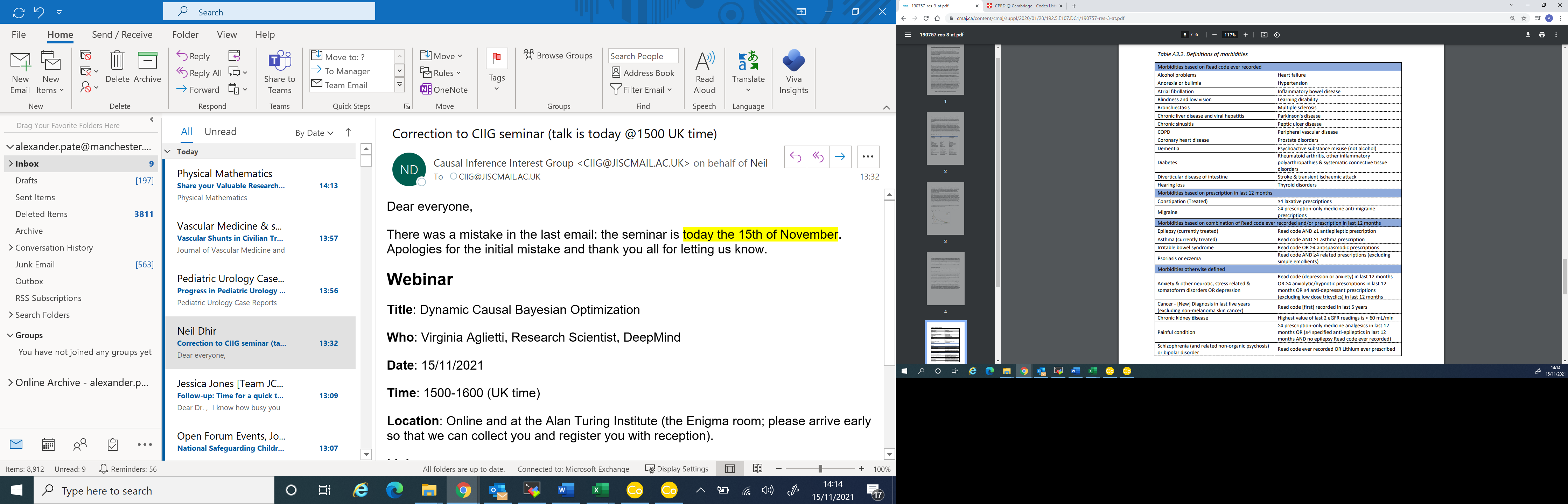
Data extraction documentation and data dictionary

Alexander Pate, Glen P Martin

# List of variables to extract

## All variables for consideration

We are planning to work with the 38 conditions from the Cambridge multimorbidity index. These are in Table 1. Table is taken from this reference1 (appendix 3: <https://www.cmaj.ca/content/cmaj/suppl/2020/01/28/192.5.E107.DC1/190757-res-3-at.pdf> ).



We will also derive age, gender, ethnicity, index of multiple deprivation (IMD) quintiles, body mass index (BMI), systolic blood pressure (SBP), cholesterol/HDL ratio and smoking status.

## Extraction process for outcomes and predictors

Potential outcomes in our model will be coronary heart disease (CHD), myocardial infarction (MI), Stroke, transient ischaemic attack (TIA), Atrial fibrillation (AF), heart failure, chronic kidney disease stage 3/4/5 (CKD), type 1 diabetes and type 2 diabetes. For outcomes we will want to identify occurrences through both primary care and secondary care (HES) data. This is because we don’t want to be modelling the time until first occurrence of an outcome, if they have already had a hospitalisation for that outcome. Another difference between outcomes and predictors is that for predictors we only want to know the value of the variable at baseline (either value of test data, or whether an individual has a record of a comorbidity prior to baseline). For outcomes conditions (all of which are chronic conditions) we need to know whether an individual had the condition prior to baseline, and for those that do have the condition, we want to extract the time since the condition was first developed. We also need to know the time until first occurrence of the outcome after the index date, or until censoring, and a censoring indicator. This means the extraction process will be significantly different for outcomes and predictors.

For predictors reliant on test data (BMI, SBP, cholesterol/HDL ratio, smoking status), we will look in the five years prior to the index date.

For predictors which are comorbidities, we will derive a variable which indicates whether an individual has a record of the comorbidity prior to their index date in their primary care record.

For outcomes, we will derive the time until first occurrence or censoring, and a censoring indicator, in both the primary care and secondary care datasets. We will also derive presence of the condition at baseline, and time since first occurrence of the condition (if present), in both the primary care and secondary care datasets. By only looking at the indicator for whether the condition is present at baseline in the primary care database, we can use an outcome variables as a predictor variables instead.

## Chronic Kidney Disease

CKD is a special case as we want to identify events from medical codes, but also from test data, namely eGFR scores and creatinie measurements. The process for doing so is below.

**Identifying CKD from eGFR/GFR, and estimating eGFR/GFR from creatinine**

KDIGO guidelines give definition of CKD: <https://kdigo.org/wp-content/uploads/2017/02/KDIGO_2012_CKD_GL.pdf>

This is from 2012, but I checked on KDIGO website and these are still the guidelines they recommend. Note how they say abnormalities of kidney function must be present for > 3 months.

This is CKD-EPI equation they recommend for converting creatinine mesaurements to GFR/eGFR scores: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2763564/> (also requires sex, ethnicity and age). Note that they state something along the lines of “this, or any equation shown to be better than this” can be used. Therefore there may be a more recent equation for this conversion, as the one they recommend was developed in 2009.

