICD-9 code`,
        no = NA
      )
  ) %>%
  fill(MaltreatmentType) %>%  # LOCF on maltreatment type
  filter(!str_detect(`ICD-9 code`, "ICD-9"))  # remove header rows
```

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.

```
table2Processed <-
  table2Processed %>%
  mutate(
    `Code description` =  # use reg.ex. to find b in final position and remove it
      gsub(`Code description`, pattern = "^(.*)b$", replacement = "\\1")
  )
```

Next we deal with rows where multiple codes are denoted in same row. This includes cases where they are delimited by commas, and cases where a range is listed.

```
table2Processed.singleCodes <-
  table2Processed %>%
  filter(!str_detect(`ICD-9 code`, pattern=",|-|-")) %>%  # mid-length dash is used in the original tab
  mutate(ICD9code = `ICD-9 code`)

table2Processed.multipleCodes <-
  table2Processed %>%
  filter(str_detect(`ICD-9 code`, pattern=", ")) %>%
  split(seq_len(nrow(.))) %>%
  map(function(x) {
      codes <- x$`ICD-9 code` %>% str_split(., pattern=", ") %>% unlist
      newRows <- cbind(
        x,
        ICD9code = codes,
        stringsAsFactors = FALSE
      ) %>% as.tibble()
    }) %>%
  bind_rows()

table2Processed.rangeCodes <-
  table2Processed %>%
  filter(str_detect(`ICD-9 code`, pattern="-|-")) %>%
  cbind(
    .,
    ICD9code =
      str_split(.$`ICD-9 code`, pattern = "-") %>%
      unlist %>%
      as.numeric %>%
      (function(x) seq(from=min(x), to=max(x), by=1)) %>%
      as.character(),
    stringsAsFactors = FALSE
    ) %>%
  as.tibble()

table2Processed <-
  bind_rows(
    table2Processed.singleCodes,
    table2Processed.multipleCodes,
    table2Processed.rangeCodes
    ) %>%
  select(-`ICD-9 code`)  # remove the unformatted ICD-9 code - we now have a clean version, one code pe
```

We also clean up the age criteria, all of which are an upper bound only.

```
table2Processed <-
  table2Processed %>%
  mutate(
    AgeMaximum = `Age included (years)`,
    ICD9description = `Code description`
    ) %>%  # while we're at it, also rename the column Code description to be in line with csv variable
  select(-`Age included (years)`, -`Code description`) %>%
  select(ICD9code, ICD9description, AgeMaximum, MaltreatmentType)
```

Now Table 2 can be exported as .csv for later use as a lookup table in including/excluding cases.

```
# TODO: include the unobserved sexual maltreatment criteria
# TODO: compile an exclusion table also!
# TODO: add more columns if needed
# TODO: change the age column to be age upper bound, and add an age lower bound, for example

table2Filename <- "icd9_inclusion_criteria_Schnitzer2011_narrow.csv"
write.csv(table2Processed, file = file.path(dirOutput,table2Filename), row.names = FALSE)
```

Exclusion criteria

Like with inclusion criteria