

ANNOTATION GUIDELINES FOR ENTITIES RELEVANT TO MENTAL ILLNESS

ANIKA OELLRICH

CONTENTS

1	INTRODUCTION	1
1.1	Motivation and background	1
1.2	Outline	2
1.3	Layout information	2
2	BRAT TO ANNOTATE ENTITIES OF INTERESTS	3
2.1	Usage of Brat	3
2.1.1	Assigning relationships between entities	4
3	ENTITIES RELEVANT TO MENTAL ILLNESSES	9
3.1	Age and age groups	9
3.2	Diseases	10
3.3	Drugs	11
3.4	Ethnicity and demographics	11
3.5	Gender	11
3.6	Measures and scales	12
3.6.1	Scales	12
3.7	Phenotypes (signs and symptoms)	13
3.8	Treatments	14
4	ANNOTATION PROCEDURE	15
4.1	Before annotation	15
4.2	During annotation	15
4.3	Annotation algorithm	16
4.4	Tree Diagram of Questions	17
5	DETAILS ABOUT QUESTIONS	21
6	THE RELEVANT ENTITIES IN DETAIL	25
6.1	Details	25
6.2	Age and age groups	26
6.2.1	Age groups as textual descriptions	27
6.3	Diseases	28
6.4	Drugs	30
6.5	Ethnicity and demographics	33
6.6	Gender	35
6.7	Measures and scales	36
6.8	Phenotypes (signs and symptoms)	38
6.9	Treatments	40
7	RESOLVING CONFLICTS DURING THE ANNOTATION PROCESS	43
7.1	Overlap of concepts	43
7.1.1	Examples	43
7.2	Missing identifiers in reference resources	44
7.2.1	Examples	44

LIST OF FIGURES

- Figure 1 Select the text span that corresponds to the entity of interest (here “cancer”) by clicking the left mouse button before the first character, keeping the mouse button pressed while moving to mark and releasing it after the last character of the entity. Note that spaces before or after the entity of interest should not be included in the annotation. 4
- Figure 2 Shortly after releasing the mouse button to highlight the entity of interest, a dialogue box appears that lets you select the entity type (here “disease”) and also pick the identifier. To use the build in normalisation support, choose the vocabulary you think the identifier is in from the drop down menu under “Normalization” (here “UMLS” as it is a disease) and click in the search box labelled “Click here to search”. 5
- Figure 3 A new dialogue box opens containing the entity that was just highlighted (here “cancer”). Activate the “Search UMLS” button to determine relevant identifiers for this term. 5
- Figure 4 The bottom half of the dialogue box has changed. It now provides a table with all possibly relevant concepts that have been identified in Unified Medical Language System (UMLS). Select the one that resembles the text span the closest (here “Primary malignant neoplasm”). If you are unsure as to which concept to pick from the list provided, you can read concept definitions using the Metathesaurus browser from the UMLS UTS service at <https://uts.nlm.nih.gov/metathesaurus.html#>. Once you have selected the appropriate concept from the list, the “ID” and “Query” field will be updated accordingly. Close this dialogue by confirming your choice and pressing “OK”. 6

Figure 5	Once you have closed the dialogue, the initial dialogue box with entity types is reappearing again, now showing the selected concept under “Normalization”. Make sure that the correct entity type is selected (here “disease” and close the dialogue box by confirming your choice with “OK”). 6
Figure 6	After confirming your choice of annotation, a green label on top of the previously selected text span appears indicating the entity type (here “disease”). The box is surrounded with an orange coloured outline; this only indicates that this was the last annotation that has been assigned. As you move on to the next text span of interest the orange outline will be moved to this one. 7
Figure 7	Click on the annotation that has been assigned beforehand (here a measure) and keep the mouse button pressed. An arrow will be shown as a result of this that then can be connected to the another previously assigned annotation (here the scale minimum for this measure). 7
Figure 8	Once the arrow is pointing on the second annotation, release the mouse button and a dialogue box will be opened. This dialogue box already contains the correct label for the relationship as the relationships have been defined prior to the annotation time. If no dialogue shows, you are most likely trying to connect to entities that should not be related with one another. 7
Figure 9	If you confirm the dialogue with “OK”, the arrow will be permanently added between the different parts of an annotation and the name appears with a orange-coloured background. This is only to indicate that this is the last element that has been added to the annotation display and will disappear as soon as another entity/relationship is assigned. 8
Figure 10	Annotation questions 19
Figure 11	Annotating details 26
Figure 12	Annotating age groups 26
Figure 13	Annotating diseases 29
Figure 14	Annotating diseases 29
Figure 15	Annotating drugs 31
Figure 16	Annotating drug (groups) list 32

