Final Project

Annie Lin

2022-12-16

Project Proposal

There are many websites that can convert convert ICD-9 code into ICD-10 code (and vice versa), but they can only convert one code at a time, which consumed me a lot of time when I did my BST210 regression project. Thus, I want to use R to convert a set of ICD codes (as many as you want) all at once.

For the 2nd part of this project, I will use data from Kaggle to build a regression model to predict opioids overdose. Because back in Taiwan, I was an anesthesiologist. In our daily practice, to treat patients' pain, opioids (such as morphine) are often used. However, opioids are very easily to be addictive to. Once these drugs are used overdose (very likely for those drug abusers), they would not only put people into sleep but suppress their breath, heart rate, and blood pressure – but people cannot react because they are deeply sedated! In the end, they are usually found dead. To prevent these tragedies, if we can predict people potentially with higher possibility of opioids overdose, we may avoid using (or use less) these highly addictive drugs on them and adopt other alternative treatment or medications.

##Part 1 Data Wrangling ##Introduction

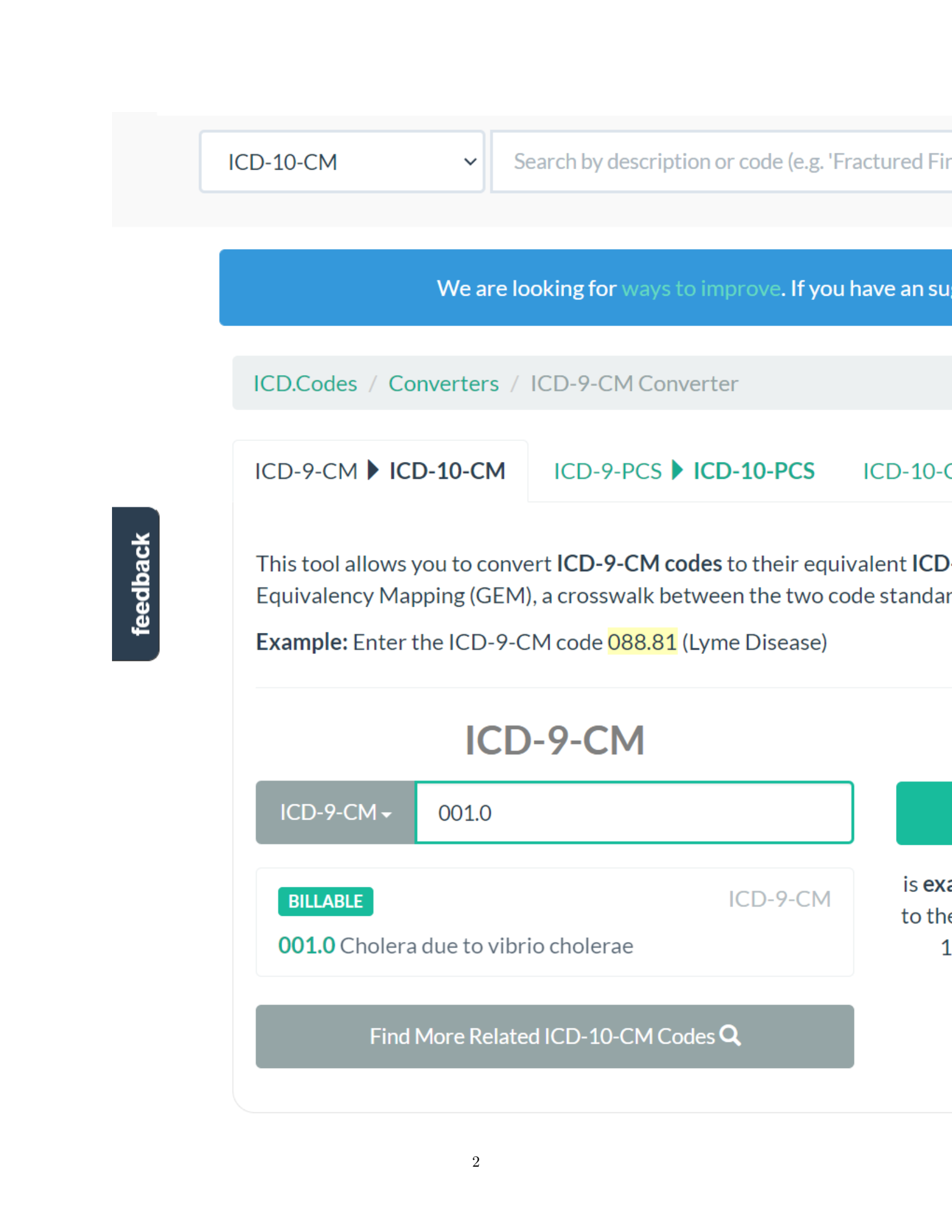
My ICD codes files are from here.

```
## [1] 23912
```

```
##   icd9cm icd10cm flags approximate no_map combination scenario choice_list
## 1     10    A000     0           0      0           0        0           0
## 2     11    A001     0           0      0           0        0           0
## 3     19    A009     0           0      0           0        0           0
## 4     20   A0100 10000           1      0           0        0           0
## 5     21    A011     0           0      0           0        0           0
## 6     22    A012     0           0      0           0        0           0
```

