MACCRs

This repository contains documentation for the set of Metadata Acquired from Clinical Case Reports (MACCRs) along with associated scripts for processing and verification of annotation metadata.

If you are accessing the data through *Figshare*, data files are available here.

If you are accessing the data through *Dryad*, some data files are available here. The remainder are provided as external links (see below).

If you are reading this file on Github, data files may be obtained from one of the above locations.

The data files are described in brief here.

Metadata Extraction Guide

A guide is provided in the ***Metadata_Extraction_Guide*** for the purposes of understanding creation of the MACCR set.

MACCR Data Files

Data Set

This file, ***MACCRs.tsv***, contains the full MACCR data set.

All annotation data are provided in this file. This is a tab-delimited file of 58 columns, one header row, and 3,100 rows of data, where each row provides metadata corresponding to a single clinical case report (CCR).

Please see the ***MACCR File Guide*** for details of all features provided in ***MACCRs.tsv***.

Citations

Citations for all clinical case reports used in the assembly of the MACCR set are available in the file ***MACCR_citations.bib***. This file is in BibTeX format. No abstract text is included.

Annotation Template

The annotation template is available at in ***TEMPLATE.xlsx***. This is an Excel format spreadsheet; we have found this format to be easiest for annotators to use.

Rare Mitochondrial Disease Subset

The Rare Mitochondrial Disease, or RMD, subset of the MACCRs includes 246 reports concerning a selection of six diseases with mitochondrial etiologies. This subset includes two tab-delimited files: ***MACCR_RMD_ICD10.tsv***, or the code file, and ***MACCR_RMD_ICD10_Categories.tsv***, the category file. Both files contain a header row, values corresponding to a sincle CCR per subsequent row, an identifying PubMed ID in the first column, and the name of a RMD in the second column. In the code

file, each subsequent column provides a binary value indicating whether the ICD-10-CM code in the header row is appropriate for the events described within the CCR. In the categories file, the header row contains names of 20 chapters of the ICD-10-CM codes, with values indicating whether at least one code in the corresponding code block has a value of '1' in the code file.

MeSH Term List

A list of all MeSH terms, with one unique term per list, is available in the tab-delimited file ***MACCR_mesh.tsv***. Each row includes a single MeSH term (as per the 2018 MeSH Terminology) in the first column and a single letter indicating the corresponding section of the MeSH Tree. Terms correspond to the 2018 version of MeSH. Please note that the U.S. National Library of Medicine is the creator, maintainer, and provider of MeSH. No proprietary rights to any MeSH content are claimed.

Named Entity List

Lists of named entities from a controlled vocabulary (specifically, combined MeSH heading names and SNOMED-CT terms) were prepared for a selection of the MACCR medical content categories - see ***MACCR_entities.tsv***. For each PMID, provided in the first column, each subsequent column provides no value if no named entities could be identified based on the text in the corresponding MACCR entry and column. Otherwise, a list of entity names is provided. Please note that these results are representative named entity recogition (NER) output only and do not take context into account. As with the MeSH Term List file, terms correspond to the 2018 version of MeSH, and please note that the U.S. National Library of Medicine is the creator, maintainer, and provider of MeSH. No proprietary rights to any MeSH content are claimed.

We thank Kushagra Rastogi for his assistance with NER.

Processing

See the project Github repository for processing scripts. Scripts for processing a set of annotation documents prepared using the annotation template are provided within the folder *Processing*. Both scripts, ***Extract.py*** and ***ExtractFunctions.py***, should be placed within and run from the same directory as a folder entitled "AnnotatedFiles" (alternatively, edit the inputDirectoryPath value of Extract.py to use a different folder name).

These scripts are compatible with Python 2 and 3. They require the following Python packages, installable through *pip*:

- **openpyxl**
- **word2number**
- **unicodecsv**
- **xlrd**
- **pandas**
- **numpy**

We thank Sarah Spendlove for developing these scripts.

Verification

See the project Github repository for verification scripts. Scripts for verifying a processed set of annotation documents are provided within the folder *Verification*. The R script ***QualityControl.R*** should be run within the same directory as the processed annotation set (i.e., MACCRs.tsv). The

functions within the script assume the existence of a comma-delimited annotation plan file containing, for each annotated record, the corresponding PMID and annotator. This allows for isolation of errors in the annotation process regarding target corpus coverage. Here, we have provided a blank annotation plan (***AnnotationPlan.csv***).

We thank Clement Feyt for developing these scripts.

Geolocation Analysis

See the Significant-Mapping Github repository. Scripts and an R Shiny app for visualizing the geographic distribution of a set of CCRs, based off locations indicated by their metadata and annotations, were developed by Kitu Komya, Clement Feyt, and Amanda Tsai.

Additional Resources

Please see also:

- heartCases, an automated system for retreiving and working with MEDLINE-indexed clinical case reports

Citation

Caufield, J.H. *et al.* A reference set of curated biomedical data and metadata from clinical case reports. *Sci. Data* 5, 180258 (2018). (Article in press.) DOI: 10.1038/sdata.2018.258