DALY CARE
NGC CLOUD

Naming Conventions

January 25, 2024

Purpose

The purpose of this document is to establish naming conventions for use when cleaning and storing data sets in the NGC cloud. Conventions are intended to provide unification across data sets and transparency for both regular users and collaborators.

Data sets

Naming scheme form:

- [origin]_[data name]_[converted/clean]_[date cutoff]

Example:

- RKKP_LYFO_CLEAN_20240101

Variables

IMPORT

For datasets in *IMPORT*, we always keep the original naming from the origin of the data. By keeping the original naming, documentation from the original sources (e.g. RKKP) can be used to understand what variables refer to and how they are encoded. For naming in CORE, see below.

General

- With two important exceptions, all names in lower case and in English.
    - Exception 1: commonly used abbreviations are capitalized e.g. “IPI_score_diagnosis”.
    - Exception 2: units are kept with their meaningful spelling, e.g. “IgA_uM_diagnosis”.
- If name includes spaces, spaces are replaced with “_”.
- Patient_id always named “patientid”
- Dates have prefix “date_”. Always “%Y-%m-%d”. E.g.
    - date_birth
    - date_diagnosis
    - date_infection
    - date_treatment_1st_line
    - date_progression
    - date_treatment_2nd_line
    - date_treatment_3rd_line
    - date_death
    - date_last_FU
    - date_last_FU_death (i.e. pmin(date_death, date_last_FU))
- Times have prefix “time_” and suffix as date-suffix. Unless otherwise specified, they are calculated from date_diagnosis. Always “hh-mm-ss".
- ICD10 codes have prefix “ICD10_”
- ATC codes have prefix “ATC_”
- SNOMED codes have prefix “SNOMED_”

- NPU codes have prefix “NPU_”
- SHAK codes have prefix “SHAK_”
    - Region (derived from SHAK) is called "region”
    - Hospital (Four first digits in SHAK) is called “hospital”
    - Department (derived from SHAK) is called “department”
- SKS codes from LPR have prefix “SKS_”

Treatment Lines, Relapse

As different data sources include variables pertaining to different lines of treatment (e.g. RKKP) along with data at diagnosis, it is necessary to distinguish clearly between these. Also, there may occur several relapses in the course of treatment, making a suffix such as “relapse” imprecise. To ensure accuracy the following suffixes should be implemented across data sets:

- “diagnosis”: data related to diagnostic investigations.
- “1st_line”: data related to 1st line treatment, even if there is only one line.
- “2nd_line”: data related to 2nd line treatment or first relapse. Because of this, a variable may be named “new_biopsy_performed_2nd_line” or “date_relapse_confirmed_2nd_line”, referring to biopsy or confirmation of first relapse.
- “3rd_line”: data related to 3rd line treatment or second relapse.
- And so on.

Though this is especially relevant for RKKP data sets, standardizing all data sets by this convention creates clarity and ease of use for all users while leaving no room for ambiguity.