Metadata Extraction Guide for Metadata Acquired from Clinical Case Reports (MACCRs)

Introduction

This Guide contains a list of data types identified in the process of assembling the MACCR data set, along with guidelines for what types of text or numerical values were used in each type and details regarding data processing and cleaning procedures. Though this process may be adapted for future studies, this Guide specifically describes how we prepared the MACCR set. As our post-processing workflow enforces additional structure on some values (e.g., Demographics), the categories in the full MACCR file do not fully correspond to those in the annotation template. All columns in the MACCR file are described in the ***MACCR_File_Guide*** in this repository.

Manual Annotation with the Template

We have arranged the annotation types into four general categories: document and annotation indentification values, case report identification values (i.e., document-level properties), medical content concepts (for the most part, these are concept-level properties) and acknowledgements (i.e., text within each document linking it to other organizations and publications). A single set of values corresponds to those identified in a single document. The data types listed here are those we have determined to be most descriptive for clinical case reports and patient-focused medical documents in general. Our goal is also to follow FAIR principles in establishing structure on case report text and have identified the corresponding principles in two of the categories provided below.

Data types without corresponding values in a given document may be left blank or specified as "NA". For most fields, unless specified otherwise, separate values are separated by semicolons. This is compatible with semicolons contained within original text in that the punctuation already denotes a distinct concept or idea.

Scoring and Template Usage

The Template contains two columns (the first and second) for scores and two columns (third and fourth) for all text values. For each CCR, annotators provided text values from the full text document in the third column and values available in the PubMed citation in the fourth column (not including abstract text, but including MeSH terms corresponding to each category). The template then provides a score for both columns and each category, such that any value provided yields a score of 1. This score is intended to provide an approximation of which categories contain any relevant text value for a given document and do not provide any metric of the quality or usability of those values. For the Case Report Identification section, all values will be functionally identical between PubMed citation details and the full text.

Document and Annotation Identification

- **Internal ID**. All documents were assigned an identifier for this project, from CCR001 to CCR3000.

- **Access date** The date (day, month, and year) on which each document was read and annotated was recorded here in a variety of formats. Example: *6/28/2017*

Case Report Identification (Findable)

Values in this category provide document-level features.

For these values, annotators recorded:

- **Title**. The title of the document. Example: *Case report: a case of severe illness.*
- **Authors**. The authors of the document. Example: *Lastname FA; Secondlastname AB*
- **Year**. The year of publication of the document. Example: *1994*
- **Journal**. The full title of the journal in which the document was published. Example: *International journal of medicine*.
- **Institution**. The address of the home institution of the authors of the document, as specified in the document. This may include departments, geographic locations, and postal address details. If multiple locations were provided (e.g., if affiliations differ between authors), specify only details for the corresponding author were specified. If a corresponding author could not be identified, that of the first author was used if possible, or in lieu of that, not specified. If a corresponding author had multiple affiliations, both were specified and separated using semicolons. Examples: *Department of paediatrics, Division of paediatric cardiology, London Health Sciences Centre, London, Ontario, Canada*; *Department of Dermatology, University of California, 1701 Divisadero St, 3rd Floor, San Francisco, CA 94115, USA*
- **Corresponding Author**. A corresponding author for the document, as specified within the document heading. This author name has the same format as that used in the Authors data type. Example: *Lastname FA*
- **PMID**. The PubMed identifier for the document. Example: *29999555*
- **DOI**. A Digital Object Identifier, resolvable to the document (through https://www.doi.org/) and provided by the publisher. Example: *10.3928/00904481-20155555-03* (not a real DOI)
- **Link**. A stable URL to the full text of the document. This may be a doi.org URL - Example: https://doi.org/10.3928/00904481-20155555-03 (again, not a real DOI)
- **Language(s)**. The document's primary language. Example: *english*

Medical content (Accessable, Interoperable, Reusable)

Values in this category identify document-level, concept-level, and text-level features. These features provide ways to observe conceptual and semantic similarities between document contents, with a focus on medical topics and events. Most categories in this section were provided with values containing multiple text statements separated using semicolons. In general, annotators included sufficient detail to contextualize each statement but not extensive detail beyond each obervation, e.g. phrases such as "the patient presented with" were omitted unless these details were semantically crucial to the statement's meaning.

