EUROPEAN
LANGUAGE
GRID

Annotation Guidelines for Clinical Entities in E3C

Manuela Speranza, Bernardo Magnini and Begoña Altuna

version 1.1, July 2021

European
Clinical
Case
Corpus

FONDAZIONE
BRUNO KESSLER

Annotation of clinical entities for the E3C Project

The goal of this document is to report the methodology adopted in the E3C project for the annotation of clinical entities in a clinical case. E3C focuses on a subgroup of clinical entities, i.e. *disorders*.
A disorder is defined in UMLS (https://uts.nlm.nih.gov/uts/umls/home) as:
Disease (C0012634) - A definite pathologic process with a characteristic set of signs and symptoms.

We make use of UMLS semantic types to identify UMLS concepts representing disorders, and restrict to the following twelve semantic types (in alphabetical order):

1. Acquired Abnormality
2. Anatomical Abnormality
3. Cell or Molecular Dysfunction
4. Congenital Abnormality
5. Disease or Syndrome
6. Experimental Model of Disease
7. Finding
8. Injury or Poisoning
9. Mental or Behavioral Dysfunction
10. Neoplastic Process
11. Pathologic Function
12. Signs or Symptoms

This is the same set of semantic types used by SHARE (Elhadad et al., 2012), except for the fact that we add the *Finding* semantic-type, as this is included by UMLS in the definition of disorder. Yet, it should be noted that not all findings are pathological, so, if they do not refer to a disorder, they should not be annotated as disorders.

The annotation task consists of individuating expressions in the clinical case that denote a disorder, and assigning them a corresponding concept in the UMLS metathesaurus. Our guidelines are adapted from the SHARE guidelines.

The SHARE guidelines include two annotation phases: (i) identifying a span of text in the clinical case that corresponds to a disorder and mapping it to a CUI (Concept Unique Identifier); and (ii) within the sentence where the disorder is mentioned, identifying additional information related to the disorder, referred to as modifiers.
In the E3C project we focus on the first phase, and leave the second phase to a future extension of our work.

1. Annotation steps

Annotators should follow these three main steps: (i) identification of a candidate disorder, (ii) search UMLS for concepts that match the candidate disorder; (iii) select a UMLS concept that can be mapped to for the candidate disorder and, if needed, adapt its textual extent[1].

1. Step 1: Candidate disorder identification

In this step the goal is to identify all textual expressions in a clinical case that are candidates to be mapped to disorders that are coded in UMLS. In order to identify candidate disorders, annotators can use both domain knowledge and information coming from the text (i.e. the whole clinical case). In this step, annotators should mainly focus on not excluding any potential disorders.

1.1 Textual extent of candidate disorders

Annotators should mark the "maximum extent" of each candidate disorder. This means that all modifiers of a noun phrase should be included, so that the candidate disorder is as specific as possible (see Ex. 1-2).

It is worth specifying that, in the following phases, annotators will check whether the candidate disorder (or a subpart of it) can actually be linked to a UMLS concept. Therefore individuating candidates that will not end up being marked (or mentions that are too specific) is not an issue.

> Example 1
> *The overall features of the [tumor] were most consistent with [metastatic melanoma] occurring in a background of [Barretts esophagus with high-grade dysplasia].*
>
> Example 2
> *Here we report a case of [aggressive metastatic cholangiocarcinoma] ([MCC]) in a 72-year-old man, sequentially treated with two targeted chemotherapies.*

1.2 Rules for candidate disorder identification

For the identification of candidate disorders, we mainly follow the SHARE guidelines as follows:

1.2.1 Each occurrence of a disorder in the text should be considered independently.

In example 2, for instance, we have two distinct candidate disorders, "aggressive metastatic cholangiocarcinoma" and the acronym "MCC".

[1] Note that, during the annotation process, annotators will likely switch back and forth between step 2 and 3.

1.2.2 A candidate disorder should be explicit[2], i.e. specifically mentioned.

In examples 1-2, we have marked between brackets all the explicit mentions of disorders. As one can see, even the acronym MCC in example 2 is marked as a disorder.

Descriptions of disorders and anaphoric references to disorders, instead, should not be marked as candidate disorders. Normally a disorder is described and not mentioned when it is presented as a full sentence or verbal phrase. In example 3, the disorder "swollen prostate" is not mentioned while a description of the status prostate is offered.

Example 3

His prostate appeared swollen in the image tests.

In example 4, the first sentence contains two candidate disorders ("Crohn flare" and "bowel obstruction") but note that "this" in the second sentence, while anaphorically referring to a disorder, is not marked as a candidate disorder.