Figure 17	Annotating ethnicities	34
Figure 18	Annotating demographics	34
Figure 19	Annotating gender	35
Figure 20	Annotating measures	37
Figure 21	Annotating continuous scales	37
Figure 22	Missing identifier in annotation tool	45

LIST OF TABLES

Table 1	Age groups	10
Table 2	Potential gender annotations	12
Table 3	Dementia-specific scoring algorithms	13

LISTINGS

ACRONYMS

AO	Age Ontology
CRIS	Central Record Interactive Search
CUI	Concept Unique Identifier
CUIs	Concept Unique Identifiers
EHR	Electronic Health Record
EHRs	Electronic Health Records
HPO	Human Phenotype Ontology
MeSH	Medical Subject Headings
UMLS	Unified Medical Language System

INTRODUCTION

1.1 MOTIVATION AND BACKGROUND

Mental illnesses are estimated to account for 11% to 27% of the disability burden in Europe (**Wykes2015**). In line with current aims to provide improved healthcare at lower costs, healthcare technologies are needed that support patient-centred healthcare, thus increasing patients' quality of life and assists clinicians and care givers in their daily work. Patient-centred healthcare includes aspects of recording treatment- and diagnosis-specific data recorded for a patient. However, the potential of including the latest research findings into treatment decisions is to date still underexplored. For example, the published scientific literature provides large amounts of peer-reviewed data that could inform treatment decisions taken by healthcare providers. Integrating Electronic Health Records (EHRs) of mentally ill patients with the scientific literature can contribute towards this goal, in particular by highlighting problems with medications, suggest medication alternatives, or foresee potential long-term issues with treatments.

In order to integrate both the published scientific literature and EHRs, methods are needed that provide the means for finding relevant literature relevant to a patient's situation. A first step to enable the search for relevant literature is the identification of health-status related information contained in publications. This health-status related information can focus on co-morbidities and related diagnoses, as well as treatments and ethnicity of the patient. The health-related information is summarised to entity groups of interests such as treatment and diseases. The entities by assessing the patient data contained in the Central Record Interactive Search (CRIS)¹. While the scope of another Electronic Health Record (EHR) system may differ, we believe that the chosen subset of entities outlined in this guide provide a good representation of mental illnesses.

The current document is the first version of annotation guideline for entities relevant to the domain of mental illnesses. These guidelines have not been used before and therefore constitute work in progress and may be refined over time.

¹ <http://www.slam.nhs.uk/about/core-facilities/cris>

1.2 OUTLINE

The annotation guidelines give first an overview about the entities of interest relevant to mental illnesses (see chapter 3), then continues with an explanation to the annotation algorithm (see chapter 4 and 5), before going into detailed explanations and examples for the different entity types (see chapter 6). Additional guidance on potential conflicts of and overlapping annotations is provided in chapter 7. It is strongly recommended to read the individual chapters in order. Chapter 2 provides an introduction to the annotation interface. We note here that the choice of annotation interface informed the development of these guidelines in places and adaptations may be required if a different annotation tool is used.

1.3 LAYOUT INFORMATION

Throughout these guidelines, we aim to adhere to a consistent layout to help with the interpretation of the content. We have inserted examples to show case the use of the specific entity types and illustrate edge case. Examples are surrounded with boxes to separate them from the rest of the text and are typically followed by a small explanation and a figure illustrating the expected annotations for the example. Entity types (Age/Age group, Diseases, Demographics/Ethnicity, Drugs, Gender, Measures and scales, Phenotypes, and Treatments) are highlighted after their introduction through an italicised font type. However, entities of interest in itself are not highlighted, only the entity types. Both layout choices are illustrated in the following:

This is an example containing a *disease* entity.