For these values, annotators recorded:

- **Key Words**. Specific terms identified within a document, usually in its header, as key terms. These were separated by semicolons. Some terms may be identical to MeSH terms provided through PubMed. Example: *barth syndrome; cardiomyopathy; 3-methylglutaconic acid*

- **Demography**. Any text statements describing the patient's background, including sex, age, ethnicity, or nationality. In practice, this nearly always includes age and sex, though not all documents provide these or further details. Example: *46-year-old; female*

- **Geographic Locations**. Any text terms or phrases denoting physical locations (i.e., not biological locations) other than those directly identifying the institution corresponding to the clinical presentation. This may include any geographic locale where the patient lives or has traveled to recently. Example: *born in Penteado, Alagoas, Brazil and residing in São Paulo*

- **Life Style**. Any text statements describing frequent patient activities or behaviors relevant to their general health. In practice, this frequently involves smoking or alcohol consumption habits, but may also include sun exposure, diet, or frequency of specific types of physical activity. Examples: *nonsmoker*, *drank alcohol in moderation*, *was physically active*

- **Family History**. Any text statements describing clinical observations of and events experienced by siblings, parents, and other family members. This includes genetic conditions and negative observations (i.e. *family history was negative for* a disease). Examples: *her mother had breast cancer at age 40*, *non-consanguineous parents with an uneventful perinatal history*, *she had family members with g6pd deficiency*.

- **Social History**. Any text statements describing patient background not covered in Demography or Life Style, though there may be overlaps in content between these categories. The statements may include occupational history and social habits. Examples: *college student*, *divorced with two children*, *he had been incarcerated for years*

- **Medical/Surgical History**. Any text statements describing any medical observations or events taking place prior to the beginning of the clinical presentation. This includes obstetric history and periods of good health. Also includes medical treatments. Examples:

  excessive fatigue; oppression over the chest; right-sided hemicolectomy; low-differentiated adenocarcinoma of the caecum; hypertension; capecitabine (five 500 mg oral tablets twice daily); chemotherapy

  with a 3-year history of chronic obstructive pulmonary disease; he had been diagnosed with stage I low-grade prostate cancer several years previously

- **Disease System**. Unlike the largely free-text values used elsewhere in the Medical Content, values for Disease System are at least one of 16 categories indicating disease type and organ system involvement, separated by semicolons. Categories are not comprehensive but indicate most systems impacted by the events described in the clinical presentation and diagnosed disease. The values were specified as follows, with additional details in parentheses:

  1. *cancer* (Any type of cancer or malignant neoplasm)
  2. *nervous* (Also referred to as Neuronal Disease or Nervous System Diseases. Includes any disease of the brain, spine, or nerves)
  3. *cardiovascular* (Also referred to as Cardiovascular Diseases)
  4. *musculoskeletal and rheumatic* (Also referred to as Musculoskeletal Diseases and Rheumatological Diseases)
  5. *digestive* (Also referred to as Digestive Diseases or Digestive System Diseases)
  6. *obstetrical and gynecological* (Also referred to as Obstetrical and Gynecological Diseases. Includes pregnancy and childbirth in addition to female reproductive organs and the breasts)
  7. *infectious* (Also referred to as Infectious Diseases)
  8. *respiratory* (Also referred to as Respiratory Diseases or Respiratory Tract Diseases)
  9. *hematologic* (Also referred to as Hematologic Diseases)
  10. *kidney and urologic* (Also referred to as Kidney Diseases and Urologic Diseases)
  11. *endocrine* (Also referred to as Endocrine Diseases)
  12. *oral and maxillofacial* (Also referred to as Oral and Maxillofacial Diseases. Includes all dental and craniofacial pathologies)

13. *eye* (Also referred to as Ophthalmological Diseases. Includes visual disturbances and blindness)
14. *otorhinolaryngologic* (Also referred to as Otorhinolaryngologic Diseases)
15. *skin* (Also referred to as Skin Diseases)
16. *rare* (A special category reserved for reports of rare diseases, defined as those impacting fewer than 200,000 individuals in the United States; see https://rarediseases.info.nih.gov/diseases)

Example: *cancer; skin; oral and maxillofacial* are used for a case entitled "Cutaneous horn: case report." (Akram et al. 2011, PMID 20226577).