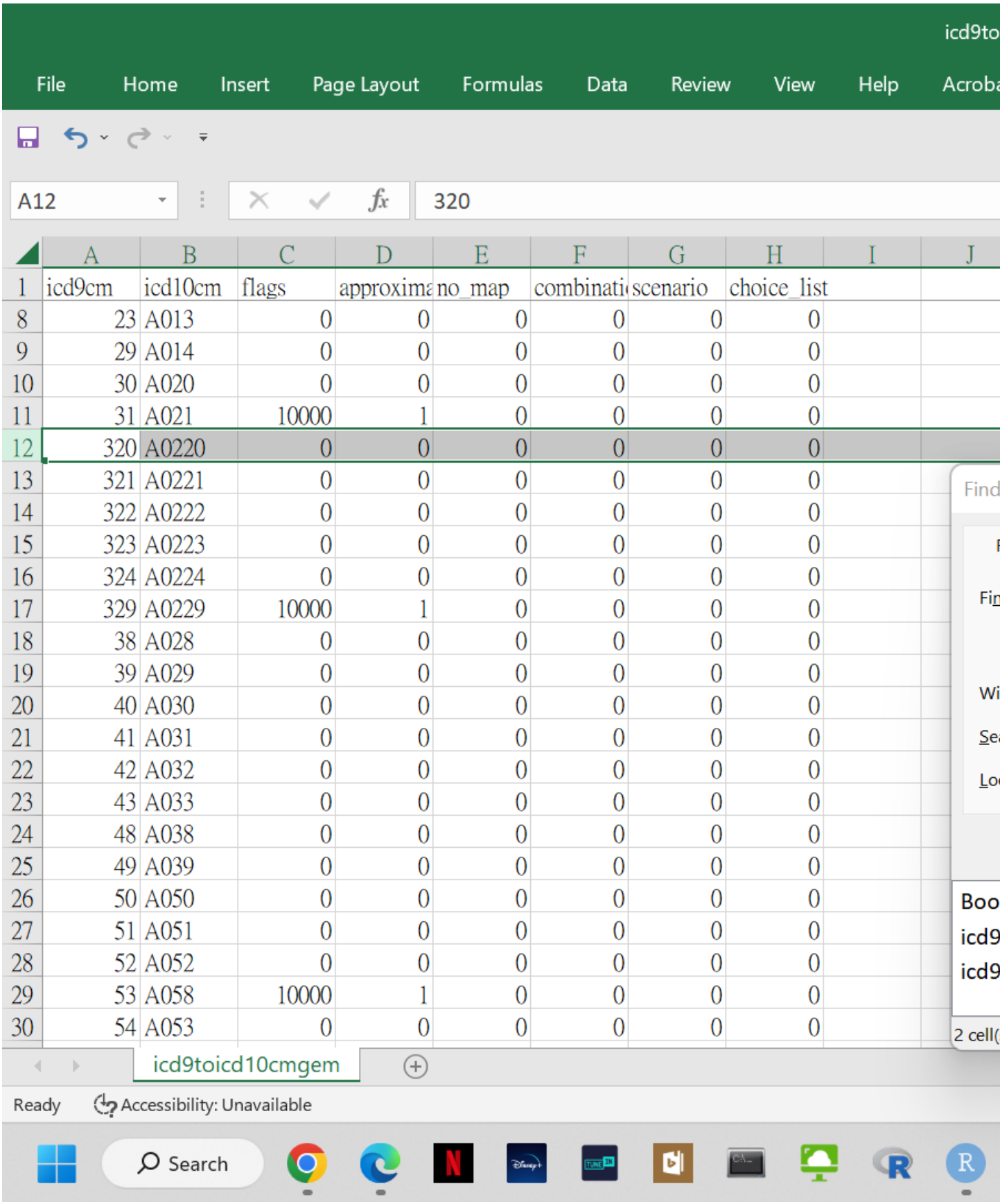
There are 23912 codes in this file, whereas the ICD-9 and ICD-10 codes are not in the correct form. Take the first row for example, there is no ICD-9 code = 10, instead, it should be 001.0, while the corresponding ICD-10 code = A00.0, rather than A000.

```
knitr::include_graphics(file.path(img_path,"icd.png"))
```


ICD-10-CM
Search by description or code (e.g. 'Fractured Fir
We are looking for ways to improve. If you have an su
ICD.Codes / Converters / ICD-9-CM Converter
ICD-9-CM ▶ ICD-10-CM
ICD-9-PCS ▶ ICD-10-PCS
ICD-10-C
feedback
This tool allows you to convert ICD-9-CM codes to their equivalent ICD
Equivalency Mapping (GEM), a crosswalk between the two code standar
Example: Enter the ICD-9-CM code 088.81 (Lyme Disease)
ICD-9-CM
ICD-9-CM
001.0
BILLABLE
ICD-9-CM
001.0 Cholera due to vibrio cholerae
is exa
to the
1
Find More Related ICD-10-CM Codes