This comparison shows the CKD-EPI equation is better than the MDRD, which is another alternative: <https://pubmed.ncbi.nlm.nih.gov/22570462/>

The attached code looks for two entries below a certain value (60) that are more than 90 days (3 months) apart. Dataset should be in format of:

* 1 row per eGFR score
* Variables: person\_id(identifier for individuals), EntryDate (date of code), CodeValue (value of eGFR), num (increasing integer indicating the observation number for each individual)

There is then a bunch of code at the bottom which parallelises the function (to run separately on lots of subsets of the dataset), as otherwise it takes a very long time to go through the whole cohort. Note first half of this code was taken from somewhere and I’ve forgot to record the source.



**Overall strategy:**

1. Identify CKD stages 3/4/5 using medical codes
2. Extract eGFR/GFR + creatinine scores from test data
3. Convert creatinine measurements to eGFR scores
4. Use algorithm to identify if individuals meets criteria for CKD stages 3/4/5 using algorithm provided
5. CKD = 1 if either medical code, or CKD identified from test data.

# Code lists

## Code lists used for primary care extraction

For this study, we opt to focus only on medical codes, and ignore prescription data. Table 1 contains all the names of the code lists used for extraction of the primary care data and the source from which they were obtained. Code lists from source “AH” are available at the referenced Github page.2,3 Code lists from source “Cambridge mapped” were taken from the Cambridge primary care unit website,4 and then mapped by author TVS from CPRD GOLD to CPRD Aurum. Ethnicity code list was available at on the LSHTM Data Compass.5,6 The groupings for the medical codes were obtained from a Read code code list on another GitHub respository7, developed for another study.8 Some additional ethnicity codes were identified from this second code list, found in the CPRD Aurum code browser, and added to the code list. The smoking status code list was available on the LSHTM Data Compass.9,10 Self-generated code lists we developed by searching the CPRD Aurum code browser for \*height\*, \*weight\*, \*body mass\*, \*systolic\* and \*cholesterol\* respectively, and selecting the appropriate codes. Finally, we opted to use code lists for Psoriasis (as opposed to Psoriasis and eczema) and Rheumatoid arthritis (as opposed to rheumatoid arthritis and connective tissue disorders), due to the availability of code lists from source AH.

*Table 1: Code lists file names and source*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variable | Code list name | Source |
| **Medical history** | | | |
| 1 | Alcohol problems | Alcohol\_misuse | AH |
| 2 | Anorexia or bulimia | Eating\_disorders | AH |
| 3 | Asthma (currently treated) | Asthma | AH |
| 4 | Atrial fibrillation | Atrial\_fibrillation | AH |
| 5 | Anxiety or depression | Anxiety\_disorders | AH |
| 5 | Anxiety or depression | Depression | AH |
| 6 | Blindness and low vision | Visual\_impairment\_and\_blindness | AH |
| 7 | Bronchiectasis | Bronchiectasis | AH |
| 8 | Cancer Diagnosis in last five years | can\_\*\*\*\* | AH |
| 9 | Chronic kidney disease | Chronic\_kidney\_disease | AH |
| 10 | Chronic Liver +Viral Hepatitis | Hepatic\_failure | AH |
| 10 | Chronic Liver +Viral Hepatitis | Chronic\_viral\_hepatitis | AH |
| 11 | Chronic sinusitis | Chronic\_sinusitis | AH |
| 12 | COPD | COPD | AH |
| 13 | Coronary heart disease | Coronary\_heart\_disease | AH |
| 14 | Dementia | Dementia | AH |
| 15 | Diabetes | Diabetes\_mellitus\_other\_or\_not\_specified | AH |
| 15 | Diabetes | Diabetes\_mellitus\_type1 | AH |
| 15 | Diabetes | Diabetes\_mellitus\_type2 | AH |
| 16 | Diverticular disease of intestine | Diverticular\_disease | AH |
| 17 | Epilepsy (currently treated) | Epilepsy | AH |
| 18 | Hearing loss | Hearing\_loss | AH |
| 19 | Heart failure | Heart\_failure | AH |
| 20 | Hypertension | Hypertension | AH |
| 21 | Inflammatory bowel disease | IBD160\_mcid | TM |
| 22 | Irritable bowel syndrome | Irritable\_bowel\_syndrome | AH |
| 23 | Learning disability | Intellectual\_disability | AH |
| 24 | Multiple sclerosis | Multiple\_sclerosis | AH |
| 25 | Parkinson disease | Parkinsons\_disease | AH |
| 26 | Peptic Ulcer Disease | PEP135\_mcid | TM |
| 27 | Peripheral vascular disease | Peripheral\_vascular\_disease | AH |
| 28 | Prostate disorders | PRO170\_mcid | TM |
| 29 | Psoriasis or eczema | Psoriasis | AH (half) |
| 30 | Psychoactive substance misuse | Substance\_misuse | AH |
| 31 | RA connective tissue disorders | Rheumatoid\_arthritis | AH (half) |
| 32 | Schizophrenia bipolar | Schizophrenia | AH |
| 32 | Schizophrenia bipolar | Bipolar\_affective\_disorder\_and\_mania | AH |
| 33 | Stroke transient ischaemic attack | Stroke\_not\_otherwise\_specified | AH |
| 33 | Stroke transient ischaemic attack | Transient\_ischaemic\_attack | AH |
| 34 | Thyroid disorders | Tyroid\_disease | AH |
| 35 | Myocardial Infarction | Myocardial\_infarction | AH |
| **Demographic, lifestyle and test data** | | | |
| 36 | Ethnicity | Ethnicity\_aurum\_wgroups\_mcid | Internet search |
| 37 | Body mass index | height | Self-generated |
| 37 | Body mass index | weight | Self-generated |
| 37 | Body mass index | bmi | Self-generated |
| 38 | Systolic blood pressure | sbp | Self-generated |
| 39 | Smoking status | cr\_smokingcodes\_aurum | Internet search |
| 40 | Cholesterol | chol\_total | Self-generated |
| 40 | Cholesterol | chol\_hdl | Self-generated |
| 40 | Cholesterol | chol\_ratio | Self-generated |
| 41 | Creatinine | creatinine | Self-generated |
| 42 | eGFR | eGFR | Self-generated |
| 43 | GFR | GFR | Self-generated |

Note that some of the code lists were edited after being moved onto incline as some formatting changes had to be made. I am keeping a table of the pre-formatted variable names for personal reference: (Table 2).