BRAT TO ANNOTATE ENTITIES OF INTERESTS

This document discusses the annotations of entities that are relevant to the domain of mental illnesses and contained in the full text of scientific publications. It details the entities relevant to mental illnesses as well as the annotation procedure to help create consistent annotations in a corpus. We note here that these annotations guidelines have been developed to annotate 50 papers concerning research on mental illness (in a first attempt dementia) of the Open Access subset of the PubMed Central articles¹. The document is addressed to potential curators and researchers interested in this work.

Even though several tools for manual text annotation have been developed (Neves2014), these guidelines have been used in conjunction with the Brat annotation tool. Brat allows for entity as well as relationship annotations and some bits of the suggested annotation structure may only make sense in the context of this annotation tool.

2.1 USAGE OF BRAT

A curator following these annotation guidelines has to fulfil three steps to assign an annotation:

1. identify the text span that corresponds to an entity of interest or parts thereof,
2. determine the identifier (either a [UMLS](#) Concept Unique Identifier (CUI) or an Human Phenotype Ontology (HPO) or DrugBank identifier) that needs to be assigned to the entity of interest or its parts and,
3. if necessary, assign a relationship between the individual parts.

To allow easy annotation in line with these guidelines, we set up the annotation tool Brat² on a server, arranging for each curator to be able to sign in. Furthermore, this Brat instance has been enriched with dictionaries from [UMLS](#), [HPO](#) and DrugBank for normalisation so that annotators can retrieve identifiers quicker from the corresponding reference resources.

The process of assigning an annotation to a text span with Brat is simple: highlight the text that corresponds to the entity of interest you have identified with your mouse cursor, once highlighted wait

¹ <http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>

² <http://brat.nlplab.org/>

for a dialogue box to open, select the entity type from the list provided and use the lookup features for choosing an identifier for the text span. Note that sometimes there may be overlapping annotations (further explained in the relevant subsections), in which case multiple annotations need to be assigned to cover all the different identifiers. However, overlapping annotations should be only used where necessary and not to express uncertainty. If you are uncertain about an annotation, choose for one according to the best of your knowledge and use this one. The process of annotating a text span in Brat is further illustrated in figures 1 to 6.

2.1.1.1 *Assigning relationships between entities*

In some case it is necessary to build an annotation from multiple parts. In order to achieve that, highlight the text first of the individual parts of this annotation (as described before), e.g. a scale and its minimum value, and then assign the relationship between both. In order to do so, click on the first part of the annotation and an arrow appears and then connect to the relevant other part of the entity. This opens a separate dialogue box with the automatically recognised relationship, which you only have to confirm with “OK”. As a result an arrow pointing from one entity to the other is visible, named with the appropriate relationship. This process is further illustrated in figures 7 to 9.

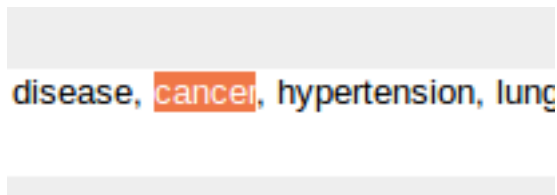


Figure 1: Select the text span that corresponds to the entity of interest (here “cancer”) by clicking the left mouse button before the first character, keeping the mouse button pressed while moving to mark and releasing it after the last character of the entity. Note that spaces before or after the entity of interest should not be included in the annotation.

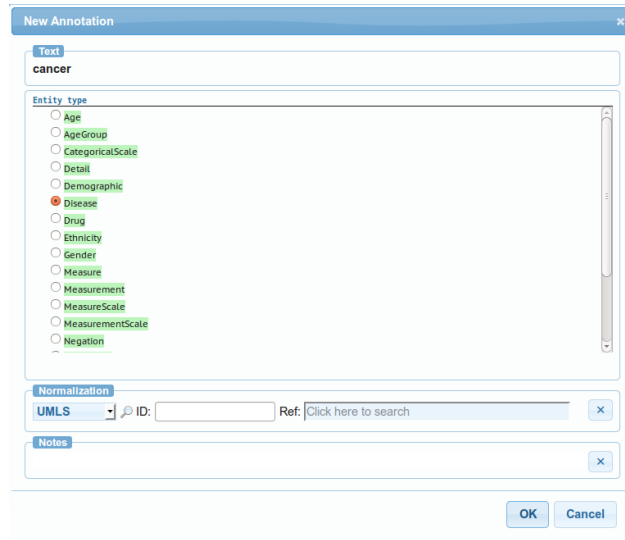


Figure 2: Shortly after releasing the mouse button to highlight the entity of interest, a dialogue box appears that lets you select the entity type (here “disease”) and also pick the identifier. To use the build in normalisation support, choose the vocabulary you think the identifier is in from the drop down menu under “Normalization” (here “UMLS” as it is a disease) and click in the search box labelled “Click here to search”.