Example 4
"She had a Crohn flare with symptoms of bowel obstruction that typically resolves with rehydration. Her current symptoms are reminiscent of this."

Example 5
Her menstrual cycle occurred irregularly.

Example 5 is another good example of what we intend for a description. The disorder irregular menstrual cycle is implicit, but it is not stated explicitly.

1.2.3 A disorder is annotated even if it does not pertain to a patient.

All the disorders present in the text are to be annotated, even if they do not pertain to the patient or are presented in hypothetical or generic contexts.

Example 6
The patient was referred to the UBC Hospital Clinic for [Alzheimer Disease] and Related Disorders (UBCH CARD).

"*Alzheimer Disease*" is a candidate disorder even though the patient may or may not have Alzheimer[3] in example 6.

1.2.4 A candidate disorder is not linked to any syntactic construct.

A candidate disorder may appear in any syntactic context in the text. In example 7, the obesity disorder is mentioned through the adjective "obese".

Example 7

[2] SHARE guidelines, Section 1.3.
[3] The fact that this is part of an institution is coded in one of the disorder's modifiers, in the second SHARE annotation phase.

The patient was [obese].

Example 8
[Glaucoma] diagnosis was excluded.

Example 9
The [COVID-19] incubation period was researched by looking at diagnosed cases that have been publicly reported.

For instance, “Glaucoma” and “COVID-19” in the examples above (8-9) are in a modifying position (from the syntactic point of view); nonetheless, they can and must be marked as disorders.

2. Step 2: Search for UMLS concepts

Given a candidate disorder identified at step 1, annotators use the UMLS browser[4] to retrieve concepts that refer to it[5].

2.1 Search with the whole candidate disorder.

The search should first be performed with the whole candidate disorder (the maximum extent considered in step 1). If annotators find a lexical form in the terminology that exactly matches the text span of the candidate disorder or they find a synonym, they go directly to Step 3 (where they will most likely find themselves in Case A).

Example 10
(*The overall features of the tumor were most consistent with metastatic melanoma) occurring in a background of [Barretts esophagus with high-grade dysplasia].*

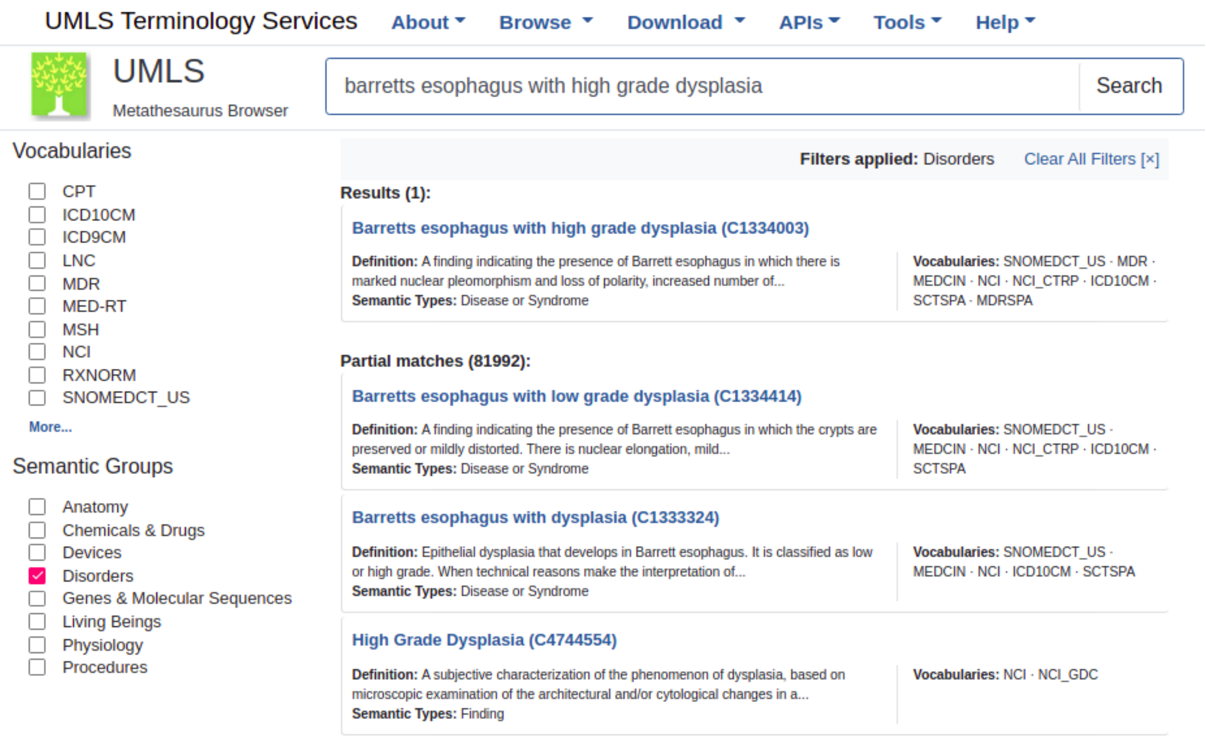
[Barretts esophagus with high grade dysplasia (C1334003)]
[Barretts esophagus with low grade dysplasia (C1334414)]
[Barretts esophagus with dysplasia (C1333324)]
[High Grade Dysplasia (C4744554)]
[Barrett Esophagus (C0004763)]