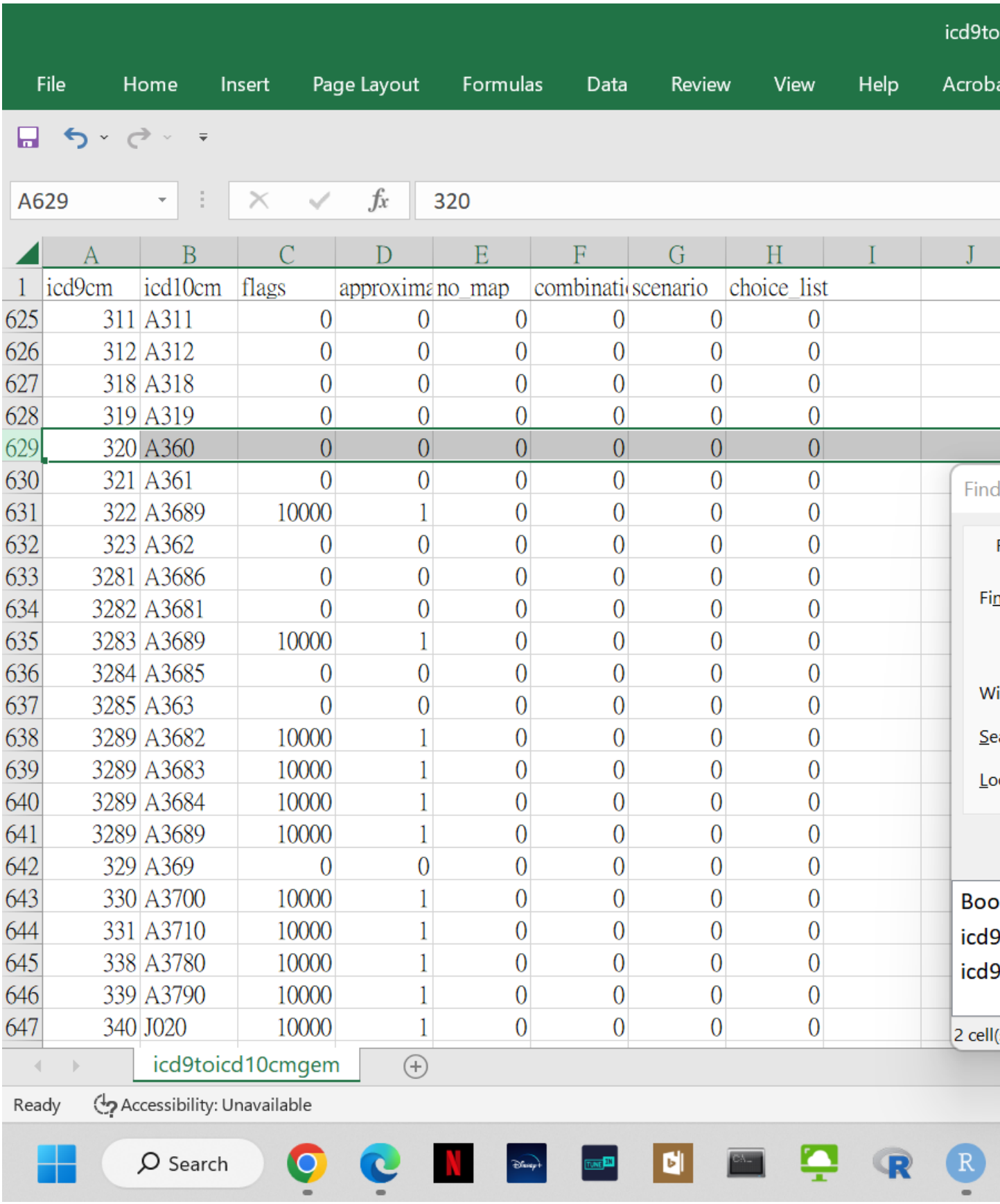
Because of this error, there are identical ICD-9 codes in the file that actually should be different and correspond to different ICD-10 codes. Take ICD-9 = 320 in this file for example:

```
knitr::include_graphics(file.path(img_path,"icd 320a.png"))
```



	A	B	C	D	E	F	G	H	I	J
1	icd9cm	icd10cm	flags	approxima	no_map	combinati	scenario	choice_list		
8	23	A013	0	0	0	0	0	0		
9	29	A014	0	0	0	0	0	0		
10	30	A020	0	0	0	0	0	0		
11	31	A021	10000	1	0	0	0	0		
12	320	A0220	0	0	0	0	0	0		
13	321	A0221	0	0	0	0	0	0		
14	322	A0222	0	0	0	0	0	0		
15	323	A0223	0	0	0	0	0	0		
16	324	A0224	0	0	0	0	0	0		
17	329	A0229	10000	1	0	0	0	0		
18	38	A028	0	0	0	0	0	0		
19	39	A029	0	0	0	0	0	0		
20	40	A030	0	0	0	0	0	0		
21	41	A031	0	0	0	0	0	0		
22	42	A032	0	0	0	0	0	0		
23	43	A033	0	0	0	0	0	0		
24	48	A038	0	0	0	0	0	0		
25	49	A039	0	0	0	0	0	0		
26	50	A050	0	0	0	0	0	0		
27	51	A051	0	0	0	0	0	0		
28	52	A052	0	0	0	0	0	0		
29	53	A058	10000	1	0	0	0	0		
30	54	A053	0	0	0	0	0	0		

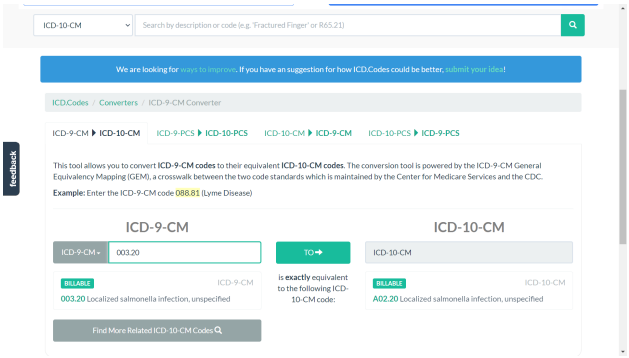
```
knitr::include_graphics(file.path(img_path,"icd 320b.png"))
```



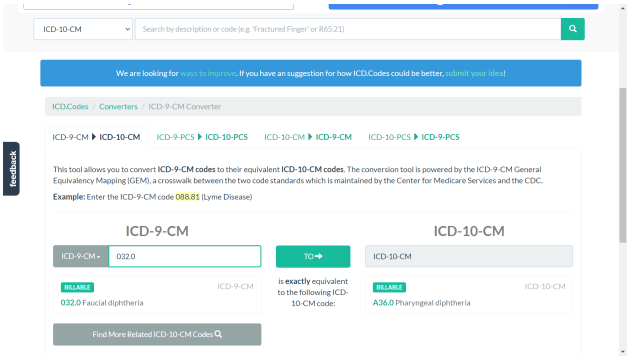
	A	B	C	D	E	F	G	H
1	icd9cm	icd10cm	flags	approxima	no_map	combinati	scenario	choice_list
625	311	A311	0	0	0	0	0	0
626	312	A312	0	0	0	0	0	0
627	318	A318	0	0	0	0	0	0
628	319	A319	0	0	0	0	0	0
629	320	A360	0	0	0	0	0	0
630	321	A361	0	0	0	0	0	0
631	322	A3689	10000	1	0	0	0	0
632	323	A362	0	0	0	0	0	0
633	3281	A3686	0	0	0	0	0	0
634	3282	A3681	0	0	0	0	0	0
635	3283	A3689	10000	1	0	0	0	0
636	3284	A3685	0	0	0	0	0	0
637	3285	A363	0	0	0	0	0	0
638	3289	A3682	10000	1	0	0	0	0
639	3289	A3683	10000	1	0	0	0	0
640	3289	A3684	10000	1	0	0	0	0
641	3289	A3689	10000	1	0	0	0	0
642	329	A369	0	0	0	0	0	0
643	330	A3700	10000	1	0	0	0	0
644	331	A3710	10000	1	0	0	0	0
645	338	A3780	10000	1	0	0	0	0
646	339	A3790	10000	1	0	0	0	0
647	340	J020	10000	1	0	0	0	0

These 320s should be 003.20 and 032.0, while the corresponding ICD-10 codes are A02.20 (not A0220) and A36.0 (not A360):

```
knitr::include_graphics(file.path(img_path,"icd 320 1.png"))
```




```
knitr::include_graphics(file.path(img_path,"icd 320 2.png"))
```



Now you may find out that (1) in the correct ICD-codes, there should be 3 numbers or 1 alphabet with 2 numbers before the decimal; (2) for ICD-9 codes, we may need to add 1 zero or 2 zeros to some of the original codes in our file.