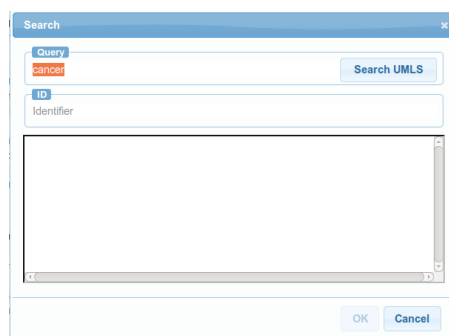


Figure 3: A new dialogue box opens containing the entity that was just highlighted (here “cancer”). Activate the “Search UMLS” button to determine relevant identifiers for this term.

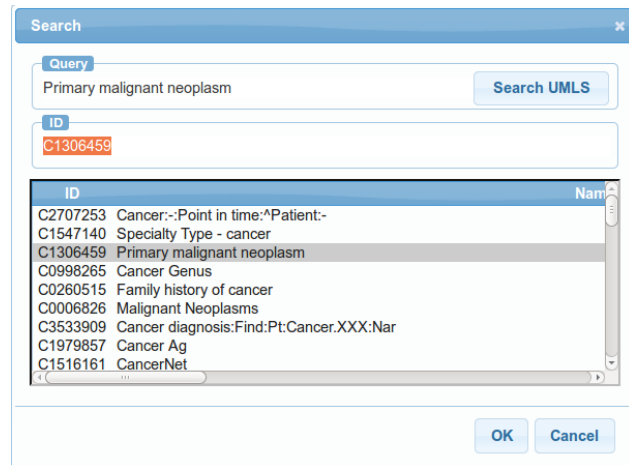


Figure 4: The bottom half of the dialogue box has changed. It now provides a table with all possibly relevant concepts that have been identified in [UMLS](#). Select the one that resembles the text span the closest (here “Primary malignant neoplasm”). If you are unsure as to which concept to pick from the list provided, you can read concept definitions using the Metathesaurus browser from the [UMLS](#) UTS service at <https://uts.nlm.nih.gov/metathesaurus.html#>. Once you have selected the appropriate concept from the list, the “ID” and “Query” field will be updated accordingly. Close this dialogue by confirming your choice and pressing “OK”.

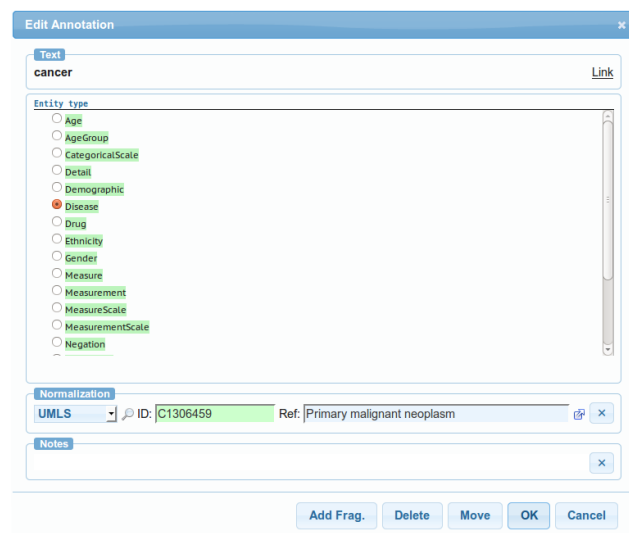


Figure 5: Once you have closed the dialogue, the initial dialogue box with entity types is reappearing again, now showing the selected concept under “Normalization”. Make sure that the correct entity type is selected (here “disease” and close the dialogue box by confirming your choice with “OK”).

