**DEVELOPING CLINICAL CODE LISTS**

**Contents Pg**

**1** General Principles 3

**2** Training on developing code lists for research 3

**3** Defining the clinical feature of interest 3

**4** Creating a list of synonyms for a condition 4

**5** Identifying existing code lists 4

**6** Different medical coding systems used in the CPRD, linked data and code browsers 5

6.1 ICD-10 5

6.2 Read Codes 5

6.3 SNOMED CT 5

6.4 BNF codes for products 6

6.5 Dm+d codes 7

**7** Searching for codes using the CPRD Code Browsers 7

 7.1 Available guides to using the Code Browsers 7

 7.2 Conducting a search 7

 7.3 Searching the Code Browsers for medication codes 8

 7.4 Processing codes used in the CPRD 9

 7.5 Searching the dictionary files directly 10

**8** Matching medcodes/medcodeids to Read or SNOMED codes in pre-existing code lists 10

**9** Reviewing code lists 10

**10** Considering the number of clinical events 10

**11** Validating code lists 10

References 11

Appendix 1. Overview of medical coding systems 12

 ICD-10 12

 Read Codes 13

 SNOMED CT 14

 Converting between ICD-10, Read and SNOMED CT 16

Appendix 2. Example R Script to search the dictionary for codes based on words in Read Term 17

Appendix 3. Example R Script to match medcodes to Read Codes using the dictionary file 19

Appendix 4. Example R Script to match medcodeids to SNOMED codes using the dictionary file 20

**Developing a code list**

**1. General principles**

*“Clinical code set construction is the process of assembling a set of clinical codes that represent a single clinical concept such as a diagnosis, a procedure, an observation or a medication.”* (Williams et al. 2017)[1]

Creation of code lists (also referred to as code sets or value sets) will generally be the first step in analyses using routinely collected data. Any errors in code lists resulting from the omission of important codes or inclusion of irrelevant codes can result in selection biases that can have a major impact on analyses[2]. For this reason the RECORD statement on the Reporting of Studies Conducted using Observational Routinely Collected Health Data[3] states that all code lists and algorithms used to derive any variables used in analyses should be reported, and a number of clinical code list repositories are available to do this.

In general it’s recommended by CPRD and others to develop (and document) a reusable search strategy to develop code lists - this allows updating of code lists for future analyses with new codes using the same strategy. If both CPRD Gold and CPRD Aurum are being used, the same search strategy can be applied to each database.

General approaches/steps that can be used in the creation of code lists[1,4] include:

* Clearly defining the clinical feature (e.g. symptom, disease, medication for a particular condition) of interest
* Creation of a list of synonyms for the condition of interest
* Identifying and reusing/updating an existing code list
* Utilizing the hierarchical nature of coding terminologies

**2. Training on developing code lists for research**

The CPRD training on MHRA is very useful and includes information on developing code lists,

[UpsideLMS - Learner Area (mhra.gov.uk)](http://learning.mhra.gov.uk/mhra/lr_register.jsp?clientid=4)

(select MHRA as the client, click learner login, then there is a link to register for the training in the log in box)./.

**3. Defining the clinical feature of interest**

Ideally, inclusion and exclusion criteria for symptoms and diseases should be defined a priori. eg. should prescriptions for relevant medications or test results be used as well as diagnostic codes to identify patients with a condition, should a symptom of appetite loss include patients with anorexia nervosa, should lung cancer also include mesothelioma? If possible a clinician should be involved in this process.

The International Classification of Primary Care (ICPC) can be useful for this – it provides definitions of symptoms and diagnoses, lists exclusions from the definitions, and also gives ICD-10 and SNOMED codes. A link to an ICPC-3 browser is given below.

[ICPC-3 Browser](https://browser.icpc-3.info/)

It’s also useful to review any available definitions used in publications which have looked at the same symptoms/diseases.

**4. Creating a list of synonyms for the condition of interest**

A number of clinical code browsers, including the ICPC-3 browser and the NHS SNOMED CT browser (see Appendix 1) can be used to find synonyms for particular conditions or symptoms that can be used in any code searches. The Code Browsers supplied by CPRD can also help with finding synonyms eg. by looking at related terms in the the Read Code hierarchy. If a pre-existing code list is available, this can also be useful for identifying synonyms. If possible check the list of synonyms with a clinican.