In example 10 we can see the original sentence in which “Barretts esophagus with high-grade dysplasia” appears as well as the candidate concepts UMLS offers to the search of “barretts esophagus with high grade dysplasia”. In this case, the concept in bold is the most adequate as it is a perfect match with the candidate disorder.

Tip: In order to ensure that they are choosing a concept that refers to a disorder, annotators are supposed to use the UMLS filter DISORDERS to exclude concepts that are not disorders. See in Figure 1 how the filter “Disorder” is selected (in the lower left side of the

[4] https://uts.nlm.nih.gov/uts/umls/home
[5] Note that the UMLS browser performs a number of normalizations on the input string and that results may include partial matches as well as lexical or syntactic variants.

image) and how the mention is entered in the search bar for example 10, where annotators will find the concept [Barrets esophagus with high grade of dysplasia (C1334003)].



barretts esophagus with high grade dysplasia

Vocabularies

☐ CPT
☐ ICD10CM
☐ ICD9CM
☐ LNC
☐ MDR
☐ MED-RT
☐ MSH
☐ NCI
☐ RXNORM
☐ SNOMEDCT_US

More...

Semantic Groups

☐ Anatomy
☐ Chemicals & Drugs
☐ Devices
☑ Disorders
☐ Genes & Molecular Sequences
☐ Living Beings
☐ Physiology
☐ Procedures

Filters applied: Disorders Clear All Filters [x]

Results (1):

Barretts esophagus with high grade dysplasia (C1334003)

Definition: A finding indicating the presence of Barrett esophagus in which there is marked nuclear pleomorphism and loss of polarity, increased number of...
Semantic Types: Disease or Syndrome
Vocabularies: SNOMEDCT_US · MDR · MEDCIN · NCI · NCI_CTRP · ICD10CM · SCTSPA · MDRSPA

Partial matches (81992):

Barretts esophagus with low grade dysplasia (C1334414)

Definition: A finding indicating the presence of Barrett esophagus in which the crypts are preserved or mildly distorted. There is nuclear elongation, mild...
Semantic Types: Disease or Syndrome
Vocabularies: SNOMEDCT_US · MEDCIN · NCI · NCI_CTRP · ICD10CM · SCTSPA

Barretts esophagus with dysplasia (C1333324)

Definition: Epithelial dysplasia that develops in Barrett esophagus. It is classified as low or high grade. When technical reasons make the interpretation of...
Semantic Types: Disease or Syndrome
Vocabularies: SNOMEDCT_US · MEDCIN · NCI · ICD10CM · SCTSPA

High Grade Dysplasia (C4744554)

Definition: A subjective characterization of the phenomenon of dysplasia, based on microscopic examination of the architectural and/or cytological changes in a...
Semantic Types: Finding
Vocabularies: NCI · NCI_GDC

Figure 1. Search of “barretts esophagus with high grade dysplasia” (Ex.10) in the “Disorders” category and concepts retrieved.

2.2 Searches with substrings of the candidate disorder.

If, on the other hand, no lexical form that exactly matches the text span of the candidate disorder or synonym is found, annotators should perform further searches with substrings of the candidate disorder. The span of text to be searched in UMLS can be reduced by removing syntactic modifiers. Modifiers are the elements in the candidate disorder that add extra specification to the head or main element of the disorder.

> Example 11
> *Here we report a case of [aggressive metastatic cholangiocarcinoma] in a 72-year-old man, sequentially treated with two targeted chemotherapies.*
> ***[Cholangiocarcinoma metastatic (C40874710)]***

In the case of example 11, the query for “*aggressive metastatic cholangiocarcinoma*” yields no results, so annotators leave out the modifier “aggressive” and, searching for *“metastatic cholangiocarcinoma”,* find a matching concept.