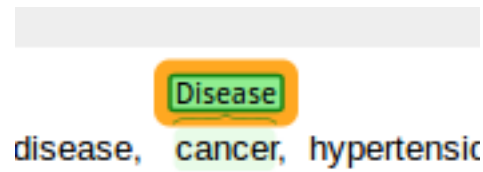


Figure 6: After confirming your choice of annotation, a green label on top of the previously selected text span appears indicating the entity type (here “disease”). The box is surrounded with an orange coloured outline; this only indicates that this was the last annotation that has been assigned. As you move on to the next text span of interest the orange outline will be moved to this one.

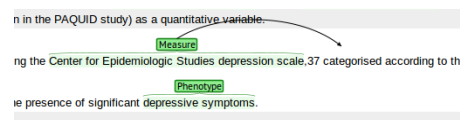


Figure 7: Click on the annotation that has been assigned beforehand (here a measure) and keep the mouse button pressed. An arrow will be shown as a result of this that then can be connected to the another previously assigned annotation (here the scale minimum for this measure).

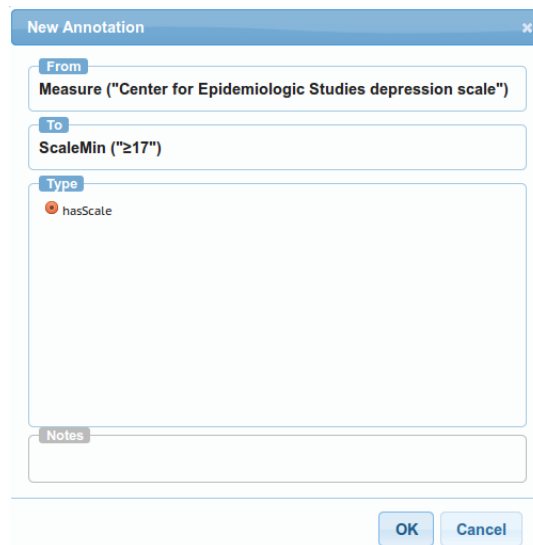


Figure 8: Once the arrow is pointing on the second annotation, release the mouse button and a dialogue box will be opened. This dialogue box already contains the correct label for the relationship as the relationships have been defined prior to the annotation time. If no dialogue shows, you are most likely trying to connect to entities that should not be related with one another.

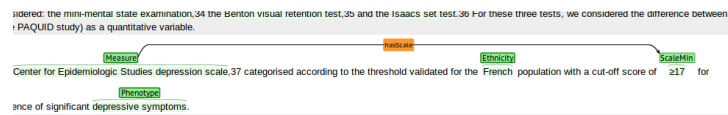


Figure 9: If you confirm the dialogue with “OK”, the arrow will be permanently added between the different parts of an annotation and the name appears with a orange-coloured background. This is only to indicate that this is the last element that has been added to the annotation display and will disappear as soon as another entity/relationship is assigned.

ENTITIES RELEVANT TO MENTAL ILLNESSES

The entities described in this chapter have been identified as relevant to the description and treatment of patients suffering from mental illnesses. These entities have been determined using the structured and unstructured information provided in [CRIS](#) and cross-referenced for their occurrence in the published scientific literature. Due its wide application and existing software for automation, [UMLS](#) was chosen as the terminology to normalise to. An exception are drugs and phenotypes that should be normalised to DrugBank¹ and the [HPO](#)² respectively. Any curator employing these guidelines is expected to be familiar with [UMLS](#) (the concepts contained and the semantic types), and HPO and DrugBank. It is further expected that curators hold their own [UMLS](#) licence as there is none issued together with these guidelines. As a point of reference and for browsing concept definitions, the UMLS Metathesaurus browser³ can be used.

3.1 AGE AND AGE GROUPS

The age of a patient or study object refers to the time that has passed since their birth ([Geifman2011](#)). In studies that concern more than one individual, ages of individuals may be summarised into an age group, e.g. “adults” or “70-74”. Age groups, depending on the way they are presented in the article text, can refer to an age range, e.g. the age group *adult* could correspond to the age range 19 to 44 years ([Geifman2011](#)). However, while [Geifman2011](#) developed an Age Ontology ([AO](#)) and the Medical Subject Headings ([MeSH](#))⁴ also contain age groups, the information provided in the published scientific literature may differ from both. For example, while the [MeSH](#) terminology defines the term *elderly* as someone aged 65 years and older, in the literature for study objects older than 60 years, the term *elderly* may be used (see more information and examples on this in [6.2](#)).