*Table 2: Pre-formatted code list names and post-formatted code list names*

|  |  |  |
| --- | --- | --- |
| **Variable** | **Pre-format name** | **Post-format name** |
| Inflammatory bowel disease | IBD160\_mapped | IBD160\_mcid |
| Peptic Ulcer Disease | PEP135\_mapped | PEP135\_mcid |
| Prostate disorders | PRO170\_mapped | PRO170\_mcid |
| Ethnicity | Ethnicity\_aurum\_wgroups | Ethnicity\_aurum\_wgroups\_mcid |
| Body mass index | height | Height\_mcid |
| Body mass index | weight | Weight\_mcid |
| Body mass index | bmi | bmi\_mcid |
| Systolic blood pressure | sbp | sbp\_mcid |
| Smoking status | cr\_smokingcodes\_aurum | cr\_smokingcodes\_aurum\_mcid |
| Cholesterol | chol\_total | chol\_total\_mcid |
| Cholesterol | chol\_hdl | chol\_hdl\_mcid |
| Cholesterol | chol\_ratio | chol\_ratio\_mcid |

## Code lists used for HES/ONS extraction

Table 3 contains the variables and ICD 10 codes used to extract it. For chronic kidney disease, the full 5 digit ICD 10 code is required to separate CKD stages 1 and 2 from stages 3, 4 and 5. All others only require the initial 3 digits for extraction. Following the process in QRISK3, CHD and MI were grouped, and Stroke and TIA were grouped into one outcome. A massive sheet of oncology codes can be found here, which may be useful for future reference: <https://www.ambrygen.com/material/oncology/icd-10-code-reference-sheets/icd-10-codes-non-breast/631>

*Table 3: ICD 10 codes used for extraction*

|  |  |  |
| --- | --- | --- |
| Variable | ICD 10 codes | # characters |
| AF | 'I48' | 3 |
| CHD/MI | 'I20','I21','I22','I23','I24','I25' | 3 |
| Stroke/TIA | 'G45','I63','I64' | 3 |
| Heart failure | 'I50' | 3 |
| Type 1 diabetes | 'E10' | 3 |
| Type 2 diabetes | 'E11' | 3 |
| Chronic kidney disease stage 3/4/5 | 'N18.3','N18.30','N18.31','N18.32','N18.4','N18.5','N18.6','N18.9' | 5 |

## Code list repositories

Table 4 contains all the different repositories from which code lists can be obtained.

*Table 4: Table of codelists resources*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **Link** | **Codelist type** | **Disease area** | **Notes** |
| UoM Clinical codes repository | [Link](https://clinicalcodes.rss.mhs.man.ac.uk/) | Mixed | Mixed |  |
| OpenCodelists | [Link](https://www.opencodelists.org/) | Mixed | Mixed | Lots of the SNOMED code lists on here are automatically generated from read code lists and have not been verified |
| Cambridge Primary Care Unit | [Link](https://www.phpc.cam.ac.uk/pcu/research/research-groups/crmh/cprd_cam/codelists/v11/) | Read code | Multimorbidity |  |
| GitHub page: Multimorbidity research | [Link](https://github.com/annalhead/CPRD_multimorbidity_codelists) | SNOMED | Multimorbidity |  |
| GitHub page: Ethnicity research | [Link](https://github.com/opensafely/ethnicity-covid-research/tree/main/codelists) | Read code | Mixed | Authors say these are available on OpenCodelists, but I couldn’t find them |
| LSHTM Data Compass (search by keyword: Code list) | [Link](https://datacompass.lshtm.ac.uk/view/keywords/Code_list.html) | Mixed | Mixed |  |

# Data dictionary and list of all programs for extraction

## Computer programs used for extraction

Table 5 contains all the programs used for the extraction of the cohort from CPRD Aurum. Conversion of raw .txt files into sas datasets had already taken place. Restriction of cohort to those with > 1 day follow up aged > 65 had also already taken place.