As many modifiers as needed can be removed from a candidate disorder in order to try to find a good match with a UMLS concept, as long as it still refers to a disorder (and still complying with the rules in Section 1.2).

Therefore, in regards to searches with substrings, proceeding to Step 3, annotators will then most likely find themselves in Case B or Case C. In some (rare) cases, none of the searches will give any results (Case D).

3. Step 3: Assign a code to a candidate disorder and select the final textual extent of the annotated Disorder

Annotators should select one (and only one) UMLS concept for each candidate disorder (i.e. the closest synonym)[6] and then identify the final extent according to the selected UMLS concept; they can be faced with one of the four following main cases:

Case A: Extent of the candidate disorder remains unchanged
If a perfect match (oor synonym) for the candidate disorder is found on UMLS, the CUI of the concept is assigned to the disorder and the extent does not vary.

Example 12
(*The overall features of the tumor were most consistent with metastatic melanoma) occurring in a background of [Barretts esophagus with high-grade dysplasia].*

[Barretts esophagus with high grade dysplasia (C1334003)]
[Barretts esophagus with low grade dysplasia (C1334414)]
[Barretts esophagus with dysplasia (C1333324)]
[High Grade Dysplasia (C4744554)]
[Barrett Esophagus (C0004763)]

Barrett esophagus with high-grade dysplasia = [Barretts esophagus with high grade dysplasia (C1334003)]

In the case of example 12, the UMLS concept matches the extent of the candidate disorder and the extent is therefore left unchanged.

Case B: Extent of the candidate disorder is reduced (by removing syntactic modifiers) to fit the UMLS concept.

If the UMLS concept is synonym to only a subpart of the candidate disorder, annotators reduce the extent of the candidate disorder accordingly.

Example 13
Here we report a case of [aggressive metastatic cholangiocarcinoma] in a 72-year-old man, sequentially treated with two targeted chemotherapies.

[Cholangiocarcinoma metastatic (C40874710)]
metastatic cholangiocarcinoma = [Cholangiocarcinoma metastatic (C40874710)]

[6] Annotators should use all the information provided by UMLS including, for example, definitions and synonyms.

The extent of the candidate disorder is reduced to “metastatic cholangiocarcinoma” in order to fit the UMLS concept [Cholangiocarcinoma metastatic (C40874710)] (remember from Section 2.2 that the query for “*aggressive metastatic cholangiocarcinoma*” yields no results).

Be aware that removing too many modifiers may lead to “losing” a candidate disorder mention. “Abnormal bronchial artery” in example 14 is potentially a disorder, but when a search is conducted on UMLS, no satisfactory results are obtained. However, if the modifier “abnormal” is removed, the remaining “bronchial artery” candidate does not refer anymore to a disease. In this case, the annotator may consider to annotate “abnormal bronchial artery” as a CUILESS disorder (as in Case D).

> Example 14
>
> *The TC showed an [abnormal bronchial artery].*

Case C: Candidate disorder is split into separate Disorders

If no synonym is found for the whole text span of the candidate disorder, but different subparts of it are disorders and do have a synonym concept in UMLS, then the candidate disorder can be split into two distinct mentions.

> Example 15
> *The histopathology revealed [teratoma] with [strumal carcinoid] IA stage according to AJCC 2010 of the right ovary.*
>
> Teratoma = [teratoma (C0039538)]
> Strumal carcinoid = [strumal carcinoid (C0334526)]

In example 15, “teratoma with strumal carcinoid” did not get any satisfactory results on UMLS, but its subparts “teratoma” and “strumal carcinoid” do. The annotator can choose to annotate both parts as individual disorders.

Case D: Candidate disorder is discarded or marked as CUIless

As stated above, all final annotated mentions should be linked with a UMLS concept. So, if no UMLS concept is found that can be linked to a candidate disorder, it means that this candidate is not really a disorder (the annotator erroneously marked it in Step 1) and it is discarded.

Only in a few cases, however, CUILESS mentions are admitted[7], i.e. when annotators are sure and find proof (in the text[8] or from external sources) that the candidate is a disorder even if it is not listed in UMLS. Note that CUIless mentions are extremely rare for English because the coverage of UMLS is very good.

> Example 16
> Subsequent gastroscopy revealed diffuse [edema], [ulcers] and [errhysis].

[7] This is very similar to what is described in the SHARE guidelines
[8] In this case, the whole text can be used

Errhysis = [CUILESS]

The search for "errhysis" of example 16 in UMLS does not produce any results. The annotator however is certain that "errhysis" is a mention of disorder; in fact, the context in which it appears, i.e. the result of a test, and domain knowledge (errhysis is a type of small bleeding) confirm this. "errhysis" is CUILESS.