The age group definitions used as a basis for these guidelines are listed in table [1](#). Both the name of the age group and the age range may occur in text. In both cases, an age group annotation should be assigned with the corresponding [CUI](#). As a side note, most of the age groups defined here fall under the semantic type “Age group (T100)”, although *aged adults* (referred to as “elderly” in [UMLS](#)) appears to be a “Population group (T098)” instead of an “Age group (T100)”.

¹ <http://www.drugbank.ca/about>

² <http://human-phenotype-ontology.github.io/>

³ <https://uts.nlm.nih.gov/home.html>

⁴ <https://www.nlm.nih.gov/mesh/>

AGE GROUP	AGE RANGE	UMLS CUI
newborn	0 – 1 monts	C0021289
infant	1 – 23 months	C0021270
preschool children	2 – 5 years	C0008100
children	6 – 12 years	C0008059
adolescent	13 – 18 years	C0205653
young adult	19 – 24 years	C0238598
adult	19 – 44	C0001675
middle aged adult	45 – 64 years	C0683973
aged adult (elderly)	65 – 79 years	C0001792
adult, aged 80 and over	> 80 years	C0001795

Table 1: Age groups derived from [UMLS](#).

We note here that **Geifman2011** distinguishes different ages relevant for patients: “age of onset”, “age of diagnosis”, “age of observation”, “age of occurrence”, and “age of evaluation”. While this is a useful distinction and has been evaluated through text mining, we do not employ this distinction in the current version of these guidelines.

3.2 DISEASES

According to the Oxford Dictionary, a disease is “a disorder of structure or function in a human, animal, or plant, especially one that produces specific symptoms or that affects a specific location and is not simply a direct result of physical injury [...]”⁵. This definition already clearly highlights the difference between a disease and the exhibited phenotypes (see 3.7 for more information) that lead to the diagnosis of the disease. For example, *Dementia* would be a disease and *memory loss* could be one of the observed phenotypes. In a way, diseases are collections of phenotypes.

In an clinical environment, which may have been the environment in which a scientific study reported in a paper has been conducted, diseases in itself can be part of the symptom description of another disease, e.g. depression is in some cases mentioned as a symptom of a disease⁶. However, for the purpose of annotating mental illnesses, we aim to distinguish both types of entities. For more information on how best to distinguish both entity types, see chapter 7. As a rule of thumb, any illness (irrespective of whether this is a mental illness or

⁵ <http://www.oxforddictionaries.com/definition/english/disease>, definition 1, accessed 14 October 2015

⁶ <http://www.nhs.uk/Conditions/dementia-guide/Pages/symptoms-of-dementia.aspx>, accessed 14 October 2015

not) that is defined within the UMLS semantic types “Disease and syndromes (T047)” and “Congenital abnormality (T019)”, would need to be annotated in a publication.

3.3 DRUGS

Drug as opposed to treatment are medications that are non-herbal, synthetically engineered and reported in DrugBank or under the UMLS semantic types “Organic Chemical (T109)” or “Pharmacologic Substance (T121)”. Herbal remedies, such as e.g. administered in Traditional Chinese Medicine (TCM) fall under the entity of *treatment* (see 3.8). For example, *Citalopram* (DrugBank: DB00215) would be considered a drug, whereas *Magnolia glauca* (UMLS: C0996638) would fall under the treatment entity type. Furthermore, this group does not cover any non-medication treatments, such as psychological interventions (e.g. guided imagery, C0150627).

3.4 ETHNICITY AND DEMOGRAPHICS

Ethnicity and demographics are two separate entity types, but are summarised here in one section. Ethnicity describes the origins of an individual or group (these are mostly study objects) of individuals that with respect to their ancestral, social or cultural roots, e.g. “Caucasian” or “Asian”. We use here the ethnic groups that have been defined with the UMLS semantic type “Population Group (T098)” to refer to the ethnicity of study objects. In some cases, a country (e.g. “India”) is mentioned as opposed to ethnicity (e.g. “Indian”). If the name of a country is used to refer to the origin of a study object, an ethnicity annotation should be assigned. However, if study results are accumulated per country or a country is referenced for where the study has taken place, then a demographics annotation is expected to be assigned. The corresponding UMLS semantic type for demographics is “Geographic Area (T083)”. For more information on distinguishing both types of annotations, see section 6.5.