*Table 5: Programs used for data extraction and what they do*

|  |  |
| --- | --- |
| **Program** | **What it does** |
| **Preliminaries** | |
| Libb\_Aurum\_65plus | Library statements to be loaded at the start of each program |
| Reformat\_codelists | Second code list formatting program |
| Code\_format1 | Format codelists |
|  |  |
|  |  |
| **Section 1: Create cohort** | |
| p1.1\_format\_extract.sas | Format extracted data |
| p1.2\_format\_extract.check.sas | Check the formatted files |
| p1.3\_cohort\_create\_548\_85.sas | Create cohort where 548 days (1 year and half) of valid follow up are required prior to index date, aged 65 - 85 |
| p1.3\_cohort\_create\_548\_100.sas | Create cohort where 548 days (1 year and half) of valid follow up are required prior to index date, aged 65 - 100 |
| p1.4\_choose\_final\_cohort | Pick either cohort 548\_85 or 548\_100, merge with linkage eligibility file. |
| **Section 2: Extract variables** | |
| **Cohort checks and creation (these need to run be after all the variables have been extracted, p2.5 – p2.7), so the numbering of the programs is stupid** | |
| p2.1\_cohort\_create.sas | Combine all the derived variables into a single dataset |
| p2.2\_cohort\_check\_var | Check distributions of each variable |
| p2.2\_cohort\_check\_missing | Check amount of missing data |
| p2.2\_cohort\_check\_hist.sas | Check histograms for each variable |
| P2.3\_export\_analysis\_datasets\_for\_R.sas | Export analysis datasets into a csv file |
| **Predictors** | |
| p2.5\_varAge.sas | Extracts age at index date for cohorts A and B |
| p2.5\_varBMI .sas | Extracts BMI |
| p2.5\_varSBP.sas | Extracts SBP |
| p2.5\_varBP.sas | Extracts BP |
| p2.5\_varCholesterol.sas | Extracts cholesterol |
| p2.5\_varEthnicity.sas | Extracts ethnicity |
| p2.5\_varSmoking.sas | Extracts smoking status |
| p2.5\_varIMD.sas | Extracts IMD score |
| p2.6\_varComor\_primhist.sas | Extracts history of comorbidities from primary care record. Looks to see whether there is a medical code prior to index date. Variables are:  \*INSERT LIST OF VARIABLES\* |
| **Outcome variables** | |
| P2.7\_varDeath.sas | Extract time until death and censoring indicator. Also extracts NelsonAalen estimator of the hazard of death at the time until death/censored. |
| p2.7\_varOutcomes\_heshist\_and\_hesevent.sas | (hist) Extracts history of outcome variables prior to index date in the secondary care record.  (event) Extract time until outcome variables post index date in the secondary care record.  Variables: AF, CHD/MI, Stroke/TIA, CKD, T2D |
| p2.7\_varOutcomes\_primhist.sas | Extracts history of outcome variables prior to index date in the primary care record.  AF, CHD, MI, Stroke, TIA.  Note that variables for CHD/MI are separate, and Stroke/TIA are separate, but we will be combining into a single variable for our outcome.  Note Diabetes and CKD are done separately. |
| p2.7\_varOutcomes\_primevent.sas | Extract time until outcome variables post index date in the primary care record.  Note that variables for CHD/MI are separate, and Stroke/TIA are separate, but we will be combining into a single variable for our outcome.  Note Diabetes and CKD are done separately. |
| p2.7\_varDiabetes\_primhist.sas | Extract history of diabetes prior to index date (type 1 and type 2 seperately) in the primary care record.  It has it’s own program as an algorithm was created: for people with vague diabetes codes, and no specific type 1 diabetes codes, we assume this is type 2 diabetes. |
| p2.7\_varDiabetes\_primevent.sas | Extract time until diabetes post index date (type 1 and type 2 seperately) in the primary care record.  It has it’s own program as an algorithm was created: for people with vague diabetes codes, and no specific type 1 diabetes codes, we assume this is type 2 diabetes. |
| p2.7\_varCKD\_primhist\_1\_derive\_gfr.sas | History of CKD prior to index date in primary care record: extract gfr values and export to .txt and .csv files |
| p2.7\_varCKD\_primhist\_2\_convert\_gfr.R | History of CKD prior to index date in primary care record: convert gfr values into whether it indicates CKD or not by applying algorithm |
| p2.7\_varCKD\_primhist\_3\_combine\_medtest.sas | History of CKD prior to index date in primary care record: combine CKD events identified in medical record and from test data |
| p2.7\_varCKD\_primevent\_1\_derive\_gfr.sas | Time until CKD post index date in primary care record: extract gfr values and export to .txt and .csv files |
| p2.7\_varCKD\_primevent\_2\_convert\_gfr.R | Time until CKD post index date in primary care record: convert gfr values into whether it indicates CKD or not by applying algorithm |
| p2.7\_varCKD\_primevent\_3\_combine\_medtest.sas | Time until CKD post index date in primary care record: combine CKD events identified in medical record and from test data |

## Data dictionary

Table 6 contains a data dictionary for the analysis datasets, including variable names, what the variable is, and what program it’s derived from.