Synonyms can also be identified using the United Medical Language System Metathesaurus browser which is available here:

[UMLS Metathesaurus Browser (nih.gov)](https://uts.nlm.nih.gov/uts/umls/home?_gl=1*ql91sd*_ga*MTY1ODIxNzg5Ni4xNjUxMzMyOTI3*_ga_P1FPTH9PL4*MTY2NzMxNDg2OS40LjEuMTY2NzMxNTY2Ni4wLjAuMA..)

**5. Identifying existing code lists**

Existing code lists produced by other researchers can be identified, either by a Google search and/or by checking in a number of repositories including:

[LSHTM Data Compass](https://datacompass.lshtm.ac.uk/)

[Phenotype Library | HDRUK (healthdatagateway.org)](https://phenotypes.healthdatagateway.org/)

[ClinicalCodes Repository (man.ac.uk)](https://clinicalcodes.rss.mhs.man.ac.uk/)

[CPRD @ Cambridge - Codes Lists (GOLD) - Primary Care Unit](https://www.phpc.cam.ac.uk/pcu/research/research-groups/crmh/cprd_cam/codelists/v11/)

[OpenCodelists](https://www.opencodelists.org/)

The code lists used in the business rules for the Quality and Outcomes Framework are available here:

[Quality and Outcomes Framework (QOF) business rules v45.0 2020-2021 baseline release - NHS Digital](https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof/quality-and-outcome-framework-qof-business-rules/quality-and-outcomes-framework-qof-business-rules-v45.0-2020-2021-baseline-release)

(code lists are in the Excel expanded cluster list).

The ATLAS tool from OHDSI also has a number of SNOMED code lists (under “Concepts”)

<https://atlas-demo.ohdsi.org/#/home>

Code lists for the same condition can vary widely in what is included. The list of codes and as a result the number of cases picked up by the list can differ substantially depending on how the condition/exposure was defined in that particular study and the relative importance of having a very broad definition which will pick up all possible cases versus a narrower more specific code list. Published code lists may also contain errors, such as codes indicating the absence of a condition[4]. As a result, it is best, if possible to:

* review multiple code lists if available, considering the clinical definition used in each study
* check any codes included
* supplement the results with your own code searches

**6. Different medical coding systems used in the CPRD, linked data and code browsers**

The CPRD and linked data are based on a number of medical coding systems including ICD-10, Read and SNOMED CT. A brief overview of these coding systems and where to find various browsers that can be used to search them is given in Appendix 1.

6.1 ICD-10

This coding system is used for the linked data (hospital episode statistics and the cancer registry data tumour file).

NB. The format of the ICD-10 codes in the hospital episode statistics data files includes a decimal point, while the ICD-10 codes in the cancer registry tumour file do not include the decimal point.

i.e. correct format for an icd10 code list for hospital episodes statistics would be e.g ("C33","C34.0", "C34.1", "C34.2", "C34.3"), while for the cancer registry data it would be ("C33","C340", "C341", "C342", "C343").

In the documentation for the tumour file it recommends looking at the variable on coding system to interpret the codes for data prior to 1995 (all data from 1995 has been converted to icd10 codes. However, in practice this variable is set to missing for all tumour diagnosis dates prior to 1995. A reasonable assumption is that this data uses ICD-9 codes.

ICD-10 codes can be converted to ICD-9 for the cancer registry data prior to 1995 at the following site.

[ICD-10-CM to ICD-9-CM](https://icd.codes/convert/icd10-to-icd9-cm)

6.2 Read Codes

Read Codes are the basis of the “medcode” coding system used in the **CPRD Gold event files** (i.e. the clinical and referral files). The Read codes can be mapped to medcodes using the Gold Code Browser (or the medical.txt and product.txt dictionaries found in the Code Browser folder).

NB. Read hierarchies will include codes for the **absence** of a condition within them, e.g. the first term in the Read hierarchy after 171..00 “Cough” is 1711.00 “No Cough”, so it is important that these terms are excluded from any code lists aiming to detect the presence of a condition.

6.3 SNOMED CT codes

SNOMED CT is the current standard for clinical terms used in the National Health Service.