3.5 GENDER

Another patient-specific aspect that may be relevant to treatment and may be reported in the published literature is gender. Text spans belonging into this group are not only *female* and *male*, but also *intersex* and *transgender* based on the exhibition of gender-specific reproductive organs as well as a felt belonging to a gender group. In particular, a *transgendered* person is defined as someone that feels belonging to a different gender than being born with, but does not have to undergo any gender-altering therapies. As there may be a prevalence for certain mental illnesses that are relevant to the extended gender

groups, all four groups need to be distinguished. If a study does not report about gender, then no gender-specific annotations are required. However, sometimes multiple genders may be included in the study in which all genders need to be identified that have been reported. The following table (table 2) lists the genders that should be annotated within publications and contains also the corresponding Concept Unique Identifiers (CUIs) for reference.

GENDER	UMLS CUI
female	C0015780
male	C0086582
intersex	C1704620
transgender	C0558141

Table 2: Illustrates the genders of study objects that may be mentioned and should be annotated within a publication.

3.6 MEASURES AND SCALES

A *measures* as defined in the context of these annotation guidelines is the standard, system, or method by which something is measured. While there are lots of different measures existing, we are limiting ourselves here to measures that are used in the context of mental illness, such as those for the diagnosis or the determination of severity of an illness, e.g. the Personal Health Questionnaire-9 (PHQ-9) for depression. To ease the annotation process and because the initial application of these guidelines to a selection of papers reporting on Dementia research, table 3 lists a selection of measures relevant to this domain. Note that this list is incomplete though and other measures may well be used in the literature that still need to be annotated even though not specifically listed here.

Measures commonly used for reporting the outcome of an assessment of a patient with mental illness, mostly fall under the UMLS semantic type “Diagnostic procedure (T060)” and the semantic type “Intellectual Product (T170)”. However, not all the measures reported are included in UMLS as concepts, in which case the measures, measurements and values should be still annotated, but the concept mapping should be left open.

3.6.1 Scales

For some of the measures, a scale is provided to aid the interpretation of the results/findings in the paper or to specify the criteria for in-/exclusions of patients into certain studies. As this information is

TEST	ACRONYM	UMLS CUI
Mini Mental State Examination	MMSE	C0451306
Montreal cognitive assessment	MoCA	C3496286
Abbreviated mental test score	AMTS	C2960765
General practitioner assessment of cognition	GPCOG	
Addenbrookes cognitive examination-III	ACE-III	
6-Item cognitive impairment test	6CIT	C1319423
Mini-cog		

Table 3: Lists the algorithms used for scoring specific signs and symptoms of dementia. Also used to report progression of dementia.

crucial to the results/findings, measurement scales need to be also annotated. Scales used to describe measures in a mental health context can be both continuous and categorical. To annotate continuous scales, the scale minimum and maximum need to be identified in text as well as the corresponding explanation. For categorical scales, all categories with their explanation need to be identified if provided within the text. Note that only scales relevant to the employed measures need to be annotated, but not other scales that may be defined in the paper, e.g. possible answers in a questionnaire. More information and examples on how to annotate measures and scales is provided in section 6.7.

3.7 PHENOTYPES (SIGNS AND SYMPTOMS)

Phenotypes are the signs and symptoms of an illness that are either patient- or clinician-reported, but both aim at the characterisation of a patient for diagnostic purposes and severity assessments. For example, a “blue eye colour” is a phenotype as well as “fatigue”. Phenotypes can also help to report the progress of a disease (which may be either an improvement or worsening of the initial, or added symptoms).

[UMLS](#) itself describes phenotype as a physical finding that can be counted as objective evidence of disease, observed by a health care provider. The same definition is to be applied here and phenotypes should be clearly distinguished from diseases (see 3.2). Note that [UMLS](#) has not yet completely included [HPO](#), both [UMLS](#) and [HPO](#) should be used for assigning phenotype annotations to the published scientific literature with a priority on [HPO](#) annotations if both resources contain the phenotype. If neither resource contains the relevant phenotype, an annotation still needs to be assigned, just without an identifier from either of the resources. As a rule of thumb, phenotypes

conform roughly to the [UMLS](#) semantic type “Finding (T033)” and “Sign or Symptom (T184)”. More information and examples for this group of entities are provided in section [6.8](#).

