

GUIDELINE FOR MEDICAL CONCEPT NORMALIZATION

CT-EBM-SP CORPUS

Vs 1

Leonardo Campillos Llanos
Adrián Capllonch Carrión
Ana Valverde Mateos

2021-23

Work funded by the CLARA-MeD project (MICIN, PID2020-116001RA-C33) developed at the CSIC; and by a JAE Intro scholarship (2021) funded by CSIC.

1. INTRODUCTION	3
UMLS METATHESAURUS BROWSER	4
2. NORMALIZATION CRITERIA.....	4
ENTITIES EXPRESSING THE TYPE OF STUDY OR TRIAL	7
DRUG REGIMEN.....	7
TEMPORAL ENTITIES	8
CRITERIA FOR AMBIGUOUS CONTEXTS	8
3. REFERENCES	9

1. INTRODUCTION

The normalization task is applied to 1200 texts about clinical trials: 500 journal abstracts and 700 clinical trial announcements published at the European Register of Clinical Trials (EudraCT, www.clinicaltrialsregister.eu). The normalization is applied to the semantic entities annotated, as shown in Figure 1 (sample of an annotated clinical trial announcement).

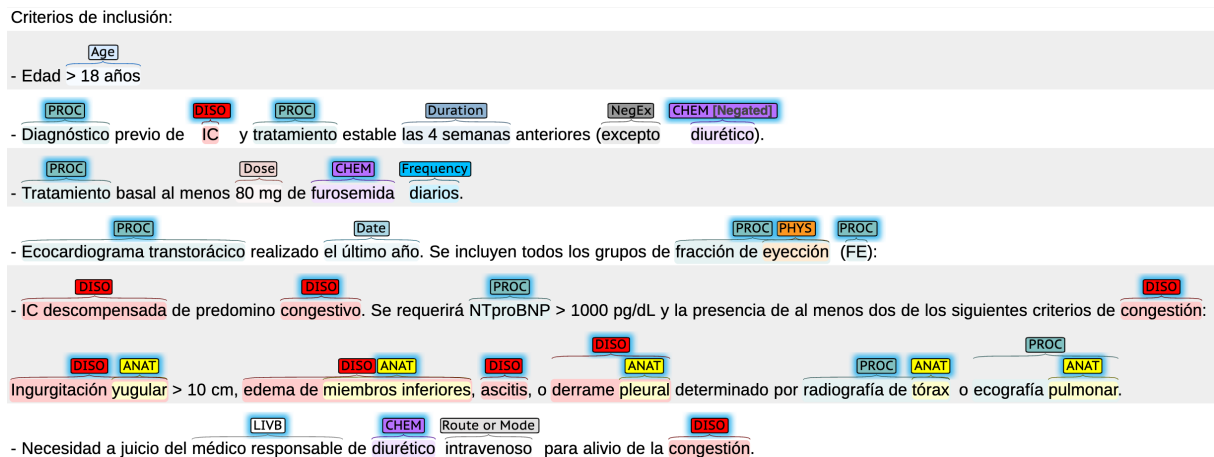


Figure 1. Annotation sample

The task consists in assigning **UMLS Concept Unique Identifiers (CUIs) codes to the annotated entities (normalisation)**. Normalisation facilitates access to medical information represented in variant forms expressing the same concept. For example, the UMLS code for *hipertensión arterial* (C0020538) is the same for variant forms such as *tensión arterial alta*, *tensión arterial elevada*, or the abbreviation *HTA*. We normalize the following entity types corresponding to semantic groups in the Unified Medical Language System (UMLS):

- Anatomic entities (ANAT): e.g. *brazo*
- Chemical or pharmacological substances (CHEM): e.g. *antibióticos*
- Medical devices (DEVI): e.g. *muletas*
- Pathological conditions, abnormalities and neoplasms (DISO): e.g. *diabetes*
- Genes and genetic material (GENE): e.g. *BRAF*
- Living beings and microorganisms (LIVB): e.g. *Staphylococcus aureus*
- Physiological processes (PHYS): e.g. *digestión*
- Therapeutic or diagnostic procedures and laboratory analyses (PROC): e.g. *radiografía*
- Entities expressing route or mode of drug administration (Route; e.g. *intravenoso*) or dosage form (Form; e.g. *píldora*)
- Entities expressing concepts (CONC): e.g. *esperanza de vida*
- Exceptionally, temporal entities with an UMLS CUI: e.g. *postprandial* (C0376674; Postprandial Period; Temporal Concept)