A SNOMED concept represents a unique clinical meaning, associated with a unique Concept ID number. Each SNOMED concept may have a number of alternative descriptions associated with it, each of which has its own SNOMED description id. For example, the different descriptions associated with the SNOMED concepts for “Asthma attack” and “Myocardial infarction” are shown below.





6.4 BNF codes for products

A useful explanation of the structure of BNF codes is available here:

[Prescribing Data: BNF Codes | Bennett Institute for Applied Data Science (ox.ac.uk)](https://www.bennett.ox.ac.uk/blog/2017/04/prescribing-data-bnf-codes/#:~:text=The%20codes%20are%20in%20a%20hierarchy%3A%201%20The,give%20more%20detailed%20information%20about%20a%20drug.%20)

The codes are alphanumeric, with the first 7 digits representing the chapter, section, paragraph and subparagraph of the BNF the medicine belongs to, and the rest of the code indicating the chemical substance and information about the product and formulation. The BNF codes given in the code browsers supplied with the CPRD include information up to the sub-paragraph level.

A searchable directory of the BNF is available here.

<https://openprescribing.net/bnf/>

Eg., a search for “selective serotonin re-uptake inhibitors” (SSRIs) brings up this link with a list of chemicals and their BNF codes at the bottom of the page

[BNF 4.3.3: Selective serotonin re-uptake inhibitors | OpenPrescribing](https://openprescribing.net/bnf/040303/)



6.5 Dm+d codes

Dm+d is the Dictionary of Medicines and Devices – the NHS standard used to uniquely identify medicines and medical devices. It is one of the search fields given for products in the CPRD Browsers. More information dm+d is available here.

[Dictionary of medicines and devices (dm+d) | NHSBSA](https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/dictionary-medicines-and-devices-dmd)

**7. Searching for codes using the code browsers provided by CPRD**

7.1 Available guides to using the Code Browsers

A guide to using the CPRD code browser can be found here:

[CPRD Code Browser user guide September 2022](https://cprd.com/sites/default/files/2022-09/CPRD%20Code%20Browser%20quick%20user%20guide.pdf)

There is also further information on using the guide in the “index” HTML file is the Resources/Help folder of the Code Browser Folder.

7.2 Conducting a search

CPRD Gold uses Medical Codes/medcodes (linked to Read codes) and CPRD Aurum uses Medical Code ID/medcodeid (linked to SNOMED codes and in some cases also Read codes). The Aurum browser has columns for both “Read Code” and “Cleansed Read Code” – “Cleansed Read Code” in the Aurum Browser is in the same format as the “Read Code” given in the Gold Browser.

When searching the browser make sure that the correct search fields (eg. Read Term, Read Code or medcode for CPRD Gold) and dictionary (medical or product) are selected. If searching Read Terms for a keyword it’s best to use enclose it with “\*” or only exact matches with the search term will be returned, eg.:



The results of a search for haemoptysis in the Read Term are shown below.



Any relevant terms under “Found terms” can be selected and added to the Selected terms box using **Tools , Add Terms**.

Looking at the list of terms returned may identify additional synonyms, eg. the list of Read Terms returned above shows that “blood in sputum” is a synonym for haemoptysis, so a search for \*sputum\* in the Read Term may return some additional relevant codes.

Related Read Terms in the hierarchy can also be searched for using “\*”. E.g., the list above suggests Read codes beginning R063 are likely to be related to haemoptysis so a search on R063\* may find additional relevant codes.



(In general, if searching on Read Code it is better to finish the search term with “\*” as varying numbers of “0” may be included at the end of the term depending on the format used in different codelists.)

Once all relevant codes have ben added to the selected terms list, the final list can be exported as a text file using **File, Export**.

7.3 Searching the browsers for medication codes

Information on prescribed medications is given in the therapy files in CPRD Gold, and in the drug issue files in CPRD Aurum. Different medications are identified by “prodcode” in the CPRD Gold therapy files, and by “prodcodeid” in the CPRD Aurum drug issue files.

The format of the BNF codes given in the BNF searchable directory described earlier differs slightly from that in the CPRD browsers and product dictionary. To search for the equivalent prodcodeids to BNF codes in the Aurum Browser omit the first 0 and use an \* in place of the final two characters. Eg for the list of BNF codes for SSRIs shown earlier this would be. 403030\*.