Thus, after checking with the correct codes, I found out that in our file: (1) for the first 1-81 ICD-9 codes, we need to add "00" before the original number, and then add "." after the 3rd number; (2) for the first 82-1211 ICD-9 codes, we need to add "0" before the original number, and then add "." after the 3rd number; (3) for the rest 1212-23912 ICD-9 codes, we need to add "." after the 3rd number; (4) for all the ICD-10 codes, we just need to add "." after the 3rd number.

So let's start data wrangling!

```
## [1] "001.0" "001.1" "001.9" "002.0" "002.1" "002.2"
```

```
## [1] "099.54" "099.55" "099.56" "099.59" "099.8"  "099.9"
```

```
## [1] "100.0"  "100.81" "100.89" "100.9"  "101."   "101."
```

```
## [1] "A00.0"  "A00.1"  "A00.9"  "A01.00" "A01.1"  "A01.2"
```

```
##   icd9cm icd10cm flags approximate no_map combination scenario choice_list
## 1     10    A000     0           0      0           0        0           0
```

```
## 2     11    A001     0           0      0           0        0           0
## 3     19    A009     0           0      0           0        0           0
## 4     20   A0100 10000           1      0           0        0           0
## 5     21    A011     0           0      0           0        0           0
## 6     22    A012     0           0      0           0        0           0
##   icd9cm_n icd10cm_n
## 1    001.0     A00.0
## 2    001.1     A00.1
## 3    001.9     A00.9
## 4    002.0    A01.00
## 5    002.1     A01.1
## 6    002.2     A01.2
```

And never forget those not-matching ones. We know that either in ICD-9 or ICD-10, there should be digits. If it's no digits, it might be "NA" or "No data" or something similar.

```
## [1] 0
```

```
## [1] 425
```

```
##   [1] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
##  [10] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
##  [19] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
##  [28] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
##  [37] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
##  [46] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
##  [55] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
##  [64] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
##  [73] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
##  [82] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
##  [91] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [100] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [109] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [118] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [127] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [136] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [145] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [154] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [163] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [172] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [181] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [190] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [199] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [208] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [217] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [226] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [235] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [244] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [253] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [262] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [271] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [280] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [289] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
```

```
## [298] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [307] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [316] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [325] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [334] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [343] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [352] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [361] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [370] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [379] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [388] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [397] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [406] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [415] "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x" "NoD.x"
## [424] "NoD.x" "NoD.x"
```

So there is no NA in our corrected ICD-9 codes (GOOD!), but there are 425 "NoD.x" in the corrected ICD-10 codes. Let's replace it with NA.

Now let's add the disease description into our new dataset.

```
## [1] "CODE"
## [2] "LONG.DESCRIPTION..VALID.ICD.9.FY2023."
```

```
##   CODE LONG.DESCRIPTION..VALID.ICD.9.FY2023.
## 1 0010        Cholera due to vibrio cholerae
## 2 0011 Cholera due to vibrio cholerae el tor
## 3 0019                   Cholera, unspecified
## 4 0020                          Typhoid fever
## 5 0021                    Paratyphoid fever A
## 6 0022                    Paratyphoid fever B
```

```
## [1] "character"
```

```
## [1] 13521
```

```
##    CODE                                    SHORT.DESCRIPTION
## 1  A000 Cholera due to Vibrio cholerae 01, biovar cholerae
## 2  A001    Cholera due to Vibrio cholerae 01, biovar eltor
## 3  A009                                 Cholera, unspecified
## 4 A0100                          Typhoid fever, unspecified
## 5 A0101                                  Typhoid meningitis
## 6 A0102               Typhoid fever with heart involvement
##                                      LONG.DESCRIPTION
## 1 Cholera due to Vibrio cholerae 01, biovar cholerae
## 2    Cholera due to Vibrio cholerae 01, biovar eltor
## 3                               Cholera, unspecified
## 4                         Typhoid fever, unspecified
## 5                                 Typhoid meningitis
## 6              Typhoid fever with heart involvement
```

```
## [1] FALSE
```

```
##      CODE                                              SHORT.DESCRIPTION
## 41  A0472   Enterocolitis d/t Clostridium difficile, not spcf as recur
## 46   A052               Foodborne Clostridium perfringens intoxication
## 123 A1883                  Tuberculosis of digestive tract organs, NEC
## 177  A288    Oth zoonotic bacterial diseases, not elsewhere classified
## 189  A312     Dissem mycobacterium avium-intracellulare complex (DMAC)
## 217 A3710 Whooping cough due to Bordetella parapertussis w/o pneumonia
##                                                          LONG.DESCRIPTION
## 41  Enterocolitis due to Clostridium difficile, not specified as recurrent
## 46    Foodborne Clostridium perfringens [Clostridium welchii] intoxication
## 123       Tuberculosis of digestive tract organs, not elsewhere classified
## 177  Other specified zoonotic bacterial diseases, not elsewhere classified
## 189         Disseminated mycobacterium avium-intracellulare complex (DMAC)
## 217       Whooping cough due to Bordetella parapertussis without pneumonia
```

```
## [1] 72836
```