*Table 6: Data dictionary for the analysis dataset*

|  |  |  |
| --- | --- | --- |
| **Variable name** | **What the variable is** | **Program it’s derived from** |
| **Demographic and test data** | | |
| Age | Age at index date | p2.5\_varAge.sas |
| BMI | BMI at index date (measurements within 5 years prior) | p2.5\_varBMI.sas |
| Cholhdl\_ratio | Cholesterol/HDL ratio at index date | p2.5\_varCholesterol.sas |
| Ethnicity6 | Ethnicity | p2.5\_varEthnicity.sas |
| Ethnicity16 | Ethnicity | p2.5\_varEthnicity.sas |
| SBP | Systolic blood pressure at index date (measurements within 5 years prior) | p2.5\_varSBP.sas |
| Smoking | Smoking status at index date (measurements within 5 years prior) | p2.5\_varSmoking.sas |
| Smoking\_anyhist | A binary indicator for whether an individual has any history of smoking (may be useful for imputation, to know someone definitely isn’t a never smoker). | p2.5\_varSmoking.sas |
| IMD | Index of multiple deprivation quintile | p2.5\_varIMD.sas |
| **‘History of’ comorbidities:** | | |
| Alcohol\_misuse\_primhist | Alcohol misuse | p2.6\_varComor\_primhist.sas |
| Eating\_disorders\_primhist | Anorexia or eating disorder | p2.6\_varComor\_primhist.sas |
| Asthma\_primhist | Asthma | p2.6\_varComor\_primhist.sas |
| Anxiety\_disorders\_primhist | Anxiety | p2.6\_varComor\_primhist.sas |
| Depression\_primhist | Depression | p2.6\_varComor\_primhist.sas |
| Visual\_impairment\_primhist | Blindness and visual impairment | p2.6\_varComor\_primhist.sas |
| Bronchiectasis\_primhist | Bronchiectasis | p2.6\_varComor\_primhist.sas |
| Hepatic\_failure\_primhist | Hepatic failure | p2.6\_varComor\_primhist.sas |
| Viral\_hepatitis\_primhist | Chronic viral hepatitis | p2.6\_varComor\_primhist.sas |
| Sinusitis\_primhist | Chronic sinusitis | p2.6\_varComor\_primhist.sas |
| COPD\_primhist | Chronic obstructive pulmonary disease | p2.6\_varComor\_primhist.sas |
| Dementia\_primhist | Dementia | p2.6\_varComor\_primhist.sas |
| Diverticular\_primhist | Diverticular disease | p2.6\_varComor\_primhist.sas |
| Epilepsy\_primhist | Epilepsy | p2.6\_varComor\_primhist.sas |
| Hearing\_loss\_primhist | Hearing loss | p2.6\_varComor\_primhist.sas |
| Hypertension\_primhist | Hypertension | p2.6\_varComor\_primhist.sas |
| IBS\_primhist | Irritable bowel syndrome | p2.6\_varComor\_primhist.sas |
| Intellectual\_dis\_primhist | Intellectual/learning disability | p2.6\_varComor\_primhist.sas |
| MS\_primhist | Multiple sclerosis | p2.6\_varComor\_primhist.sas |
| Parkinsons\_primhist | Parkinsons disease | p2.6\_varComor\_primhist.sas |
| Perip\_vascular\_primhist | Peripheral vascular diease | p2.6\_varComor\_primhist.sas |
| Psoriasis\_primhist | Psoriasis | p2.6\_varComor\_primhist.sas |
| Substance\_misuse\_primhist | Psychoactive substance misuse | p2.6\_varComor\_primhist.sas |
| RA\_primhist | Rheumatoid arthritis | p2.6\_varComor\_primhist.sas |
| Schizophrenia\_primhist | Schizophrenia | p2.6\_varComor\_primhist.sas |
| Bipolar\_primhist | Bipolar disorder | p2.6\_varComor\_primhist.sas |
| Thyroid\_primhist | Thyroid disease | p2.6\_varComor\_primhist.sas |
| Peptic\_ulcer\_primhist | Peptic ulcer disease | p2.6\_varComor\_primhist.sas |
| IBD\_primhist | Inflammatory bowel disease | p2.6\_varComor\_primhist.sas |
| Prostate\_primhist | Prostate disorders | p2.6\_varComor\_primhist.sas |
| **Outcome: Death** | | |
| Death\_t | Time until death or censored | p2.7\_vaDeath.sas |
| Death\_c | Censoring indicator | p2.7\_vaDeath.sas |
| Death\_NelsonAalen | Nelson Aalen estimate of cumulative hazard at Death\_t | p2.7\_vaDeath.sas |
| **Outcome: Atrial Fibrillation** | | |
| AF\_heshist | Binary indicator for presence of AF prior to index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| AF\_heshist\_t | Time since first occurrence of AF prior to index date in secondary care (HES) data.  Set = -1 for individuals where AF\_heshist = 0 | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| AF\_hesev\_c | Binary indicator for presence of AF post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| AF\_hesev\_t | Time until first occurrence of AF, or censoring, post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| AF\_primhist | Binary indicator for presence of AF prior to index date in primary care data | p2.7\_varOutcomes\_primhist.sas |
| AF\_primhist\_t | Time since first occurrence of AF prior to index date in primary care data.  Set = -1 for individuals where AF\_primhist = 0 | p2.7\_varOutcomes\_primhist.sas |
| AF\_primev\_c | Binary indicator for presence of AF post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| AF\_primev\_t | Time until first occurrence of AF, or censoring, post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| **Outcome: Heart failure** | | |
| HF\_heshist | Binary indicator for presence of HF prior to index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| HF\_heshist\_t | Time since first occurrence of HF prior to index date in secondary care (HES) data.  Set = -1 for individuals where HF\_heshist = 0 | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| HF\_hesev\_c | Binary indicator for presence of HF post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| HF\_hesev\_t | Time until first occurrence of HF, or censoring, post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| HF\_primhist | Binary indicator for presence of HF prior to index date in primary care data | p2.7\_varOutcomes\_primhist.sas |
| HF\_primhist\_t | Time since first occurrence of HF prior to index date in primary care data.  Set = -1 for individuals where HF\_primhist = 0 | p2.7\_varOutcomes\_primhist.sas |
| HF\_primev\_c | Binary indicator for presence of HF post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| HF\_primev\_t | Time until first occurrence of HF, or censoring, post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| **Outcome: Coronary heart disease/MI** | | |
| CHD\_MI\_heshist | Binary indicator for presence of CHD or MI prior to index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| CHD\_MI\_heshist\_t | Time since first occurrence of CHD or MI prior to index date in secondary care (HES) data.  Set = -1 for individuals where CHD\_heshist = 0 | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| CHD\_MI\_hesev\_c | Binary indicator for presence of CHD or MI post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| CHD\_MI\_hesev\_t | Time until first occurrence of CHD or MI, or censoring, post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| CHD\_primhist | Binary indicator for presence of CHD prior to index date in primary care data | p2.7\_varOutcomes\_primhist.sas |