- **Signs and Symptoms**. Any text statements describing any medical observations of signs or symptoms beginning at initial presentation but not including those in the outcome. May overlap with other types if symptoms continue from history to initial presentation. Examples:

  headaches; confusion; photophobia; neck stiffness; pyrexia (38.5 degrees C) and bilateral third nerve palsies

  a papulo-pustular rash and Raynaud's phenomenon; extreme fatigue, pale stools, dark urine and pruritus

- **Comorbidity**. Any terms or phrases describing distinct diseases present at the time of initial clinical presentation. There is often overlap between these values and those in clinical history, though Comorbidity should not include terms identical to those in the Diagnosis. Examples: *hypertension*, *becker muscular dystrophy*

- **Diagnostic Techniques and Procedures**. Any text statements describing medical procedures done for diagnostic purposes, including examinations, tests, and imaging. Quantitative results of lab tests are preferentially placed in Laboratory Values. Similarly, pathology results go in the Pathology section, but the procedures performed to obtain samples (e.g., biopsy) and those used in the course of analyses are included in Diagnostic Techniques and Procedures. Examples:

  xray; spinal MRI; abdominal ultrasound; iliac bone biopsy

  electrocardiogram; contrast enhanced computed tomography; coronary angiography; endomyocardial biopsy

- **Diagnosis**. Any text statements describing diagnoses of disease, even if the final diagnosis is ambiguous. Examples:

  aspergillosis; maxillary sinusitis

  diffuse right-sided facial soft tissue infection, mastoid effusion and temporal lobe cerebritis; septic lung metastases

- **Laboratory Values**. Names of diagnostic tests, their values, and conditions under which they were performed. This involves overlap with terms used in the Diagnostic Techniques and Procedures data type. Some overlap with Signs and Symptoms may be present, particularly in cases where names of diagnostic tests were not provided but terms describing the results (e.g., *leukopenia*) were present. Examples:

  white blood cell count, 6,200 /µl; hemoglobin level, 12.5 g/dl; mean corpuscular volume, 88.4; platelet count, 230,000 /µl; sodium level, 136 meq/l

laboratory tests (complete blood count, urine analysis, blood electrolytes, liver, and renal function tests) were normal

- **Pathology**. Any text statements describing results of pathology and histology studies, including gross pathology, immunology, and microscopy studies. Terms may overlap with those used in Diagnostic Techniques and Procedures. Examples:

  the left superior parathyroid was small, was partially removed, and was confirmed by biopsy to be normal parathyroid tissue

  uterine tumor (measuring 26×12×12 cm in size and weighing 1.8 kg) occupied a large part of the pelvic cavity; giant tumor showed necrotic changes, and had infiltrated the gall-bladder. there was neither the presence of a lymph node nor distant metastasis

- **Pharmacological Therapy**. Any text statements describing pharmaceutical therapies used in the course of treatment, including general terms such as *antibiotics* or specific drug terms. Values should also include descriptions of when and how drug therapies were stopped. Examples:

  methylprednisolone; prednisone

  therapy was initiated with meropenem hydrate (1.5 g/day), dopamine hydrochloride (3 μg/kg/min) and nafamostat mesylate (0.07 mg/kg/hr). meropenem hydrate was used until day 14, after which sulfamethoxazole (administered through the njt/gd) was employed. dopamine hydrochloride was discontinued from day 3 and nafamostat mesylate from day 5

- **Interventional Therapy**. Any text statements describing therapeutic procecures used in the course of treatment, including surgeries, implantation of medical devices, and procedures done to facilitate other therapies. Values should also include descriptions of when and how ongoing therapeutic procedures were stopped, if necessary. Examples:

  venesection; thrombectomy of the right coronary arteria; coronary stenting

  spinal tap; trephination was performed, with partial resection of the right frontal lesion; radiotherapy

- **Patient Outcome Assessment**. Any text statements describing health of the patient as of the end of the clinical presentation described in the report, including any follow-up tests. Examples:

  to date, 29 months post surgery, the patient is without clinical evidence of disease and is functionally performing well. She is followed as an outpatient every 6 months

  died 36 hours later due to progressive cardiac failure

- **Diagnostic Imaging/Figures/Videotape Recording/Tables**. All counts of visual media included in the report, in the following format:

  Count of images;Count of figures;Count of videos or animations;Count of tables

  We make the distinction between images and figures in this way: images include any products of clinical diagnostics, including photographs, micrographs, electrocardiogram rhythm images, and other products of diagnostic imaging, while figures are all other images, generally including those data plots and illustrations.