There are two disease descriptions in the icd_cm_10d file, I will use the more detailed one (the long description). And in these two files, the codes also should be corrected just like above. After correction, we can join the tables.

```
## [1] "001.0" "001.1" "001.9" "002.0" "002.1" "002.2"
```

```
##   CODE LONG.DESCRIPTION..VALID.ICD.9.FY2023. icd9cm_n
## 1 0010        Cholera due to vibrio cholerae    001.0
## 2 0011 Cholera due to vibrio cholerae el tor    001.1
## 3 0019                   Cholera, unspecified    001.9
## 4 0020                         Typhoid fever    002.0
## 5 0021                   Paratyphoid fever A    002.1
## 6 0022                   Paratyphoid fever B    002.2
```

```
## [1] "A00.0"  "A00.1"  "A00.9"  "A01.00" "A01.01" "A01.02"
```

```
##    CODE                                    SHORT.DESCRIPTION
## 1  A000 Cholera due to Vibrio cholerae 01, biovar cholerae
## 2  A001    Cholera due to Vibrio cholerae 01, biovar eltor
## 3  A009                                 Cholera, unspecified
## 4 A0100                            Typhoid fever, unspecified
## 5 A0101                                  Typhoid meningitis
## 6 A0102               Typhoid fever with heart involvement
##                                     LONG.DESCRIPTION icd10cm_n
## 1 Cholera due to Vibrio cholerae 01, biovar cholerae     A00.0
## 2    Cholera due to Vibrio cholerae 01, biovar eltor     A00.1
## 3                                Cholera, unspecified     A00.9
## 4                          Typhoid fever, unspecified    A01.00
## 5                                Typhoid meningitis    A01.01
## 6              Typhoid fever with heart involvement    A01.02
```

```
##  [1] "icd9cm"      "icd10cm"     "flags"       "approximate" "no_map"
##  [6] "combination" "scenario"    "choice_list" "icd9cm_n"    "icd10cm_n"
```

```
## [1] "CODE"
## [2] "LONG.DESCRIPTION..VALID.ICD.9.FY2023."
## [3] "icd9cm_n"
```

```
##   icd9cm icd10cm flags approximate no_map combination scenario choice_list
## 1     10    A000     0           0      0           0        0           0
## 2     11    A001     0           0      0           0        0           0
## 3     19    A009     0           0      0           0        0           0
## 4     20   A0100 10000           1      0           0        0           0
## 5     21    A011     0           0      0           0        0           0
## 6     22    A012     0           0      0           0        0           0
##   icd9cm_n icd10cm_n CODE LONG.DESCRIPTION..VALID.ICD.9.FY2023.
## 1    001.0     A00.0 0010        Cholera due to vibrio cholerae
## 2    001.1     A00.1 0011 Cholera due to vibrio cholerae el tor
## 3    001.9     A00.9 0019                  Cholera, unspecified
## 4    002.0    A01.00 0020                         Typhoid fever
## 5    002.1     A01.1 0021                   Paratyphoid fever A
## 6    002.2     A01.2 0022                   Paratyphoid fever B
```

```
##   icd9cm icd10cm flags approximate no_map combination scenario choice_list
## 1     10    A000     0           0      0           0        0           0
## 2     11    A001     0           0      0           0        0           0
## 3     19    A009     0           0      0           0        0           0
## 4     20   A0100 10000           1      0           0        0           0
## 5     21    A011     0           0      0           0        0           0
## 6     22    A012     0           0      0           0        0           0
##   icd9cm_n icd10cm_n CODE.x                      ICD9 Description CODE.y
## 1    001.0     A00.0   0010        Cholera due to vibrio cholerae   A000
## 2    001.1     A00.1   0011 Cholera due to vibrio cholerae el tor   A001
## 3    001.9     A00.9   0019                  Cholera, unspecified   A009
## 4    002.0    A01.00   0020                         Typhoid fever  A0100
## 5    002.1     A01.1   0021                   Paratyphoid fever A   A011
## 6    002.2     A01.2   0022                   Paratyphoid fever B   A012
##                                   SHORT.DESCRIPTION
## 1 Cholera due to Vibrio cholerae 01, biovar cholerae
## 2    Cholera due to Vibrio cholerae 01, biovar eltor
## 3                              Cholera, unspecified
## 4                        Typhoid fever, unspecified
## 5                               Paratyphoid fever A
## 6                               Paratyphoid fever B
##                                      LONG.DESCRIPTION
## 1 Cholera due to Vibrio cholerae 01, biovar cholerae
## 2    Cholera due to Vibrio cholerae 01, biovar eltor
## 3                              Cholera, unspecified
## 4                        Typhoid fever, unspecified
## 5                               Paratyphoid fever A
## 6                               Paratyphoid fever B
```

```
##  [1] "icd9cm"            "icd10cm"           "flags"
##  [4] "approximate"       "no_map"            "combination"
##  [7] "scenario"          "choice_list"       "icd9cm_n"
## [10] "icd10cm_n"         "CODE.x"            "ICD9 Description"
## [13] "CODE.y"            "SHORT.DESCRIPTION" "LONG.DESCRIPTION"
```

```
##  [1] "icd9cm"            "icd10cm"           "flags"
##  [4] "approximate"       "no_map"            "combination"
##  [7] "scenario"          "choice_list"       "icd9cm_n"
```

```
## [10] "icd10cm_n"         "CODE.x"            "ICD9 Description"
## [13] "CODE.y"            "SHORT.DESCRIPTION" "ICD10 Description"
```