| CHD\_primhist\_t | Time since first occurrence of CHD prior to index date in primary care data.  Set = -1 for individuals where CHD\_primhist = 0 | p2.7\_varOutcomes\_primhist.sas |
| CHD\_primev\_c | Binary indicator for presence of CHD post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| CHD\_primev\_t | Time until first occurrence of CHD, or censoring, post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| MI\_primhist | Binary indicator for presence of MI prior to index date in primary care data | p2.7\_varOutcomes\_primhist.sas |
| MI\_primhist\_t | Time since first occurrence of MI prior to index date in primary care data.  Set = -1 for individuals where MI\_primhist = 0 | p2.7\_varOutcomes\_primhist.sas |
| MI\_primev\_c | Binary indicator for presence of MI post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| MI\_primev\_t | Time until first occurrence of MI, or censoring, post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| **Outcome: Stroke/Transient Ischaemic Attack** | | |
| Stroke\_TIA\_heshist | Binary indicator for presence of Stroke or TIA prior to index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Stroke\_TIA\_heshist\_t | Time since first occurrence of Stroke or TIA prior to index date in secondary care (HES) data.  Set = -1 for individuals where Stroke\_heshist = 0 | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Stroke\_TIA\_hesev\_c | Binary indicator for presence of Stroke or TIA post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Stroke\_TIA\_hesev\_t | Time until first occurrence of Stroke or TIA, or censoring, post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Stroke\_primhist | Binary indicator for presence of Stroke prior to index date in primary care data | p2.7\_varOutcomes\_primhist.sas |
| Stroke\_primhist\_t | Time since first occurrence of Stroke prior to index date in primary care data.  Set = -1 for individuals where Stroke\_primhist = 0 | p2.7\_varOutcomes\_primhist.sas |
| Stroke\_primev\_c | Binary indicator for presence of Stroke post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| Stroke\_primev\_t | Time until first occurrence of Stroke, or censoring, post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| TIA\_primhist | Binary indicator for presence of TIA prior to index date in primary care data | p2.7\_varOutcomes\_primhist.sas |
| TIA\_primhist\_t | Time since first occurrence of TIA prior to index date in primary care data.  Set = -1 for individuals where TIA\_primhist = 0 | p2.7\_varOutcomes\_primhist.sas |
| TIA\_primev\_c | Binary indicator for presence of TIA post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| TIA\_primev\_t | Time until first occurrence of TIA, or censoring, post index date in primary care data | p2.7\_varOutcomes\_primevent.sas |
| **Outcome: Diabetes type 1** | | |
| Diab\_t1\_heshist | Binary indicator for presence of Diab\_t1 prior to index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Diab\_t1\_heshist\_t | Time since first occurrence of Diab\_t1 prior to index date in secondary care (HES) data.  Set = -1 for individuals where Diab\_t1\_heshist = 0 | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Diab\_t1\_hesev\_c | Binary indicator for presence of Diab\_t1 post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Diab\_t1\_hesev\_t | Time until first occurrence of Diab\_t1, or censoring, post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Diab\_t1\_primhist | Binary indicator for presence of Diab\_t1 prior to index date in primary care data | p2.7\_varDiabetes\_primhist.sas |
| Diab\_t1\_primhist\_t | Time since first occurrence of Diab\_t1 prior to index date in primary care data.  Set = -1 for individuals where Diab\_t1\_primhist = 0 | p2.7\_varDiabetes\_primhist.sas |
| Diab\_t1\_primev\_c | Binary indicator for presence of Diab\_t1 post index date in primary care data | p2.7\_varDiabetes\_primevent.sas |
| Diab\_t1\_primev\_t | Time until first occurrence of Diab\_t1, or censoring, post index date in primary care data | p2.7\_varDiabetes\_primevent.sas |
| **Outcome: Diabetes type 2** | | |
| Diab\_t2\_heshist | Binary indicator for presence of Diab\_t2 prior to index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Diab\_t2\_heshist\_t | Time since first occurrence of Diab\_t2 prior to index date in secondary care (HES) data.  Set = -1 for individuals where Diab\_t2\_heshist = 0 | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Diab\_t2\_hesev\_c | Binary indicator for presence of Diab\_t2 post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Diab\_t2\_hesev\_t | Time until first occurrence of Diab\_t2, or censoring, post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| Diab\_t2\_primhist | Binary indicator for presence of Diab\_t2 prior to index date in primary care data | p2.7\_varDiabetes\_primhist.sas |
| Diab\_t2\_primhist\_t | Time since first occurrence of Diab\_t2 prior to index date in primary care data.  Set = -1 for individuals where Diab\_t2\_primhist = 0 | p2.7\_varDiabetes\_primhist.sas |
| Diab\_t2\_primev\_c | Binary indicator for presence of Diab\_t2 post index date in primary care data | p2.7\_varDiabetes\_primevent.sas |
| Diab\_t2\_primev\_t | Time until first occurrence of Diab\_t2, or censoring, post index date in primary care data | p2.7\_varDiabetes\_primevent.sas |
| **Outcome: Chronic kidney disease stage 3/4/5** | | |
| CKD\_heshist | Binary indicator for presence of CKD prior to index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| CKD\_heshist\_t | Time since first occurrence of CKD prior to index date in secondary care (HES) data.  Set = -1 for individuals where CKD\_heshist = 0 | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| CKD\_hesev\_c | Binary indicator for presence of CKD post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| CKD\_hesev\_t | Time until first occurrence of CKD, or censoring, post index date in secondary care (HES) data | p2.7\_varOutcomes\_heshist\_and\_hesevent.sas |
| CKD\_primhist | Binary indicator for presence of CKD prior to index date in primary care data | p2.7\_varCKD\_primhist\_3\_combine\_medtest.sas |
| CKD\_primhist\_t | Time since first occurrence of CKD prior to index date in primary care data.  Set = -1 for individuals where CKD\_primhist = 0 | p2.7\_varCKD\_primhist\_3\_combine\_medtest.sas |
| CKD\_primev\_c | Binary indicator for presence of CKD post index date in primary care data | p2.7\_varCKD\_primhist\_1\_derive\_gfr.sas  p2.7\_varCKD\_primhist\_2\_convert\_gfr.R  p2.7\_varCKD\_primevent\_3\_combine\_medtest.sas |
| CKD\_primev\_t | Time until first occurrence of CKD, or censoring, post index date in primary care data | p2.7\_varCKD\_primevent\_1\_derive\_gfr.sas  p2.7\_varCKD\_primevent\_2\_convert\_gfr.R  p2.7\_varCKD\_primevent\_3\_combine\_medtest.sas |