We do not normalize the following types of annotations:

- Temporal expressions (Timex) according to the TimeML standard: Duration, Frequency, or Time (excepting those cases where a CUI exists); Age expressions are not normalized either.
- Expressions of drug dose (Dose): e.g. *4 mg*
- Negation cues (e.g. *no*, *sin*) or speculation cues (e.g. *probablemente*)

The following sources are recommended to check the codes, in addition to UMLS:

- ATHENA: vocabulary browser of the Observational Medical Outcomes Partnership (OMOP): <https://athena.ohdsi.org/search-terms/start>
- CIMA (Centro de información online de medicamentos): <https://cima.aemps.es>
- International Classification of Diseases vs. 10, electronic edition (3rd ed, 2010), Ministerio de Sanidad, Consumo y Bienestar Social: https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_10_mc.html
- SEDOM Medical abbreviation dictionary: <http://www.sedom.es/diccionario/>
- DrugBank: <https://www.drugbank.ca>
- MedlinePlus: <https://medlineplus.gov/spanish/>
- Vademecum: <https://www.vademecum.es/>
- Wikipedia

UMLS METATHESAURUS BROWSER


To check the automatically pre-assigned codes, or to manually add missing codes, consult the UMLS Metathesaurus Browser (free register is required): <https://uts.nlm.nih.gov/metathesaurus.html>.

2. NORMALIZATION CRITERIA

The CUI code, the preferred term in the UMLS and the semantic type are indicated in a comment (using BRAT), separated by a semicolon (;), as follows:

CUI; ENTITY; SEMANTIC TYPE

For example, for the entity *fiebre*: C0015967; fever; Sign or Symptom.

 Any comment about the annotation that is not related to the normalization is written between square brackets ([]):

(1.) toman **1 c/día** (S1130-63432014000400005)
 Annotate *1 c* as Dose, with the comment [1 pill]

(2.) **BLcG** positivo (2016-004114-99)
 Annotate *BLcG* as PROC in this way:
 C0430103; Chorionic gonadotrophin stimulation test; Laboratory Procedure **[Mispelling]**

We use the most semantically precise concept code with respect to the annotated entity. If other codes exist, only the closest concepts are included, not the most general ones, nor lower subcategories:

- (3.) **diabetes mellitus** (2013-001229-15)
C0011849; Diabetes Mellitus; Disease or Syndrome
Do not normalize to the more general code (C0011847; Diabetes; Disease or Syndrome)
nor to more specific codes (C0011854; Diabetes Mellitus, Insulin-Dependent; Disease or Syndrome)

Annotated entities are normalized: **noun** or **adjective phrases**, and even **adverbs** or **verbs**, if it is possible to map them to the noun from which they are derived.

- (4.) **dificultad para respirar** (2019-002852-17)
dificultad para respirar: C0013404; Dyspnea; Sign or Symptom
respirar: C0035203; Respiration; Physiologic Function
- (5.) **asmáticos** (2011-005030-19)
C0004096; Asthma; Disease or Syndrome
- (6.) se les administró **aleatoriamente** (S1139-67092004000100009)
C0034656; Randomization; Research Activity

Entities in plural number may be normalised to concept codes corresponding to terms in the singular, or vice versa.

- (7.) **estudios de intervención** (2016-002077-35)
C3274035; Interventional Study; Research Activity
- (8.) **estudio prospectivo** (2014-001325-33)
C0033522; Prospective Studies; Research Activity

The annotated entities are also standardised to concept codes of **synonymous** terms or even **abbreviations** or **acronyms** documented in reliable texts.

