Setting KEEPER parameters

Anna Ostropolets

2025-09-11

Contents

1 Introduction

This vignette describes how one sets input parameters for Keeper to generate patient summaries.

2 Executing phenotype definition of interest

Keeper extracts and summarizes patient level data for patients in a given cohort to enable examination of patient summaries. Such examination can be used to interatively develop a phenotype (cohort definition) of a disease or (primary use case) determine if patients have the disease and subsequently calculate positive predictive value. The review should be done by a person familiar with the disease of interest. Alternatively, it can be done by LLM, which additionally enables calculation of sensitivity and specificity. Please refer to the Using Keeper with LLMs vignette for the latter use case.

First step is to create a phenotype of interest and execute it against database(s). Cohorts can be created using ATLAS, R or SQL. The cohort table should contain cohort_id (number of cohort), subject_id (patient identifier), cohort_start_date and cohort_end_date and will be subsequently fed into KEEPER input. More information on creating cohorts can be found here. One can then specify a sample size to randomly select a given number of patients from the cohort (parameter sample_size) or input a comma-separated vector of person_ids to select specific patients (parameter personIds). One can further de-identify patients by replacing OMOP personId with new random ids (parameter assignNewId = TRUE).

3 Creating concept code vectors

KEEPER extracts data based on user input. If a code is found in patient data, KEEPER will extract it along with the date relative to the index date (cohort_start_date). Therefore, code selection is very important.

We will use example of Type I Diabetes Mellitus (T1DM) to illustrate a strategy for code selection. The first step is to create a clinical definition. For this exercise, we will use a brief version of it. T1DM is an autoimmune condition characterized by decreased production of insulin by pancreas. Common onset is in childhood or adolescence but can occurr in adults. Symptoms include weight loss, polyuria and polydipsia, fatigue and others. Common differential diagnoses (those that need to be ruled out) include type II diabetes, pancreatic disorders such as cystic fibrosis, pancreanecrosis, steroid-induced diabetes, renal glucosuria and other conditions. Diagnostic procedures include glucose measurements, C-peptide, pancreatic and insulin antibodies as well as HbA1C. It is primarily treated with insulin. Complications include hypo and hyperglycemia, neuropathy, nephrophaty, cerebrovascular disease and peripheral arthery disease. We will use this definition to construct our inputs. The notion of differential diagnosis is important as for each input except for disease of interest we will consider T1DM and differential diagnoses to be able to see evidence for and rule out other diagnoses.

The full iput we will use as an example looks as follows:

3.1 DOI

Doi or disease of interest is the disease or state of interest itself. Here, we select two concepts with their descendants:

- 201254 Type 1 diabetes mellitus
- 435216 Disorder due to type 1 diabetes mellitus

First code is the code of T1DM itself and second code is the code denoting the diseases occurring due to T1DM which implies the patients have T1DM as well. Common strategy is to select codes used the index event criteria in the phenotype. If useAncestor is set to TRUE (default behaviour), KEEPER will use the hierarchy to pull in descendants of selected concepts.

DOI is looked up in CONDITION_OCCURRENCE table.

3.2 Symptoms

Here we input symptoms typically occurring in T1DM and differential diagnoses. These are signs and symptoms occuring in a short time window before the disease onset.

Based on our clinical definition, we selected the following codes:

- 79936 Polyuria
- 432454 Excessive thirst
- 254761 Cough (symptom of cystic fibrosis)
- 4229881 Weight loss
- 4232487 Weight gain

These codes are SNOMED codes that represent broad codes for the symptoms we are intrested in and source codes of the corresponding conditions map either to them directly or to their descendants. The decendants will be pulled in as we set useAncestor to TRUE. A good approach for selecting codes for this section as well as for the subsequent sections is to input your term in Atlas Search and click on the green shopping cart (Phoebe initial code selection) to get the starting point and use Phoebe (Recommend tab in Atlas Concept Set tab) to explore recommendations. Instructions on how to use Phoebe can be found here. Alternatively, you can explore your data to find appropriate SNOMED codes for symptoms using string search. It should be noted that with local data exploration you are more likely to miss the relevant codes.

Symptoms are looked up in OBSERVATION and CONDITION_OCCURRENCE tables within 30 days prior to the index date.

For this and subsequent categories we want to select the codes relevant to the doi as well as to other alternative (competing, differential) diagnoses. For example, cough is a symptom of cystic fibrosis. Similarly, hereon observing symptoms and other categories for T1DM will increase our confidence in the diagnosis while observing symptoms and other catogories for differential diagnoses will decrease our confidence in the diagnosis of T1DM.

3.3 Comorbidities

Comorbidities are conditions associated with the disease of interest or differential diagnoses. As opposed to symptoms, they can occurr within a longer time period before the disease onset.

For T1DM we selected the following comorbidities:

- 141253 Disorder of thyroid gland (comorbidity of T1DM)
- 432867 Hyperlipidemia (comorbidity of T2DM)
- 436670 Metabolic disease (comorbidity of T2DM)
- 433736 Obesity (comorbidity of T2DM)
- 255848 Pneumonia (comorbidity of cystic fibrosis)

Comorbidities are looked up in OBSERVATION and CONDITION_OCCURRENCE tables any time prior to the index date.