To do the equivalent to find prodcodes in the Gold Browser, include the first 0 (i.e. 0403030\*).



Not all medications in the dictionaries supplied have a BNF code associated with them so searches should also be conducted using the Drug Substance Name and the Product Name.

Dm+d codes are one of the search fields provided in both the CPRD Gold and CPRD Aurum product dictionaries and some code lists for medicines are provided in this format (e.g on opencodelists). However, the format of the dm+d codes in these codelists does not generally agree with that provided in the browsers/dictionaries provided by CPRD, so they are probably not very useful for generating codelists to use with the CPRD data.

7.4 Processing codes used in the CPRD

CPRD strongly recommends that any long numeric codes in the datasets are processed as t**ext/string variables** rather than integers. This is because some of these codes are very long, resulting in integers which are too large for many standard software packages to deal with. Treating them as text means these codes may be converted to numeric approximations resulting in errors. This is particularly important for the numeric codes provided with CPRD Aurum.

7.5 Searching the dictionary files provided by CPRD directly

It is also possible to read the dictionary files provided in the Code Browser folders for Gold (medical.txt/product.txt) and Aurum (CPRDAurumMedical.txt/CPRDAurumProduct.txt) into a package of your choice and search them directly.

Appendix 1 contains some example R script for code searching the dictionaries based on words in the Read term.

**8. Matching medcodes/medcodeids to Read or SNOMED codes in pre-existing code lists**

Many of the code lists available in the various repositories only include Read codes or SNOMED IDs (rather than medcode/medcodeid used in CPRD Gold and Aurum). The corresponding medcode/medcodeid can be found using the code browser for each Read code/SNOMED ID individually. Alternatively If the format of the Read codes in these lists agrees with the format in the Code Browsers it can be possible to match them directly with the medcode/medcodeids using the dictionary files provide in the same folder as the Code Browser. Some example R code to do this for lists of Read Codes and SNOMED IDs is provided in Appendix 3 and 4 respectively.

**9. Reviewing code lists**

The list of codes obtained by searching using one of the methods above needs to be carefully reviewed to ensure all identified codes are suitable to define the clinical concept. For example, a search on “\*cough\*” in the CPRD Gold code browser returns codes for “Night cough absent”, and a search for “\*diabetes\*” returns a number of non-diagnostic screening codes. If possible, a clinician should be involved in reviewing the code list. Any published code lists used should also be checked for possible errors. If published code lists using Read Codes or SNOMED codes are matched with medcodes/medcodeids using the dictionary files, it is worth retaining the Read Term/Description in the dictionary in the matched output, to check that no codes have been included erroneously.

**10. Considering the number of clinical events.**

The CPRD Gold code browser gives information on the total number of clinical and referral events associated with each medcode in the CPRD Gold datasets, and if the set of patient ids for inclusion has been defined it is also worth looking up how often any particular medcode occurs in that set of patients. This can save time agonising over whether particular medcodes that may occur at very low frequencies (or not at all) should be included in the final code list. Conversely, codes that occur very frequently have high potential to introduce bias into analyses if they are erroneously included in code lists.

**11. Validating code lists**

This can be done by

1. Sensitivity analyses comparing the number of patients identified with broader and narrower code lists to the number of patients that would be expected based on other information, e.g. QOF data
2. Checking whether the observed associations between lifestyle variables, symptoms and diseases defined by the code lists are as expected (e.g. looking at the association between smoking and coronary heart disease, or between smoking and IMD, or age and chronic conditions).
3. Checking the concordance between conditions identified in the primary care data, and in the linked data (HES and cancer registry data).

Data on prevalence from the QOF is available here:

<https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/2021-22#resources>

Other sources of information on disease incidence/prevalence include the Health Survey for England, Cancer Research UK, British Heart Foundation etc.

[Health Survey for England - NHS Digital](https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england)

[Cancer Statistics for the UK (cancerresearchuk.org)](https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk)

[Heart statistics - Heart and Circulatory Diseases in the UK | BHF](https://www.bhf.org.uk/what-we-do/our-research/heart-statistics)

**References**

1. Williams R, Kontopantelis E, Buchan I, Peek N. Clinical code set engineering for reusing EHR data for research: A review. Journal of Biomedical Informatics 2017;70:1-13.

2. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, Reeves D. ClinicalCodes: An online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. PLoS One 2014;9(6):e99825. Doi:10.1371/journal.pone.0099825.

3. Nicholls SG, Quach P, von Elm E et al. The Reporting of Studies Conducted using Observational Routinely-Collected Health Data (RECORD) Statement: Methods for Arriving at Consensus and Developing Reporting Guidelines. PLoS One 2015:10:e0125620.

4. Watson J, Nicholson BD, Hamilton W, Price S. Identifying clinical features in primary care electronic health record studies: methods for codelist development. BMJ Open 2017;7:e019637 doi 10.36/bmjopen-2017-019637.

.

Appendix 1. Overview of medical coding systems

**ICD-10**

ICD-10 is the International Classification of Diseases 10th Revision coding system. The ICD is the international standard diagnostic classification system used in epidemiology. It was originally used to classify causes of mortality and later extended to cover diagnoses in morbidity.

Detailed information on the ICD-10 is available here.

[ICD10Volume2\_en\_2010.pdf (who.int)](https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2010.pdf)

The latest edition of ICD-10 includes 21 chapters.

|  |  |  |
| --- | --- | --- |
| Chapter | Code Range | Description |
| 1 | A00-B99 | Certain infectious and parasitic diseases |
| 2 | C00-D49 | Neoplasms |
| 3 | D50-D89 | Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism |
| 4 | E00-E89 | Endocrine, nutritional and metabolic disorders |
| 5 | F01-F99 | Mental, behavioural and neurodevelopmental disorders |
| 6 | G00-G99 | Diseases of the nervous system |
| 7 | H00-H59 | Diseases of the eye and adnexa |
| 8 | H60-H65 | Diseases of the ear and mastoid process |
| 9 | I00-I99 | Diseases of the circulatory system |
| 10 | J00-J99 | Diseases of the respiratory system |
| 11 | K00-K95 | Diseases of the digestive system |
| 12 | L00-L99 | Diseases of the skin and subcutaneous tissue |
| 13 | M00-M99 | Diseases of the musculoskeletal system and connective tissue |
| 14 | N00-N99 | Diseases of the genitourinary system |
| 15 | O00-O9A | Pregnancy, childbirth and the puerperium |
| 16 | P00-P96 | Certain conditions originating in the perinatal period |
| 17 | Q00-Q99 | Congenital malformations, deformations and chromosomal abnormalities |
| 18 | R00-R99 | Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified |
| 19 | S00-T88 | Injury, poisoning and certain other consequences of external causes |
| 20 | V00-Y99 | External causes of morbidity |
| 21 | Z00-Z99 | Factors influencing health status and contact with health services |

These are a hierarchical series of 3 or 4 character codes, the first character being a letter and the 2nd through 4th numbers. Where a 4th character is present there is a decimal point between the third and 4th character. The 4th character may identify different sites or varieties of a disease, or different individual diseases within a group of conditions.

Eg. C34 Malignant neoplasm of bronchus or lung

 C34.1 Malignant neoplasm of upper lobe, bronchus or lung

In some cases 5th or 6th characters may be available – this is the case for **Chapter 13**, where they indicate subdivisions by anatomic site, and **Chapter 19** to indicate open and closed fractures and injuries with and without open wound.

An ICD-10 browser is available here:

[ICD-10 Version:2010 (who.int)](https://icd.who.int/browse10/2010/en)

This can be searched either with search terms or by selecting down through the hierarchies.

**READ CODES**

The Read clinical coding system was widely used in the NHS from 1985 until 2018 when it was superseded in most contexts by SNOMED CT. Two versions (version 2 and version 3/CTV3) are in use. In addition to coding diagnoses and symptoms there are also codes for other phenomena, e.g. codes for patient lifestyle variables and social class and administrative codes. Both version 2 and version 3 are 5 byte alphanumeric codes.

