Macro documentation: identify_CCS_health_status

Date: Aug 26, 2024

- **Purpose**

To ascertain the Clinical Classification System (CCS) chronic disease category status given a set of diagnosis codes. There are 130 categories of chronic diseases, as well as 30 higher-level disease clusters that are mapped to ICDA-8, ICD-9-CM and ICD-10-CA.

- **Input files**

- Four provided files to input the crosswalks between ICD codes and CCS categories and clusters, CCS categories, and CCS clusters:

1. CCS_ diag_crosswalk.csv

Provides a crosswalk between three ICD versions (ICDA-8, ICD-9-CM, ICD-10-CA) and the CCS categories. It contains three variables:

- ✓ CCS_category_id (an ID assigned to each of the 130 CCS categories)
- ✓ diagnosis_coding_system (an indicator of the coding system: ICDA-8, ICD-9-CM or ICD-10-CA)
- ✓ diagnosis_code_match_pattern (lists the ICD codes that map to the CCS category, formatted into uppercase, without whitespace, and matches all precise codes under the specific parent code)

2. CCS_categories.csv

Lists the 130 CCS categories and contains two variables:

- ✓ CCS_category_id (an ID assigned to each of the 130 CCS categories)
- ✓ CCS_category_descr (provides a short description for each CCS category)

3. CCS_clusters.csv

Lists the 30 high-level CCS categories of chronic diseases and contains two variables:

- ✓ CCS_cluster_id (an ID assigned to 30 high-level clusters of CCS categories)
- ✓ CCS_cluster_descr (provides a short description for each high-level cluster)

4. CCS_cluster_mappings.csv:

Contains the mappings between CCS categories and high-level clusters and has two variables:

- ✓ CCS_category_id
- ✓ CCS_cluster_id

- Two files to input the study cohort:

1. Cohort_diagnoses:

The dataset includes IDs of the cohort members and all ICD codes identified in inpatient and/or outpatient records during a specific ascertainment window. It should include at least four variables:

- ✓ Individual ID (e.g., scrphin)
- ✓ encounter_date (date of the diagnosis code)
- ✓ diagnosis_code (the ICD code)
- ✓ coding_system (an indicator of the coding system: ICDA-8, ICD-9-CM or ICD-10-CA).

Notes:

- o Diagnosis codes must be all uppercase letters or numeric values, and whitespace must be trimmed.
- o One individual may have multiple records; however, the duplicate records should be removed by ID (e.g., scrphin), encounter_date, coding_system, diagnosis_code

2. Cohort_scrphins

A subset of the file "Cohort_diagnoses" that lists all individual IDs (one variable, e.g., scrphin)

Note: Input files were split in this format to improve computational efficiency

- One file to input ICD code labels.

This is an optional file. It could be supplied to produce labels of the ICD codes that were not mapped (i.e., skipped codes). It should have three variables: coding_system, diagnosis_code and diagnosis_descr (which provides a short description/label for the ICD code).

- **Parameters**

✓ participant_id_var_: name of the variable that uniquely identifies study participants. For example, scrphin

✓ participant_id_ds_: name of the dataset with IDs. For example, cohort_scrphins

✓ diagnosis_codes_ds_: name of dataset containing ICD diagnosis codes. For example, cohort_diagnoses

✓ concept_labels_ds_: name of dataset containing the CCS categories (CCS_categories)

✓ cluster_labels_ds_: name of dataset containing the high level CCS categories (CCS_clusters)

✓ diag_crosswalk_ds_: name of dataset containing the crosswalk between ICD codes and CCS categories (CCS_diag_crosswalk)

✓ cluster_mappings_ds_: name of dataset containing the crosswalk between CCS categories and clusters (CCS_cluster_mappings)

✓ out_wide_concept_binary_ds_name_: name of dataset where a summary will be saved of the status of each CCS condition in individuals

✓ out_wide_concept_freq_ds_name_: name of dataset where a summary will be saved of the number of CCS conditions identified per individual

✓ out_wide_concept_date_ds_name_: name of dataset where a summary will be saved of the earliest diagnosis date of each CCS condition in individuals. Wide format: each row refers to a study participant with many variables (one for each CCS condition)

✓ out_long_concept_ds_name_: name of dataset where a summary will be saved of the earliest diagnosis date of each CCS condition in individuals. Long format: each row refers to a single CCS condition (participant id, CCS_category_id, first_occurrence)

✓ out_wide_cluster_binary_ds_name_:name of dataset where a summary will be saved of the status of each high level CCS cluster in individuals

✓ out_skipped_codes_ds_name_: name of dataset where a summary of all the skipped ICD codes will be saved (i.e., the diagnosis codes that were not included in the provided ICD-to-CCS crosswalk file.

✓ diag_code_labels_ds_: name of dataset that has labels (short descriptions) of ICD codes, if available

✓ var_name_prefix_: [optional] characters that will be used to prefix the output variables. For example, if researchers would like all generated variables to begin with child_ or mom_ or dad_, this can be specified here. The prefix must be provided without quotes and is limited to a maximum of 8 characters.

✓ debug_: [default=0] a flag indicating whether intermediate datasets should be deleted (debug_=0) or retained (debug_=1)

Notes:

1. The ICDA-8 and ICD-9-CM codes in the crosswalk file are limited to 3-digits. The macro trims longer ICDA-8 and ICD-9-CM codes.
2. It's recommended to review the skipped diagnosis codes in the datasets "output_skipped_codes_ds_name_" to ensure relevant ICD codes were not excluded when ascertaining disease status.
3. Computation time varies depending on cohort size and number of diagnoses codes. In previous research with a cohort size of 125,000 individuals and about 14 million diagnosis codes over 5 decades, the computation time to run the macro was 30 minutes.