- (9.) **patología ocular** (2014-005259-20)
C0015397; Disorder of eye; Disease or Syndrome
- (10.) Indicación científica: **Artritis Reumatoide (AR)** (2011-005030-19)
Artritis reumatoide: C0003873; Rheumatoid Arthritis; Disease or Syndrome
AR: C0003873; Rheumatoid Arthritis; Disease or Syndrome

Entities with spelling errors are also normalized to CUIs:

- (11.) **quimioradioterapia** (2012-005624-15)
C0436307; Chemoradiotherapy; Therapeutic or Preventive Procedure

Some entities may have several semantic types (STY); in these cases, each STY is separated by · and white spaces:

- (12.) **mitomicina C**
C0002475; mitomycin; Antibiotic · Organic Chemical

Some terms may be normalised to several CUIs because the UMLS collects several codes for the same entity (e.g. from different sources). In this case, **as long as no concept is a superclass of the other annotated concept**, all codes are indicated **separated by a vertical bar (|)**:

(13.) **ojo**

C0015392; Eye; Body Part, Organ, or Organ Component | C1280202; Entire eye; Body Part, Organ, or Organ Component

In other cases, the annotated entity corresponds to a coordinated expression of two or three concepts; in this case, the UMLS code is added for each concept, separated by a + sign:

(14.) **Función renal y hepática**

(2013-004008-20)

C0232804; Renal function; Organ or Tissue Function + C0232741; Liver function; Organ or Tissue Function



Entities in coordination contexts where only one entity can be normalized were annotated separately (discontinuous entities):

(15.) **Función adecuada de órgano y médula**

Annotate separately *función* (..) *de órgano* (C1254358; Organ or Tissue Function; Organ or Tissue Function) and *función* (..) *de médula* (not normalizable).

Do not annotate anything if there is not an exact code (do not use more general or very specific ones). This may occur in medical drugs that are not registered in the UMLS (they might appear in PubChem):

(16.) **fórmulas isoglucídicas** con distintos **edulcorantes** y **fibra**

fórmulas isoglucídicas: no CUI exists, cannot be normalized

edulcorantes: C0038998; Sweetening Agents; Chemical Viewed Functionally

fibra: C0225326 fiber; Organic Chemical; Pharmacologic Substance

(17.) **tratamiento aleatorizado**

(2013-003778-29)

Do not normalize to any code; the CUI of *tratamiento* (C0087111; Therapeutic procedure; Therapeutic or Preventive Procedure) is a superclass of this entity

(18.) **Prontogest**

(2015-000290-12)

Do not add any code, no CUI exists

If you want to include the normalisation to a doubtful code, add an interrogation mark (?) at the end of the comment:

(19.) **Capacidad vesical**

C0232840; Bladder function; Organ or Tissue Function (?)

Medical abbreviations and acronyms are often ambiguous: e.g., CV can stand for *carga viral* (C1261478; Viral Load Measurement; Laboratory Procedure), *test de capacidad* (C0430511; Vital capacity test; Diagnostic Procedure), or *calidad de vida* (C0281588; Assessment of quality of life; Diagnostic Procedure):

(20.) **CV** < 50 copias/mL

CV (PROC): C1261478; Viral Load Measurement; Laboratory Procedure

ENTITIES EXPRESSING THE TYPE OF STUDY OR TRIAL

Normalize those entities with an UMLS CUI:

(21.) Ensayo Clínico Aleatorio

C0206034; Clinical Trials, Randomized; Research Activity

(22.) Estudio de observación

C1518527; Observational Study; Research Activity

In longer entities, some CUIs may be concatenated:

(23.) Estudio prospectivo aleatorizado

C0033522; Prospective Studies; Research Activity + C0206034; Clinical Trials, Randomized; Research Activity

However, do not normalize too complex entities expressing too much information or made up of aspects without a code in the UMLS:

(24.) ensayo clínico aleatorizado y controlado (Fase-IIb) a doble ciego

Do not normalize the entity.