Lastly, we need to add some warning signs because sometimes ICD-9 codes cannot exactly match with the ICD-10 codes. Notice those flags? When the flag = 0, it means we can find the exact ICD-10 codes; when the flag = 10000, it means we can only find the most similar meaning ICD-10 codes; when the flag = 11000, sadly there's no such ICD-10 codes. This is our last step of data wrangling!

```
##   icd9cm icd10cm flags icd9cm_n icd10cm_n                     ICD9 Description
## 1     10    A000     0    001.0     A00.0        Cholera due to vibrio cholerae
## 2     11    A001     0    001.1     A00.1 Cholera due to vibrio cholerae el tor
## 3     19    A009     0    001.9     A00.9                   Cholera, unspecified
## 4     20   A0100 10000    002.0    A01.00                          Typhoid fever
## 5     21    A011     0    002.1     A01.1                    Paratyphoid fever A
## 6     22    A012     0    002.2     A01.2                    Paratyphoid fever B
##                                    ICD10 Description              matching
## 1 Cholera due to Vibrio cholerae 01, biovar cholerae      Exactly matching
## 2    Cholera due to Vibrio cholerae 01, biovar eltor      Exactly matching
## 3                               Cholera, unspecified      Exactly matching
## 4                          Typhoid fever, unspecified Approxiately matching
## 5                                Paratyphoid fever A      Exactly matching
## 6                                Paratyphoid fever B      Exactly matching
```

```
##   icd9cm_n icd10cm_n                     ICD9 Description
## 1    001.0     A00.0        Cholera due to vibrio cholerae
## 2    001.1     A00.1 Cholera due to vibrio cholerae el tor
## 3    001.9     A00.9                   Cholera, unspecified
## 4    002.0    A01.00                          Typhoid fever
## 5    002.1     A01.1                    Paratyphoid fever A
## 6    002.2     A01.2                    Paratyphoid fever B
##                                    ICD10 Description              matching
## 1 Cholera due to Vibrio cholerae 01, biovar cholerae      Exactly matching
## 2    Cholera due to Vibrio cholerae 01, biovar eltor      Exactly matching
## 3                               Cholera, unspecified      Exactly matching
## 4                          Typhoid fever, unspecified Approxiately matching
## 5                                Paratyphoid fever A      Exactly matching
## 6                                Paratyphoid fever B      Exactly matching
```

Results

Finally, we can start to search the matching ICD-10 codes! For example, if I want to convert ICD-9 = "E93.00","003.1","032.0":

```
icd_cm_final |>
  filter(icd9cm_n %in% c("E93.00","003.1","032.0")) |>
  summarise(icd9 = icd9cm_n, icd10 = icd10cm_n, footnote = matching)
```

```
##     icd9 icd10              footnote
## 1  003.1 A02.1 Approxiately matching
## 2  032.0 A36.0      Exactly matching
## 3 E93.00  <NA>           No matching
```

We can find the corresponding ICD-10 codes along with their matching extent in the summarize (footnote). And we can directly copy the corresponding ICD-10 codes into our word files or slides by using codes below:

```
exp1 = icd_cm_final |>
  filter(icd9cm_n %in% c("E93.00","003.1","032.0")) |>
  summarise(icd9 = icd9cm_n, icd10 = icd10cm_n, footnote = matching) |>
  pull(icd10)

exp1 |>
  paste(collapse = " ") |>
  str_replace_all(" ", ", ")
```

```
## [1] "A02.1, A36.0, NA"
```

By using the codes below, we can directly copy a number of ICD codes from word files and paste them into " " and search!! No need to spend time to further separate them with " "!

```
exp2 = c("E93.00, 003.1, 032.0")
e2 = unlist(str_split(exp2, ", "))

icd_cm_final |>
  filter(icd9cm_n %in% c(e2[1:length(e2)])) |>
  summarise(icd9 = icd9cm_n, icd10 = icd10cm_n, footnote = matching) |>
  pull(icd10)
```

```
## [1] "A02.1" "A36.0" NA
```

This is what I want!

##Part 2 Regression model

As mentioned above, I want to build a regression model to predict the possibility of opioids overdose.

Firstly, the mean of opioids overdose rate is 8. I defined opioids overdose rate >8 as more likely to have opioids overdose, and <=8 as less likely, which becomes overdose_p in the data.

Secondly, Using overdose_p as outcome, putting all the possible covariates into the model as our full model (logistic regression).

Though the performance of the full model is good (AIC= 417.32, AUC = 0.99), because there are many covariates related to finance, such as thealthspend (total health spend) and totalrealhcspend (total real hospital and clinics spend), considering collinearity and simplicity/parsimony, stateid (state), totalrealhcspend (total health spend), labor_participation_pct (labor or not), grad_hs_pct (education), and cpi (consumer price index) are kept in my final model (s_model).

The performance of this final model is nice, with AIC: 551.68 and AUC = 0.9755.

```
op = read.csv("c:\\Users\\user\\Downloads\\Opioid.csv")
head(op)
```

```
##     state stateid year t mcare_millions medicaid_spend_actual medicaidspending
## 1 Alabama       1 2000 0           3690                2719.15          2.7e+09
## 2 Alabama       1 2001 1           4065                2901.74          2.9e+09
## 3 Alabama       1 2002 2           4394                3115.61          3.1e+09
## 4 Alabama       1 2003 3           4756                3505.83          3.5e+09
```

```
## 5 Alabama       1 2004 4           5274               3664.08          3.7e+09
## 6 Alabama       1 2005 5           5698               3864.14          3.9e+09
##   thealthspend totalrealhcspend overdoses population overdose_rate
## 1         6410             9410        43    4500000         0.956
## 2         6970             9860        57    4500000         1.270
## 3         7510            10500        71    4500000         1.580
## 4         8260            11300        49    4500000         1.090
## 5         8940            12000        83    4500000         1.840
## 6         9560            12400        80    4600000         1.740
##   mdhhincomereal stategdpml realstategdp unemployment_pct
## 1          35424     119242     175097.8              4.6
## 2          35160     122449     173337.7              5.1
## 3          37603     127792     178858.3              5.9
## 4          37255     133739     182443.0              6.0
## 5          36629     146525     196107.7              5.7
## 6          37150     155970     202728.3              4.5
##   labor_participation_pct insured_pct grad_hs_pct is_manufacturing_state
## 1                    60.3        87.5        77.5                      1
## 2                    59.2        87.6        80.2                      1
## 3                    58.2        87.8        78.9                      1
## 4                    58.2        87.5        79.9                      1
## 5                    58.5        88.0        82.4                      1
## 6                    58.9        86.0        80.9                      1
##   post_recession   cpi
## 1              0 168.8
## 2              0 175.1
## 3              0 177.1
## 4              0 181.7
## 5              0 185.2
## 6              0 190.7
```

```
summary(op$overdose_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.365   4.170   6.410   8.173  10.000  46.300
```

```
op$overdose_p = ifelse(op$overdose_rate >8, 1,0)
names(op)
```

```
##  [1] "state"                   "stateid"
##  [3] "year"                    "t"
##  [5] "mcare_millions"          "medicaid_spend_actual"
##  [7] "medicaidspending"        "thealthspend"
##  [9] "totalrealhcspend"        "overdoses"
## [11] "population"              "overdose_rate"
## [13] "mdhhincomereal"          "stategdpml"
## [15] "realstategdp"            "unemployment_pct"
## [17] "labor_participation_pct" "insured_pct"
## [19] "grad_hs_pct"             "is_manufacturing_state"
## [21] "post_recession"          "cpi"
## [23] "overdose_p"
```