3.1 Rules for the identification of matching UMLS concepts

The following rules should be followed when matching disorder candidates with UMLS concepts.

3.1.1 Synonymy does not imply exact match.

Synonymy[9] between a candidate disorder and a UMLS concept can hold even if there is no exact match, as in the case in example 17 between "arm swelling" and "swelling of arm".

Example 17
The same patient apparently was admitted three years back for excision of the [arm swelling] which was relatively small that time.

[Swelling of arm (C0577598)]
[Large arm swelling (C0542067)]
[swelling; arm, localized (C1411773)]

3.1.2 Domain knowledge and external sources of information can be used.

Domain knowledge can be used to find the matching UMLS concept. In some cases, especially if there is no exact match, a certain level of domain knowledge is actually needed to identify synonymy. To this aim, non domain-experts annotators may need to consult external sources of information.

Example 18
A [tumor] was found in the left ovary.

[Neoplasms (C0027651)]
- New abnormal growth of tissue

[Tumor Mass (C3273930)]
- A benign or malignant pathologic structure in any part of the body, resulting from a neoplastic accumulation of cells.

[Malignant Neoplasms (C0006826)]
- A term for diseases in which abnormal cells divide without control and can invade nearby tissues.

[Hydatidiform Mole (C0020217)]
- Trophoblastic hyperplasia associated with normal gestation, or molar pregnancy.

[Renal Cell Carcinoma (C0007134)]

[9] According to the SHARE guidelines, we interpret synonymy in terms of "reasonable synonymy".

- A heterogeneous group of sporadic or hereditary carcinoma derived from cells of the KIDNEYS.

For “tumor” in example 18, we choose “neoplasms” based on our research on external sources. Neoplasms is a synonym of tumor, as the definition of neoplasms is “new abnormal growth of tissue” on UMLS and “type of abnormal and excessive growth, called neoplasia, of tissue” in Wikipedia[10]. In addition, redirection to “tumor” is suggested.

Example 19
A 70yrs old Female was admitted in our institution diagnosed with [severe bilateral Osteoarthritis].

Osteoarthritis = [Degenerative polyarthritis (C0029408)]
Osteoarthritis of hip (C0029410)
Osteoarthritis, Knee (C0409959)
Generalized osteoarthritis (C1384584)
Osteoarthritis, Spine (C2350242)

In example 19, the initial candidate mention “severe bilateral Osteoarthritis” (between brackets) has been reduced to “Osteoarthritis”. [Degenerative polyarthritis (C0029408)] is selected because it is a synonym of *Osteoarthritis*[11].

3.1.3 Only the closest context of the candidate disorder can be used as context to make inferences

The context that annotators can use to find the matching UMLS concept for a Disorder is restricted to the closest context of the candidate disorder. More specifically, annotators are not allowed to use:

- knowledge extracted/derived/inferred from other disorders mentioned in the text;
- knowledge extracted/derived/inferred from other words intext that are not immediately next to the candidate disorder;
- Information about coreference.

For “closest context” we intend the modifiers inside the NP of the candidate disorder that are not part of the disorder mention.

Example 20
Further workup of the patient demonstrated multiple radiologic lesions consistent with metastases. Molecular studies demonstrated that the melanoma was positive for the mutation in the BRAF gene. The overall features of the [tumor] were most consistent with metastatic melanoma.

For the candidate disorder “tumor” in example 20, annotators should not use the context outside the span of the candidate disorder to infer that it is a “melanoma”. The correct concept to be selected is Neoplasms as in example 18.

Example 21
The x-rays showed [Varus Malalignment].

[10] https://en.wikipedia.org/wiki/Neoplasm
[11] https://radiopaedia.org/articles/investigation-of-polyarthritis-summary

[Bone Malalignment (C0206231)]
[Knee malalignment (C5195157)]
[Elbow malalignment (C3806546)]
[Dental malalignment (C4693873)]
[Misalignment of teeth (C1852504)]

In example 21 we have the candidate mention “Varus Malalignment”. There is no ULMS disorder either for “Varus Malalignment” or “Malalignment”. Domain knowledge tells us that a “varus malalignment” is necessarily a “bone malalignment” and therefore this is selected as the most informative concept for “Varus Malalignment”.

3.1.4 Links between a Disorder and a more specific UMLS concept must be avoided

Annotators must **never** link a Disorder to a more specific UMLS concept as this would add information (all the information must in fact be contained in the text span of the candidate disorder).

Example 22
The ultrasound examination revealed [ovarian enlargement].