## A few specific notes about the code for variable extraction and variable definitions, a list of all dataset names, code for how some variables are combined to create final dataset, and a final list of variable names

### Specific notes

* When extracting from HES, Stroke and TIA are one variable, but from primary care, they are separate. The primary care variables must be combined into one variable.
* When extracting from HES, CHD and MI are one variable, but from primary care, they are separate. The primary care variables must be combined into one variable.
* When extracting from HES, variables for diabetes type 1 and type 2 are stored in separate datasets, when extracting from primary care they are stored in the same dataset (so there are more \_hes datasets for diabetes, but same amount of variables for diabetes type1/type2.

### List of dataset names

%macro list\_datasets;

cohort\_in (in=ina)

/\* Predictors \*/

datint.varAge\_&cohort

datint.varEthnicity\_&cohort

datint.varBMI\_&cohort

datint.varCholhdl\_ratio\_&cohort

datint.varSBP\_&cohort

datint.varSmoking\_&cohort

datint.varIMD\_&cohort

datint.varAlcohol\_misuse\_primhist\_&cohort

datint.varEating\_disorders\_primhist\_&cohort

datint.varAsthma\_primhist\_&cohort

datint.varAnxiety\_disorders\_primhist\_&cohort

datint.varDepression\_primhist\_&cohort

datint.varBronchiectasis\_primhist\_&cohort

datint.varVisual\_impairment\_primhist\_&cohort

datint.varHepatic\_failure\_primhist\_&cohort

datint.varViral\_hepatitis\_primhist\_&cohort

datint.varSinusitis\_primhist\_&cohort

datint.varCOPD\_primhist\_&cohort

datint.varDementia\_primhist\_&cohort

datint.varDiverticular\_primhist\_&cohort

datint.varEpilepsy\_primhist\_&cohort

datint.varHearing\_loss\_primhist\_&cohort

datint.varHypertension\_primhist\_&cohort

datint.varIBS\_primhist\_&cohort

datint.varIntellectual\_dis\_primhist\_&cohort

datint.varMS\_primhist\_&cohort

datint.varParkinsons\_primhist\_&cohort

datint.varPerip\_vascular\_primhist\_&cohort

datint.varPsoriasis\_primhist\_&cohort

datint.varSubstance\_misuse\_primhist\_&cohort

datint.varRA\_primhist\_&cohort

datint.varSchizophrenia\_primhist\_&cohort

datint.varBipolar\_primhist\_&cohort

datint.varThyroid\_primhist\_&cohort

datint.varPeptic\_ulcer\_primhist\_&cohort

datint.varIBD\_primhist\_&cohort

datint.varProstate\_primhist\_&cohort

/\* Outcomes \*/

/\* Death \*/

datint.varDeath

/\* CKD \*/

datint.varCKD\_primev\_&cohort datint.varCKD\_primhist\_&cohort datint.varCKD\_hesev\_&cohort datint.varCKD\_heshist\_&cohort

/\* Diabetes \*/

datint.varDiab\_primev\_&cohort datint.varDiab\_primhist\_&cohort datint.varDiab\_t1\_hesev\_&cohort datint.varDiab\_t1\_heshist\_&cohort datint.varDiab\_t2\_hesev\_&cohort datint.varDiab\_t2\_heshist\_&cohort

/\* AF \*/

datint.varAF\_primev\_&cohort datint.varAF\_primhist\_&cohort datint.varAF\_hesev\_&cohort datint.varAF\_heshist\_&cohort

/\* HF \*/

datint.varHF\_primev\_&cohort datint.varHF\_primhist\_&cohort datint.varHF\_hesev\_&cohort datint.varHF\_heshist\_&cohort

/\* CHD/MI raw \*/

datint.varCHD\_primev\_&cohort datint.varCHD\_primhist\_&cohort datint.varMI\_primev\_&cohort datint.varMI\_primhist\_&cohort datint.varCHD\_MI\_hesev\_&cohort datint.varCHD\_MI\_heshist\_&cohort

/\* Stroke/TIA raw \*/

datint.varStroke\_primev\_&cohort datint.varStroke\_primhist\_&cohort datint.varTIA\_primev\_&cohort datint.varTIA\_primhist\_&cohort datint.varStroke\_TIA\_hesev\_&cohort datint.varStroke\_TIA\_heshist\_&cohort;

%mend;

### Code to combine Stroke and TIA, CHD and MI in primary care

We take the maximum time of the “time since first event”, but we take the minimum time of “time until first event”.

CHD\_MI\_primhist = max(CHD\_primhist, MI\_primhist)

CHD\_MI\_primhist\_t = max(CHD\_primhist\_t, MI\_primhist\_t)

CHD\_MI\_primev\_c = max(CHD\_primev\_c, MI\_primev\_c)

CHD\_MI\_primev\_t = min(CHD\_primev\_t, MI\_primev\_t)

Stroke\_TIA\_primhist = max(Stroke\_primhist, TIA\_primhist)

Stroke\_TIA\_primhist\_t = max(Stroke\_primhist\_t, TIA\_primhist\_t)

Stroke\_TIA\_primev\_c = max(Stroke\_primev\_c, TIA\_primev\_c)

Stroke\_TIA\_primev\_t = min(Stroke\_primev\_t, TIA\_primev\_t)

### List of variable names from initial extraction

This is a text list of all outcomes that can be copy and pasted into keep statements, etc, for personal use.