DRUG REGIMEN

Entities expressing mode or route of administration of drugs, substances or food (Route) are normalised to a UMLS CUI if a concept exists (usually Functional Concept, from the CONC semantic group):

(25.) recibieron 5 mg orales de ácido fólico (0211-699501013259)

C1527415; Oral Route of Drug administration; Functional Concept

(26.) infusión de dexmedetomidina intravenosa (S1134-80462015000100002)

C1522726; Intravenous Route of Drug Administration; Functional Concept

(27.) nutrición parenteral (S0370-41062014000300006)

C1518896; Parenteral Route of Drug Administration; Functional Concept

Likewise, entities expressing the dosage form (Form) are also normalized: *aerosol*, *cápsulas*, *comprimidos*, *crema*, *formulación*, *gel*, *gotas*, *tabletas*... They are generally UMLS entities of STYs Quantitative Concept (unit of presentation) or Biomedical or Dental Material.

(28.) comparar la efectividad de 800 UI/día de D2 (gotas) y D3 (comprimidos)

(S0025-76802012000300002)

gotas: C4318619; Drop (unit of presentation); Quantitative Concept

comprimidos: C0039225; Tablet Dosage Form; Biomedical or Dental Material

C4319774; Tablet (unit of presentation); Quantitative Concept

- (29.) **gel** con 250 mgr de hidroclicuro de lidocaína intra-rectal (S0004-06142009000500003)
Normalizate *gel* (annotated as Form): C0017243; gel; Biomedical or Dental Material
- (30.) **montelukast** en **tabletas masticables** de 5 mg (S0120-41572012000300010)
tabletas masticables: C0304290; Chewable Tablet; Biomedical or Dental Material

TEMPORAL ENTITIES

Temporal expressions (TIMEX) are not normalized to CUIs, except in cases where an UMLS CUI exists (Temporal Concept):

- (31.) un **seguimiento** de **dos años** (S1698-69462006000200015)
Do not normalize *dos años* (Duration) because no CUI exists.
Normalize *seguimiento* (PROC) to: C1522577; Follow-up; Health Care Activity
- (32.) pico de glicemia **postprandial**
C0376674; Postprandial Period; Temporal Concept
- (33.) náuseas y vómito en el **período postoperatorio**
C0032790; Postoperative Period; Temporal Concept

CRITERIA FOR AMBIGUOUS CONTEXTS

In some contexts, the annotated entity is a general and unspecific reference, but the context allows interpreting that it refers to a specific entity. In these cases, the more specific code may be added instead of the general one.

- (34.) infección por **coronavirus** (covid-19) (2020-001162-12)
C5203676; SARS-CoV-2; Virus
Here we use the more specific code for SARS-CoV-2 coronavirus, because it refers to the virus of COVID-19, rather than the code for the general class of coronavirus (C0206419; Genus: Coronavirus; Virus)

When in doubt, use the code that can be strictly interpreted by the context, without over-interpreting the information. But if the context allows you to infer a more specific concept, the more specific code will be used.

Some entities (mostly of the LIVB type) were annotated to record as much information as possible. However, no exact CUI for such a mention may exist, but several codes. In these cases, normalize using all concept unique identifiers available, separated by plus sign (+).

- (35.) **mujeres voluntarias**
C0043210; Woman; Population Group + C1520061; Volunteer Group; Group

3. REFERENCES

Bada, M., & M. Eckert. (2010) CRAFT Concept Annotation Guidelines. Vs 2010/03/14. https://bionlp-corpora.sourceforge.net/CRAFT/guidelines/CRAFT_concept_annotation_guidelines.pdf (Fecha de consulta: 4/7/2021)

Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004 Jan 1; 32(Database issue): D267–D270. doi: 10.1093/nar/gkh061

Mas, T., A. Villarrubia, N. Vittini (2020) *Guía de anotación de Codificación e Indexado en CIE-10 [CODIESP Challenge Annotation Guidelines]*. <https://temu.bsc.es/codiesp/index.php/2019/09/19/annotation-guidelines/> (Fecha de consulta: 31/05/2020)

Rabal, O, A. Intxaurreondo & M. Krallinger (2018) *Guía de anotación y normalización de compuestos químicos. [PharmaCoNER Challenge Annotation Guidelines]*. Disponible en: <https://bit.ly/395PLTE> Ministerio de Economía y Empresa, Plan de Impulso de Tecnologías del Lenguaje, área Sanidad.

Real Academia Nacional de Medicina de España – RANME (2011) *Diccionario de términos médicos (DTM)*. Madrid: Editorial Panamericana.