[Ovarian enlargement (C0392039)]
[right ovarian enlargement (C3468873)]
[left ovarian enlargement (C3468856)]
[bilateral ovarian enlargement (C3468209)]
[Enlargement (morphologic abnormality) (C2711450)]

For the candidate disorder “ovarian enlargement” in example 22, annotators discard [right ovarian enlargement (C3468873)] (“right” adds extra information), [left ovarian enlargement (C3468856)] (“left” adds extra information), [bilateral ovarian enlargement (C3468209)] (“bilateral” adds extra information). Annotators can only choose between Ovarian enlargement (C0392039) and Enlargement (morphologic abnormality) (C2711450).

3.1.5 Priority should be given to the most informative UMLS concept

When in doubt, annotators should select the most informative UMLS concept.

In Ex. 22 (see above, Section 3.1.4), after discarding the too-specific UMLS concepts (which discretionally add information), annotators are left with two potential options:

- Link “enlargement” to [Enlargement (morphologic abnormality) (C2711450)]
- Link “ovarian enlargement” to [Ovarian enlargement (C0392039)]

As [Ovarian enlargement (C0392039)] is more informative than [Enlargement (morphologic abnormality) (C2711450)], [Ovarian enlargement (C0392039)] is the correct choice.

Example 23

Intraoral examination revealed an [ulcerative lesion] around the upper left first and second molars.

[Ulcerative cytomegalovirus lesion (C3874324)]
[ESOPHAGEAL ULCERATIVE LESION (C0743593)]
[FACE SKIN LESION ULCERATIVE (C0743788)]
[Ulcer (C0041582)]
[Lesion (C0221198)]

In example 23, [Ulcer (C0041582)] and [Lesion (C0221198)] are rather good UMLS concepts for "ulcerative lesion". However, as [Ulcer (C0041582)] gives more information about the type of disorder, the annotator chooses it.

3.1.6 Links between a Disorder and a more generic UMLS concept are admitted in certain cases

Annotators can link a Disorder to a more generic UMLS concept, provided that this is the most informative concept available and that no synonym is available for the candidate disorder or any of the reduced-extent options.

Example 24

X-ray showed [complete destruction of the upper humerus].

[Abnormality of the humerus (C4021742)]
[Bowed humerus (C1859460)]
[Fracture of upper end of humerus (C0435531)]
[Deformed humerus (C4025539)]
[Bone destruction (C0238790)]
[Periodontal destruction (C0236023)]

No synonym is available for the candidate disorder "complete destruction of the upper humerus" of example 24 or for any of the reduced-extent options (including the very generic concept "destruction") in spite of the long list of disorders retrieved from UMLS (the reported list is abbreviated for simplicity).
Based on the knowledge that the upper humerus is a bone, annotators should define the final extent "*destruction of the upper humerus"* and link it to the more generic concept [Bone destruction (C0238790)].
Note that [Bone destruction (C0238790)] is preferred to [Abnormality of the humerus (C4021742)] because it is more informative.

3.1.7 UMLS concepts containing negations and reference to the past are admitted

Annotators are allowed to link Disorders to UMLS concepts containing negations and reference to the past.[12]

Example 25
The baby presented [no light reflex at the tympanic membrane]

[12] In this, our annotation guidelines differ from the SHARE guidelines.

In example 25, the span “no light reflex at the tympanic membrane” is a disorder, and is linked to [No light reflex at tympanic membrane (C0576887)]. In fact, having no reflex is a disorder and having it is a healthy situation.

> Example 26
> [*Hx of stroke*].

The span “Hx of stroke” in example 26 is a disorder, linked to [History of cerebrovascular accident (C0559159)].

3.1.8 The text span of a disorder can be discontinuous

The span for a disorder can be discontinuous.

> Example 27
> *A 25-year-old man presented to the hospital with* [epigastric persistent pain].
>
> [Epigastric pain (C0232493)]
> [Pain (C0030193)]

“Epigastric persistent pain” from example 27 is not available in UMLS, so the extent of the candidate disorder must be reduced. Removing the modifier “persistent” from the extent would give “Epigastric… pain”, which is a discontinuous mention.

A single tag (a unique tag, not splitted) is assigned to the discontinuous disorder mention. The tag covers all the elements that should effectively be inside the span of the disorder mention. In the case of “epigastric persistent pain”, “epigastric” and “pain” are part of the mention and thus should be annotated. “Persistent”, though is between both tokens and, as a single tag needs to be assigned to “epigastric” and “pain”, “persistent” is consequently inside the span of the tag.

Hence, the disorder mention “epigastric persistent pain” is annotated as [epigastric persistent pain].