Example: *4;1;0;1*

- **Relationship to other Case Reports**. Identifiers of other reports in the data set cited by or referencing this report. Example: *27746431*

- **Relationship with Clinical Trial**. Identifiers of clinical trials referencing this report, specifically their ClinicalTrials.gov identifiers, preceded by NCT. Example: *NCT00000555*

- **Crosslink with Database**. Identifiers, preferably as database names and stable URLs, linking to this report. Example: *2 MedlinePlus Health Information:https://medlineplus.gov/lupus.html; 1 Genetic Alliance:http://www.diseaseinfosearch.org/result/4336*

Acknowledgements

Values provided in this category provide additional details about the document (i.e., metadata).

For these values, annotators recorded:

- **Funding Source**. Any text indicating government entities or organizations providing funding for the publication, or explicit statements that no outside funding was provided. Organization names are spelled out. Specific subgroups of large research organizations such as NIH are specified with a slash if they are noted within the source text. Multiple organizations are delimited with semicolons. Examples: *National Institutes of Health/National Center for Advancing Translational Sciences*, *American Cancer Society*, *The authors have no funding*

- **Award number**. Identifiers, if any, corresponding to specific financial support from the entities specified in Funding Source. Where appropriate and specified in the document, these values include initials of grant recipients in parentheses. Examples: *HD32062 (to AB)*, *Faculty research grant from the Fictional University College of Medicine (to ZP)*

- **Disclosures/Conflict of interest**. Any text descriptions or statements regarding financial disclosures and those of conflicts of interest, or explicit statements that authors have no conflicts of interest to declare. Example: *Dr. Lastname has received research support from MedicalCo and DrugCo*

- **References**. A numerical value providing a count of all references cited. Reference text is not included. Example: *12*.

Post-processing

Automated processing

The automated processing workflow performed the following functions:

- Aggregated all individual annotation files into a single, tab-delimited file.
- Converted all text values, with the exception of those in Title and Author, to all lowercase characters to improve word-level comparisons.
- Processed Demography values into Age and Gender using regular expressions. Processing followed the following rules:
    1. Ages are integers indicating number of years of age. Ages expressed as words (e.g. *twelve*) are converted to integers.
    2. Patients less than 1 year of age are assigned an age of 0.

3. If a decade category of age is provided, estimate the patient's age to be in the middle of the decade, e.g. a patient in his or her *50's* or *fifties* is estimated to be 55.
4. Where specified, sex should be annotated as *male* or *female*. This binary categorization may be expanded as needed.
5. Assigned disease category membership based on category annotation.
6. Separated media counts into four categories.

Manual cleaning

The following was done to ensure consistency within the dataset:

- CCR ID numbers (e.g., CCR123) were checked to ensure consistency.
- All dates were converted to a consistent format of YYYY-MM-DD.
- Titles were checked against PubMed citations to ensure they were identical.
- Author names were checked against PubMed citations, normalized for format (separated with semicolons, such that *Firstname A. Lastname, Authortwo B. Secondlastname* is rendered as *Lastname FA;Secondlastname AB*)
- Years were checked against the publication year specified in the PubMed citation.
- Journal titles were normalized to those specified by the NLM Catalog (https://www.ncbi.nlm.nih.gov/nlmcatalog). For example, *International journal of cardiology* was preferable to *Int J Cardiol* or *Int. J. Cardiol.* "The" was omitted from the prefix of journal names, such that *The Lancet* became *Lancet*.
- Institutions
- DOIs were obtained where possible and for documents without DOIs provided in the document or its PubMed citation. These DOIs correspond to the publisher's version of the full text of each CCR.
- Links were added where missing. In most cases, these are doi.org URLs, but in cases where DOIs are not available, direct links to articles on PubMed Central or publisher sites were used.
- Funding Source and Award Number values were normalized for format (in the first case, limited to organization names where possible, and in the second, limited to grant numbers/names and recipient initials).
- Disease category counts were checked to ensure they reflected the categories provided during annotation.
- Image and media counts were checked in instances where they appeared unexpectedly large.
- Demographic values were corrected where needed, to account for any remaining non-numeric age values and any non-binary gender values not normalized during automated processing.
- All fields were checked to ensure no material was in the wrong field (e.g., symptoms in the Funding Source category) and that semicolons were used as delimiters wherever possible.
- All text was confirmed to be in English.
- Common phrases were removed from the beginning and end of text segments in the Medical Content section where present, e.g. "she presented with a headache" became "headache" and "we performed a biopsy" became "biopsy".