3.4 Drugs

We selected drugs (ancestor terms with descendants) that are used to treat T1DM as well as differential diagnoses:

- 21600712 DRUGS USED IN DIABETES (T1DM and T2DM)
- 21602728 Glucocorticoids (indicative of steroid-induced diabetes)
- 21603531 OTHER RESPIRATORY SYSTEM PRODUCTS (cystic fibrosis)

Drugs are looked up in DRUG_ERA table any time prior and any time after (displayed as two different columns).

3.5 Diagnostic procedures

Diagnostic procedures are procedure codes used for diagnosis of the disease of interest or alternative disease(s). For T1DM we did not select any procedures. One could think of relevant procedures such as ultrasound of pancreas for pancreanecrosis or CT of lungs for cyctic fibrosis.

Diagnostic procedures are looked up in PROCEDURE_OCCURRENCE table within 30 days prior and after the index date.

3.6 Measurements

Measurements are lab tests used to diagnose T1DM and diffential diagnoses:

- 4229110 Insulin C-peptide measurement
- 3005673 Hemoglobin A1c/Hemoglobin.total in Blood by HPLC

- 3004410 Hemoglobin A1c/Hemoglobin.total in Blood
- 3033819 Glucose-6-Phosphate dehydrogenase [Presence] in Serum

- 4149519 Glucose measurement Measurement
- 3005131 Glucose mean value [Mass/volume] in Blood Estimated from glycated hemoglobin

- 3010084 C peptide [Mass/volume] in Serum or Plasma
- 4020120 Trypsinogen measurement

3.7 Alternative diagnosis

Alternative diagnosis is where we put the diagnoses we rule out. As we already discused, differential for T1DM are the following conditions:

- 201826 Type 2 diabetes mellitus
- 40443308 Polycystic ovary syndrome
- 192963 Disorder of pancreas
- 441267 Cystic fibrosis

Alternative diagnosis codes are looked up in CONDITION_OCCURRENCE table within 90 days before and after the index date.

3.8 Treatment procedures

Treatment procedures are procedure codes corresponding to treatment of the disease of interest or alternative disease(s). We selected the following code denoting a broad group of procedures for pancrease partial or full removal for pancreonecrosis:

- 4242748 Incision of pancreas

We did not identify any tratement procedures for T1DM.

Treatment procedures are looked up in PROCEDURE_OCCURRENCE table any time after the index date.

3.9 Complications

Complications are other conditions occurring due to the disease. We selected the following codes with descendants:

- 4299544 Acanthosis nigricans (T1DM)
- 433968 Candidiasis (T2DM)
- 375545 Cataract (T2DM)
- 380834 Coma (T1DM and T2DM)
- 442793 Complication due to diabetes mellitus (T1DM and T2DM)
- 201820 Diabetes mellitus (descendants of for T1DM and T2DM)
- 4016045 Diabetic - good control (T1DM and T2DM)
- 4209145 Ketoacidosis (T1DM and T2DM)

Complications are looked up in CONDITION_OCCURRENCE table any time before or after the index date (displayed as two separate colums).

4 Data extraction and output

The output and it looks as follows:

- patient id;
- demographics (age, gender);
- visit_context: information about visits overlapping with the index date (day 0) formatted as the type of visit and its duration;
- observation_period: information about overlapping OBSERVATION_PERIOD formatted as days prior - days after the index date;
- presentation: all records in CONDITION_OCCURRENCE on day 0 with corresponding type and status;

- comorbidities: records in CONDITION_ERA and OBSERVATION that were selected as comorbidities and risk factors within all time prior excluding day 0. The list does not inlcude symptoms, disease of interest and complications;
- symptoms: records in CONDITION_ERA that were selected as symptoms 30 days prior excluding day 0. The list does not include disease of interest and complications. If you want to see symptoms outside of this window, please place them in complications;
- prior_disease: records in CONDITION_ERA that were selected as disease of interest or complications all time prior excluding day 0;
- prior_drugs: records in DRUG_ERA that were selected as drugs of interest all time prior excluding day 0 formatted as day of era start and length of drug era;
- prior_treatment_procedures: records in PROCEDURE_OCCURRENCE that were selected as treatments of interest within all time prior excluding day 0;
- diagnostic_procedures: records in PROCEDURE_OCCURRENCE that were selected as diagnostic procedures within all time prior excluding day 0;
- measurements: records in MEASUREMENT that were selected as measurements (lab tests) of interest within 30 days before and 30 days after day 0 formatted as value and unit (if exists) and assessment compared to the reference range provided in MEASUREMENT table (normal, abnormal high and abnormal low);
- alternative_diagnosis: records in CONDITION_ERA that were selected as alternative (competing) diagnosis within 90 days before and 90 days after day 0. The list does not include disease of interest;
- after_disease: same as prior_disease but after day 0;
- after_drugs: same as prior_drugs but after day 0;
- after_treatment_procedures: same as prior_treatment_procedures but after day 0;
- death: death record any time after day 0.