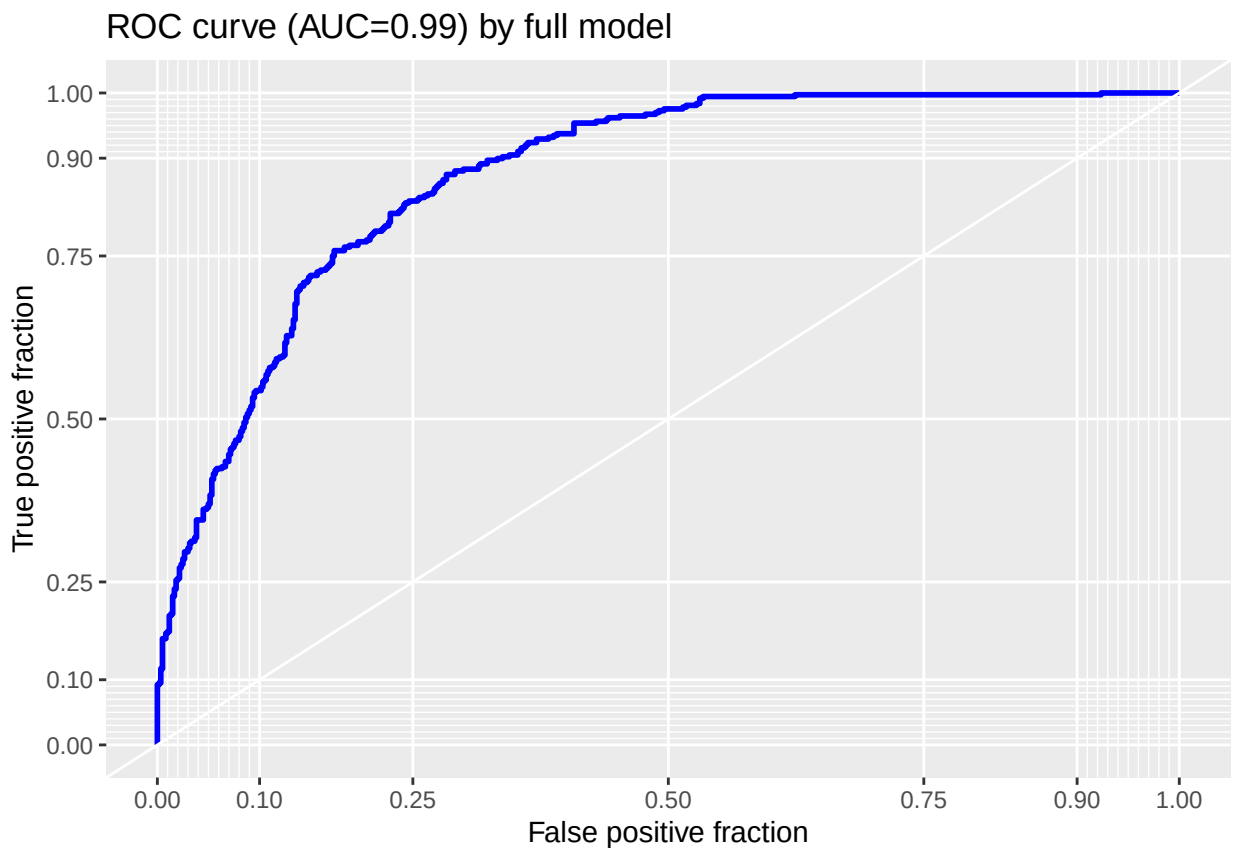
```
library(ggplot2)
library(tidyverse)
library(caret)
library(leaps)
library(MASS)
library(pROC)
library(plotROC)

full_model = glm(overdose_p ~ stateid + mcare_millions + medicaid_spend_actual + medicaidspending + thea

summary(full_model)
```

```
##
## Call:
## glm(formula = overdose_p ~ stateid + mcare_millions + medicaid_spend_actual +
##     medicaidspending + thealthspend + totalrealhcspend + mdhhincomereal +
##     stategdpml + realstategdp + unemployment_pct + labor_participation_pct +
##     insured_pct + grad_hs_pct + is_manufacturing_state + post_recession +
##     cpi, family = "binomial", data = op)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3053  -0.6741  -0.2413   0.6604   2.9909
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              6.233e+00  4.451e+00   1.400 0.161436
## stateid                  5.309e-02  6.542e-03   8.115 4.85e-16 ***
## mcare_millions           4.370e-03  3.166e-03   1.380 0.167549
## medicaid_spend_actual    4.592e-03  3.222e-03   1.425 0.154030
## medicaidspending        -2.700e-10  6.913e-10  -0.391 0.696140
## thealthspend            -3.373e-03  3.120e-03  -1.081 0.279562
## totalrealhcspend        -7.857e-04  2.282e-04  -3.444 0.000573 ***
## mdhhincomereal           1.561e-04  1.841e-05   8.475  < 2e-16 ***
## stategdpml              -4.130e-05  1.415e-05  -2.919 0.003512 **
## realstategdp             2.916e-05  1.237e-05   2.357 0.018433 *
## unemployment_pct         2.240e-01  7.051e-02   3.177 0.001487 **
## labor_participation_pct -1.519e-01  4.046e-02  -3.754 0.000174 ***
## insured_pct             -4.179e-02  2.916e-02  -1.433 0.151847
## grad_hs_pct             -7.048e-02  4.330e-02  -1.628 0.103591
## is_manufacturing_state  -7.297e-01  2.409e-01  -3.029 0.002455 **
## post_recession          -8.966e-01  4.494e-01  -1.995 0.046036 *
## cpi                      1.496e-02  1.156e-02   1.294 0.195680
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1286.74  on 968  degrees of freedom
## Residual deviance:  833.02  on 952  degrees of freedom
## AIC: 867.02
##
## Number of Fisher Scoring iterations: 6
```

```
roc_curve_f = roc(op$overdose_p ,predict(full_model, type = c("response")))
roc_curve_f$auc
```

```
## Area under the curve: 0.8729
```

```
ggplot(op, aes(m = predict(full_model, type = c("response")), d = overdose_p))+ geom_roc(n.cuts = 0, lab
```




```
s_model = glm(formula = overdose_p ~ stateid + totalrealhcspend + labor_participation_pct +
                grad_hs_pct + cpi, data = op)
summary(s_model)
```