A useful guide to the Read codes and abbreviations used within Read Terms (e.g. H/O, NOS etc.) is available here:

<https://www.scimp.scot.nhs.uk/better-information/clinical-coding/scimp-guide-to-read-codes>

The codes are organised into chapters – the different chapter headings are given in the table below.

|  |  |
| --- | --- |
| Chapter | Heading |
| 1 | History/Symptoms |
| 2 | Examinations/Signs |
| 3 | Diagnostic Procedures |
| 4 | Laboratory Procedures |
| 5 | Radiology/physics in Medicine |
| 6 | Preventive procedures |
| 7 | Operations, procedures, sites |
| 8 | Other therapeutic procedures |
| 9 | Administration |
| A | Infectious and parasitic diseases |
| B | Neoplasms |
| C | Endocrine, nutritional and metabolic disorders |
| D | Diseases of blood and blood-forming organs |
| E | Mental disorders |
| F | Nervous system and sense organ diseases |
| G | Circulatory system diseases |
| H | Respiratory system diseases |
| J | Digestive system diseases |
| K | Genitourinary system diseases |
| L | Complications of pregnancy, childbirth and the puerperium |
| M | Skin and subcutaneous tissue diseases |
| N | Musculoskeletal and connective tissue diseases |
| P | Congenital anomalies |
| Q | Perinatal conditions |
| R | [D] Symptoms, signs and ill-defined conditions |
| S | Injury and poisoning |
| T | Causes of injury and poisoning |
| U | [X] External causes of morbidity and mortality |
| Z | Unspecified conditions |

The codes are hierarchically organised becoming more specific as you move down the hierarchy.

Eg.



(Taken from Aickelin et al. SSRN Electronic Journal 2014, doi: 10.2139/ssrn.2828031

**SNOMED CT**

There are 19 hierarchies

* + Body structure
	+ Clinical findings
	+ Environmental or geographical location
	+ Event
	+ Observable entity
	+ Organism
	+ Pharmaceutical/biological product
	+ Physical force
	+ Physical object
	+ Procedure
	+ Qualifier value
	+ Record artefact
	+ Situation with explicit context
	+ SNOMED CT Model component (metadata)
	+ Social context
	+ Special concept
	+ Specimen
	+ Staging and scales
	+ Substance

Unlike Read V2 and the ICD coding systems, a SNOMED concept ID can be included in more than one hierarchy.

Further information on the SNOMED CT hierarchies is available here:

[SNOMED - 5-Step Briefing](https://www.snomed.org/snomed-ct/five-step-briefing)

[6. SNOMED CT Concept Model - SNOMED CT Starter Guide - SNOMED Confluence (ihtsdotools.org)](https://confluence.ihtsdotools.org/display/DOCSTART/6.%2BSNOMED%2BCT%2BConcept%2BModel)

[SNOMED CT Educational Resources - Delen: Home - NHS Digital (kahootz.com)](https://hscic.kahootz.com/t_c_home/view?objectID=37787312)

An NHS SNOMED CT browser is available here:

[NHSDigital SNOMED CT Browser (termbrowser.nhs.uk)](https://termbrowser.nhs.uk/)

The browser can be searched by entering a search term or clicking down through the hierarchies.

An example is given below of the information shown on the “Summary” tab for the SNOMED concept of varicella.



Parent and child concepts are shown along with the SNOMED CT ID (SCTID).

The blue box shows the preferred term (shown next to the ID) with synonyms shown underneath (eg. “Coughing up blood” for hemoptysis as shown below).



**Converting between ICD-10, Read and SNOMED CT**

There is some capacity to convert between Read Codes and SNOMED CT using the Code Browsers supplied by CPRD. In addition, the SNOMED CT UK Data Migration Workbench can be used to map between ICD-10, Read codes and SNOMED CT.

This is available from NHS TRUD <https://isd.digital.nhs.uk/trud/user/guest/group/0/home> (it is found under “Terminology Browsers and Tools”.