Nonetheless, in order to mark the discontinuous nature of the mention, annotators must add a DISCONTINUOUS flag to the tag.

In what refers to the assignment of a UMLS code, the search should be based on the actual mention (without the excluded modifiers). In the case of “epigastric persistent pain” the search should be conducted with “epigastric pain” as that is effectively the disorder mention. In consequence, the UMLS code chosen for “epigastric (persistent) pain” is the following: [Epigastric pain (C0232493)]

Possible discontinuous extents can be identified when the scope of a modifier spans over a conjunction (or disjunction) of disorders. We have identified four cases in which the conjunction (or disjunction) of two disorders may lead to discontinuous disorder spans:

- Modifier 1 and Modifier 2 “common-head-disorder” (28)

  Example 28

He had a significant [preorbital and facial oedema].
Preorobital oedema: DISCONTINUOUS; [Preorbital edema (C0151205)]
Facial oedema: [Facial edema (C0542571)]

- “Common-head-disorder” Modifier 1 and Modifier 2 (29)

  Example 29
  The patient had no [pain in walking and running].
  Pain in walking = pain provoked by walking (C1960732)
  Pain in running = pain provoked by running (C1960729)

- “CommonModifier” Disorder 1 and Disorder 2 (30)

  Example 30
  Laboratory tests revealed that she had [elevated bilirubin and liver enzyme levels].
  Elevated bilirubin levels
  Elevated liver enzyme levels
- Disorder 1 and Disorder 2 “CommonModifier”

  Example 31
  A 12-year-old arrived with [pain and swelling in the left of maxilla].
  Pain in the left of maxilla = maxilla pain (C0746434)
  Swelling in the left of maxilla = swelling of maxilla (C3693221)

Finally, note that, even if in SHARE they are considered disjoint disorders, we consider verbal phrases as disorder descriptions (see Section 1.2.2) rather than disorder mentions and we therefore do not annotate them.

Example 32
Her menstrual cycle occurred irregularly.

No disorder is mentioned in example 32 above according to our guidelines.

4. Annotation in languages other than English

These guidelines have been written focusing on the annotation of English texts, but the annotation of clinical entities in the rest of the languages of the E3C project should follow the same procedure. Nonetheless, we are aware of the resource limitations languages other than English have. More precisely, we are aware that there are less concepts in UMLS for Spanish, French, Italian and Basque. As a consequence, we have set up a little guide to help the annotators that deal with these languages.

1. Identify the full span of the candidate disorder in text
2. Search for it on UMLS
   a. For this step language specific dictionaries can be selected so as to restrict the search
3. As in English, if no satisfactory result is obtained, apply the modifier reduction steps to obtain a less specific candidate disorder and search for it.
4. If no satisfactory result is achieved for the searches in the original language, the candidate disorder should be translated into English (or other relevant language) and repeat the procedure.

5. If a satisfactory result is obtained from the search in English, the UMLS code should be assigned to the candidate disorder in the original text.

BIBLIOGRAPHY

- Noémie Elhadad, Guergana Savova, Wendy Chapman, Glenn Zaramba, David Harris, and Amy Vogel. 2012. ShARe Guidelines for the Annotation of Modifiers for Disorders in Clinical Notes. Technical report, Columbia University.

APPENDIX A

Non trivial annotation examples with explanations

Example 1

A [tumor] was found in the left ovary.

Even if [ovarian neoplasm] is available in UMLS, annotators will mark only [tumor] and map its corresponding concept [neoplasm] as they should use only the closest context of the candidate disorder as the context from which information can be extracted (it has been objected that the span of the candidate mention could have been “tumor was found in the left ovary”, but this would not be considered an acceptable candidate disorder).

Example 2

The x-rays showed bone on bone Tricompartment [OA Knee] with [Varus Malalignment].

[Bone Malalignment (C0206231)]
[Knee malalignment (C5195157)]
[Elbow malalignment (C3806546)]
[Dental malalignment (C4693873)]
[Misalignment of teeth (C1852504)]
(OA Knee = [Osteoarthritis, Knee (C0409959)])

Here we have the candidate mention “Varus Malalignment”. There is no ULMS disorder either for “Varus Malalignment” or “Malalignment”. Domain knowledge tells us that a “varus malalignment” is necessarily a “bone malalignment” and therefore this is selected as the most informative concept for “Varus Malalignment”.

Example 3

[Opacity] with cavernous lesion in the upper lobe.
[Opacity] in the upper lobe = [Decreased translucency (C0029053)]
Cavernous [lesion] in the upper lobe = [Lesion (C0221198)]
One disorder candidate is split into two disorders “opacity” and “cavernous lesion”, both in the upper lobe.