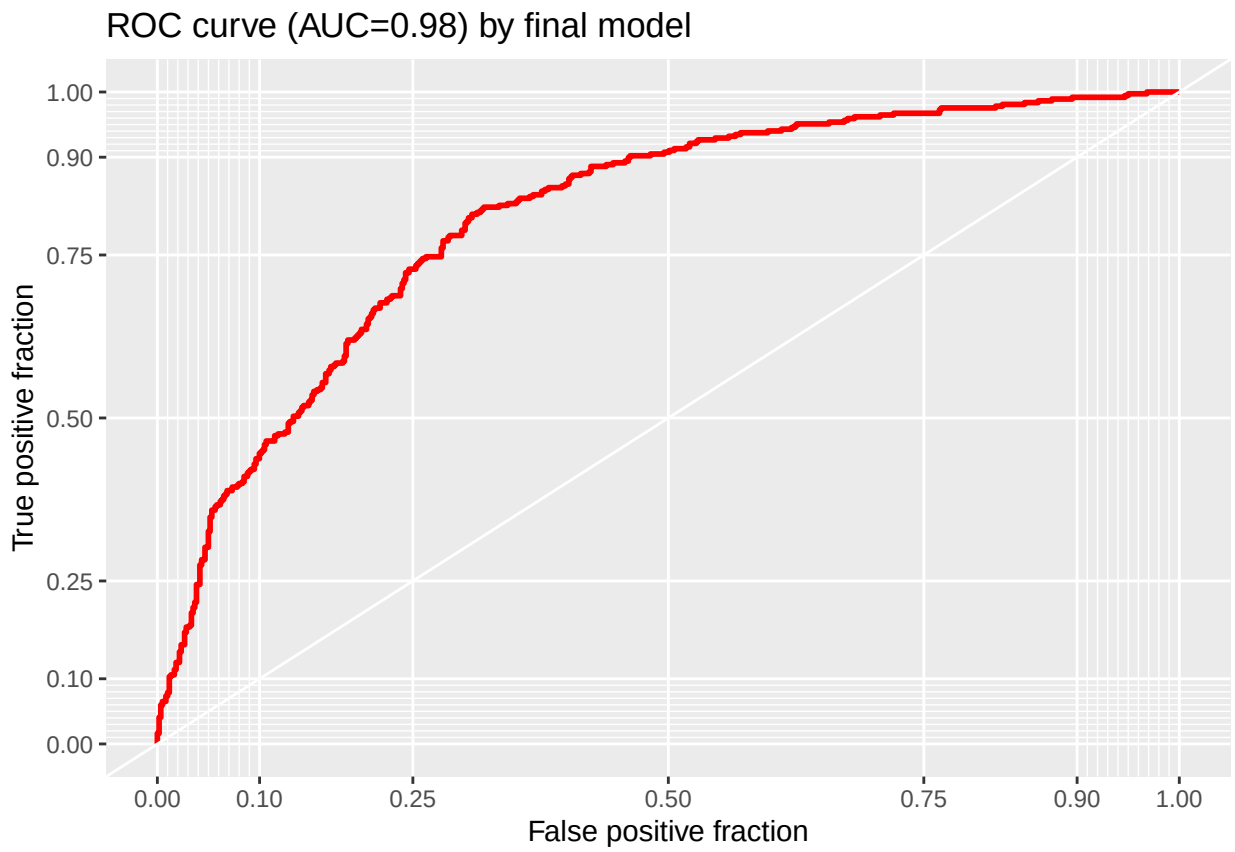
```
##
## Call:
## glm(formula = overdose_p ~ stateid + totalrealhcspend + labor_participation_pct +
##     grad_hs_pct + cpi, data = op)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.92288  -0.31954  -0.08052   0.35230   1.08393
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.089e+00  3.403e-01  -3.200  0.00142 **
## stateid                  5.956e-03  9.182e-04   6.486 1.40e-10 ***
```

```
## totalrealhcspend        -2.016e-06  6.207e-07  -3.247  0.00120 **
## labor_participation_pct -2.516e-02  4.541e-03  -5.541 3.87e-08 ***
## grad_hs_pct              1.663e-02  5.742e-03   2.896  0.00386 **
## cpi                      6.989e-03  7.577e-04   9.225  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1744889)
##
##     Null deviance: 228.24  on 968  degrees of freedom
## Residual deviance: 168.03  on 963  degrees of freedom
## AIC: 1066.1
##
## Number of Fisher Scoring iterations: 2
```

```
roc_curve = roc(op$overdose_p ,predict(s_model, type = c("response")))
roc_curve$auc
```

```
## Area under the curve: 0.8079
```

```
ggplot(op, aes(m = predict(s_model, type = c("response")), d = overdose_p))+ geom_roc(n.cuts = 0, labels
```



##Conclusion

In part 1, I did data wrangling to convert ICD-9 to ICD-10. In part 2, I built a logistic regression model to predict the possibility of opioids overdose. I think both parts are quite successful. If I have more time, I would like to apply machine learning skills in part 2.