Appendix 2 – Example R script to search the dictionary for codes

rm(list=ls())

#set the working directory to where the dictionary file is saved and read it in

setwd("C:/Users/bsms9bbh/OneDrive - University of Brighton/CPRD work/Code lists")

dictionary <- read.delim("medical.txt", header=FALSE, na.strings="", stringsAsFactors = FALSE,

 skip=1)

names(dictionary) <- c("medcode", "readcode", "clinicalevents", "immunisationevents",

 "referralevents", "testevents", "readterm", "databasebuild")

#changes read term to all lower case as searches are case sensitive

dictionary$readterm <- tolower(dictionary$readterm)

#Example searches

#1 This searches for "diabetes" OR "cancer" within the read term (| is used for OR)

codelist1 <- dictionary[(grepl("diabetes|cancer", dictionary$readterm)),]

#2 This searches for "eating" AND "disorder" within the read term

codelist2 <- dictionary[(grepl("eating", dictionary$readterm) & grepl("disorder", dictionary$readterm)),]

#3 This searchers for "diabetes" or "cancer" AND "complications" within the read term

codelist3 <- dictionary[(grepl("diabetes|cancer", dictionary$readterm) & grepl("advice", dictionary$readterm)),]

#4 Can also define the terms separately which may be easier if using long list of terms

searchterm1 <- "cancer|neoplasm|tumour|tumor|carcinoma"

codelist4 <- dictionary[(grepl(searchterm1, dictionary$readterm)),]

#5 Can combine 2 search terms defined in this way using AND

searchterm2 <- "lung|resp|trachea|bronch"

codelist5 <- dictionary[(grepl(searchterm1, dictionary$readterm) & grepl(searchterm2, dictionary$readterm)),]

#6 Can use "!" to exclude rows containing particular search terms

#Eg search below find find lines that include "diabet" but NOT "pregnan|gestat"

codelist6 <- dictionary[(grepl("diabet", dictionary$readterm) & !grepl("pregnan|gestat", dictionary$readterm)),]

#7 Not really logical for these examples but this syntax will bind several

#code lists from different searches together and remove any duplicate rows.

allcodes <- rbind(codelist1, codelist2, codelist3) #bind together

allcodes\_unique <- unique(allcodes) #remove duplicate rows

#save the final list as csv file

write\_csv(allcodes\_unique, "mycodelist.csv")

Appendix 3

rm(list=ls())

library(readr)

#set the working directory to wherever your code lists are

setwd("C:/Users/bsms9bbh/OneDrive - University of Brighton/CPRD Work/Code lists")

#read in dictionary file from code browser directory

dictionary <- read.delim("C:/Users/bsms9bbh/OneDrive - University of Brighton/CRPD Gold code browser/medical.txt", header=FALSE,

 na.strings="", stringsAsFactors = FALSE, skip=1)

names(dictionary) <- c("medcode", "readcode", "clinicalevents", "immunisationevents","referralevents", "testevents",

 "readterm", "databasebuild")

#read in code lists for merging with medcodes from .csv files

#make sure the read code column in the codelist is labelled “readcode”!

#if there is a description column in the codelist can check this agrees with read term to make sure match has worked correctly

asthma <- read.csv("asthma.csv")

asthma\_medcodes <- merge(x=dictionary, y=asthma, by="readcode", all.y=TRUE)

write\_csv(asthma\_medcodes, "asthma\_plus\_medcodes.csv")

Appendix 4

rm(list=ls())

library(readxl)

library(writexl)

#set the working directory to where the dictionary file is saved and read it in

setwd("C:/Users/bsms9bbh/OneDrive - University of Brighton/CPRD\_CodeBrowser\_202210\_Aurum/CPRD\_CodeBrowser\_202210\_Aurum")

dictionary <- read.delim("CPRDAurumMedical.txt", header=FALSE, na.strings="", stringsAsFactors = FALSE, colClasses = c(rep("character", 8)),

 skip=1)

names(dictionary) <- c("medcodid", "observations", "origreadcode", "cleansedreadcode", "term", "SCTID", "DescriptionID", "Release", "EmisCodeCategoryId")

#c("medcodid", "observations", "origreadcode", "cleansedreadcode", "term", "SCTID", "DescriptionID", "Release", "EmisCodeCategoryId")

#read in code lists for merging with "medcodes

setwd("C:/Users/bsms9bbh/OneDrive - University of Brighton/CPRD work/Code lists")

#NB Need to make sure you have saved your xlsx file code list with all long numeric codes in text format

bmi <- read\_excel("bmi\_snomed.xlsx")

bmi\_medcodeids <- merge(x=dictionary, y=bmi, by="SCTID", all.y=TRUE)

write\_xlsx(bmi\_medcodeids, "bmi\_plus\_medcodeids.xlsx")