Example 4

We spotted [a small Osteochondral Cyst in the Anterior Femur].

[Osteochondral fracture (C0476169)]
[Osteochondral defects (C3495890)]
[Closed osteochondral fracture (C1997369)]
[Cyst (C0010709)]
[Epithelial cyst (C0014511)]

Here there is no synonym of the candidate mention “*small Osteochondral Cyst in the anterior Femur”*. After removing some of the modifiers, the annotator ends with “Osteochondral cyst”, the search of which offers the results in example 4. As an exact match is preferred, the annotator selects [Cyst (C0010709)], over the more generic “[Osteochondral defects (C3495890)]”, and assigns it to the final annotated Disorder “Cyst” (with extent “Cyst”, reduced with respect to the candidate mention).

Example 5
Based on this microscopical finding, the final diagnosis was [periodontal disease].
[Periodontal Diseases (C0031090)]
[Necrotizing periodontal disease (C3873591)]
[Periodontal Disease, CTCAE (C1559300)]
[Chronic Periodontitis (C0266929)]
[Gingival and periodontal disease (C0155936)]

For periodontal disease in Ex. 5, we choose the first UMLS entry since it is a synonym and does not contain any additional information. The acronym CTCAE stands for Common Terminology Criteria for Adverse Events[13] and is used to define disorders that are caused by adverse events. In this case, the periodontal disease is not the result of any adverse event and thus, the first entry is selected.

Example 6

The patient's chem profile reveals 138 sodium, 4.0 potassium, 106 chloride, 19 bicarb, 43 and 2.2 are the BUN and creatinine, indicating probable blood in the GI tract.

In Example 6, we find a description of the disorder “GI bleed”, but we do not annotate descriptions, so no candidate disorder is marked.

Example 7

In May 2014 a 49 year gentleman was admitted for widespread mucocutaneous blistering diagnosed as [PV] by histology and immunofluorescence.

[Polycythemia Vera (C0032463)]
[Vaginal Hemorrhage (C2979982)]
[Tinea Versicolor (C0040262)]
[Pemphigus Vulgaris (C0030809)]

In example 7 we have as a candidate mention an abbreviation, “PV”, and, among the concepts retrieved from UMLS, two are potential extensions of the abbreviation. Although later in the clinical case it will be clear that the correct concept is [Pemphigus Vulgaris (C0030809)], this information (i.e., coreference with other mentions) can not be used by the annotator. Hence, the annotator should leave the annotation CUILESS or try to guess which

[13] https://en.wikipedia.org/wiki/Common_Terminology_Criteria_for_Adverse_Events

of the two options causes “widespread mucocutaneous blistering” so as to choose the adequate code only using the **closest context** of the candidate disorder and world knowledge.

Example 8

The patient was referred to the [lupus] clinic.

[Lupus Vulgaris (C0024131)]
[Lupus Erythematosus (C0409974)]
[Lupus hepatitis (C0267807)]

In example 8, as in the previous one, the annotator was not able to choose among three UMLS concepts referring to the candidate mention “lupus”. Either leaves it CUILESS or tries to assign a UMLS code based on the closest context of the candidate disorder and world knowledge.

Example 9

Ignoring the prolonged exposure of the community to [MDR-TB] strain, he decided to cease the therapy.

Here “MDR-TB” is an apposition of the syntactic head “strain”. However, as there is no restrictions on the syntactic role of a disorder mention, MDR-TB must be annotated as a disorder.

Example 10

A chorion villus biopsy revealed the [mutation] causing late [familial hyperinsulinemic hypoglycemia].
Mutation = Disease-causing mutation (C2985434)
Familial hyperinsulinemic hypoglycemia = Hyperinsulinemic hypoglycemia, familial, 6 (C1847555)

In example 10, we have decided to break “mutation causing late Familial hyperinsulinemic hypoglycemia” into two separate candidate disorders as we could not find a satisfactory result in UMLS for the full span.

Example 11
Intraoral examination revealed an [ulcerative lesion] around the upper left first and second molars.

[Ulcerative cytomegalovirus lesion (C3874324)]
[ESOPHAGEAL ULCERATIVE LESION (C0743593)]
[FACE SKIN LESION ULCERATIVE (C0743788)]
[Ulcer (C0041582)]
[Lesion (C0221198)]

In example 11, [Ulcer (C0041582)] and [Lesion (C0221198)] are rather good UMLS concepts for “ulcerative lesion”. However, as [Ulcer (C0041582)] gives more information about the type of disorder, the annotator chooses it.