**/\* Predictors \*/**

Age

BMI

Cholhdl\_ratio

Ethnicity6

Ethnicity16

SBP

Smoking

Smoking\_anyhist

IMD

Alcohol\_misuse\_primhist

Eating\_disorders\_primhist

Asthma\_primhist

Anxiety\_disorders\_primhist

Depression\_primhist

Visual\_impairment\_primhist

Bronchiectasis\_primhist

Hepatic\_failure\_primhist

Viral\_hepatitis\_primhist

Sinusitis\_primhist

COPD\_primhist

Dementia\_primhist

Diverticular\_primhist

Epilepsy\_primhist

Hearing\_loss\_primhist

Hypertension\_primhist

IBS\_primhist

Intellectual\_dis\_primhist

MS\_primhist

Parkinsons\_primhist

Perip\_vascular\_primhist

Psoriasis\_primhist

Substance\_misuse\_primhist

RA\_primhist

Schizophrenia\_primhist

Bipolar\_primhist

Thyroid\_primhist

Peptic\_ulcer\_primhist

IBD\_primhist

Prostate\_primhist

**/\* Outcomes \*/**

/\* Death \*/

Death\_t Death\_c Death\_NelsonAalen Death\_NelsonAalen\_link

/\* CKD \*/

CKD\_primhist CKD\_primhist\_t CKD\_primev\_c CKD\_primev\_t

CKD\_heshist CKD\_heshist\_t CKD\_hesev\_c CKD\_hesev\_t

/\* Diabetes \*/

Diab\_t1\_primhist Diab\_t1\_primhist\_t Diab\_t1\_primev\_c Diab\_t1\_primev\_t

Diab\_t1\_heshist Diab\_t1\_heshist\_t Diab\_t1\_hesev\_c Diab\_t1\_hesev\_t

Diab\_t2\_primhist Diab\_t2\_primhist\_t Diab\_t2\_primev\_c Diab\_t2\_primev\_t

Diab\_t2\_heshist Diab\_t2\_heshist\_t Diab\_t2\_hesev\_c Diab\_t2\_hesev\_t

/\* AF \*/

AF\_primhist AF\_primhist\_t AF\_primev\_c AF\_primev\_t

AF\_heshist AF\_heshist\_t AF\_hesev\_c AF\_hesev\_t

/\* HF \*/

HF\_primhist HF\_primhist\_t HF\_primev\_c HF\_primev\_t

HF\_heshist HF\_heshist\_t HF\_hesev\_c HF\_hesev\_t

/\* CHD/MI \*/

CHD\_primhist CHD\_primhist\_t CHD\_primev\_c CHD\_primev\_t

MI\_primhist MI\_primhist\_t MI\_primev\_c MI\_primev\_t

CHD\_MI\_heshist CHD\_MI\_heshist\_t CHD\_MI\_hesev\_c CHD\_MI\_hesev\_t

/\* Stroke/TIA \*/

Stroke\_primhist Stroke\_primhist\_t Stroke\_primev\_c Stroke\_primev\_t

TIA\_primhist TIA\_primhist\_t TIA\_primev\_c TIA\_primev\_t

Stroke\_TIA\_heshist Stroke\_TIA\_heshist\_t Stroke\_TIA\_hesev\_c Stroke\_TIA\_hesev\_t

/\* CHD/MI and Stroke/TIA, after combining the separate primary variables for CHD/MI, and Stroke/TIA \*/

CHD\_MI\_primhist

CHD\_MI\_primhist\_t

CHD\_MI\_primev\_c

CHD\_MI\_primev\_t

Stroke\_TIA\_primhist

Stroke\_TIA\_primhist\_t

Stroke\_TIA\_primev\_c

Stroke\_TIA\_primev\_t;

## CPRD data doumentation

Links to data documentation for Aurum, HES and ONS data.

Data dictionary for HES: <https://www.cprd.com/sites/default/files/Data_Dictionary_HES_set21.pdf>

Data dictionary for Aurum: <https://www.cprd.com/sites/default/files/CPRD%20Aurum%20Data%20Specification%20v2.6.pdf>

ONS data documentation: <https://www.cprd.com/sites/default/files/Documentation_Death_set21_v2.5.pdf>

# References

1. Payne RA, Mendonca SC, Elliott MN, et al. Development and validation of the Cambridge Multimorbidity Score. *Cmaj* 2020; 192: E107–E114.

2. Head A. GitHub: Annalhead/CPRD\_multimorbidity\_codelists, https://github.com/annalhead/CPRD\_multimorbidity\_codelists (accessed 26 November 2021).

3. Head A, Fleming K, Kypridemos C, et al. Inequalities in incident and prevalent multimorbidity in England, 2004–19: a population-based, descriptive study. *Lancet Heal Longev* 2021; 2: e489–e497.

4. Primary Care Unit CU. CPRD @ Cambridge - Code Lists, http://www.phpc.cam.ac.uk/pcu/cprd\_cam/codelists/ (accessed 24 January 2018).

5. Mathur R. LSHTM Data Compass: Risk factor codelist - Ethnicity, https://datacompass.lshtm.ac.uk/id/eprint/2414/ (accessed 26 November 2021).

6. Strongman H, Carreira H, De Stavola BL, et al. Factors associated with excess all-cause mortality in the first wave of COVID-19 pandemic in the UK: a time-series analysis using the Clinical Practice Research Datalink. *medRxiv* 2021; 2021.06.04.21258344.

7. Mathur R, Morton C. GitHub: opensafely/ethnicity-covid-research, https://github.com/opensafely/ethnicity-covid-research/tree/main/codelists (accessed 26 November 2021).

8. Mathur R, Rentsch CT, Morton CE, et al. Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: an observational cohort study using the OpenSAFELY platform. *Lancet* 2021; 397: 1711–1724.

9. Forbes H. LSHTM Data Compass: Clinical Code list - Smoking SNOMED Codes, https://datacompass.lshtm.ac.uk/id/eprint/1603/ (accessed 26 November 2021).

10. Forbes HJ, Bhaskaran K, Grint D, et al. Incidence of acute complications of herpes zoster among immunocompetent adults in England: a matched cohort study using routine health data\*. *Br J Dermatol* 2021; 184: 1077–1084.