3.8 TREATMENTS

Any treatment method that does not fall under the drug entity group falls into the treatment entity group. This may be herbal substances as well as psychological interventions, such as Cognitive Behaviour Therapy (CBT; [CUI](#): C0009244) or guided imagery ([CUI](#): C0150627). [UMLS](#) semantic types relevant to this entity group are: “Therapeutic or Preventative Procedure (T061)”, “Organic Chemical (T109)”, and “Pharmacologic Substance (T121)”. More information and examples are provided in section [6.9](#).

ANNOTATION PROCEDURE

4.1 BEFORE ANNOTATION

Before starting to assign annotations, it is recommended that the paper (provided as PDF file separately from annotation tool) is read entirely to understand the content of the paper and the intended meaning of entities contained. It is not necessary to understand the paper in great detail, but a knowledge about the following is required

- questions addressed in the paper and how entities relate to that
- what are the relevant entity groups addressed in the paper
- what is the context required to determine the meaning of an entity (e.g. *“feeling cold”* vs *“having a cold”*)

It is expected that curators are familiar with [UMLS](#), [HPO](#) and Drug-Bank, and have access to the online browser¹ in case clarification for a concept is needed. This entails that a curator using these guidelines must have obtained a license for the use of [UMLS](#). While the annotation interface will provide limited support, it is impossible to provide full definitions and the links to relevant resources at the time of curation. It is further expected that annotations to one paper are finished before progressing to the next. All papers are provided through the annotation interface and do not need to be uploaded by the curator.

4.2 DURING ANNOTATION

Entity annotations are to be assigned within the context, but without taking the structure of the document into consideration. The following guidelines should be kept in mind while annotating:

- Always identify the longest possible text span corresponding to a unique entity (e.g. *“breast cancer”* vs *“left breast cancer”*).
- Entities should only be annotated in an overlapping fashion if this is really necessary, i.e. if two distinct entities are mentioned and one is not encompassed by the other (see [7.1](#) for more details).
- While the annotation guide suggests reference resources for the annotation of each of the entity groups, if the reference resource

¹ <https://uts.nlm.nih.gov/home.html>

is too imprecise or does not contain a fitting concept, an annotation is to be assigned with the correct entity type, but without an identifier.

- Entities should be located on a per sentence level, i.e. entities cannot span across sentences for any of the entity types.
- A paper does not have to cover all entity types, but it is necessary to annotate all entities contained in a paper, even if it is not directly relevant to the hypothesis of the paper or mentioned multiple times.
- If an entity appears multiple times with the same intended meaning in the paper, it is **not** sufficient to annotate it only once. Multiple occurrences of one entity need to be annotated with a consistent annotation.
- Entities of interest as part of other words (e.g. “dementia-free” or “benzodiazepine-dementia association”) should not be annotated.

Note that the lists of [CUIs](#) provided in this document are by no means complete. This information has been provided to speed up curation and clarification on potentially controversial aspects of annotations. The curator is expected to check the [UMLS](#) catalogue, [HPO](#) and DrugBank to find the best fitting identifier if the text span that was recognised as one of the entities of relevance and the identifier has not been referenced in these guidelines. Only if there is no entry in the in any of the reference resources ([UMLS](#) catalogue, [HPO](#) or DrugBank) should the identifier field be left blank.

4.3 ANNOTATION ALGORITHM

As mentioned before, entity occurrences are restricted to sentences, by means that entities should only be annotated if the entire entity falls within the boundary of a sentence. However, while assigning annotations, be precise as possible on the boundaries and do not mark the entire sentence as the entity if it is not necessary. To aid the annotation process, a curation tool Brat² is set up as a Amazon Web Service. Each annotator is assigned their own login and can add annotations according to their specific guidelines, independently from annotations assigned by all the other curators.

² <http://brat.nlpplab.org/>

4.4 TREE DIAGRAM OF QUESTIONS

Questions (a)-(b) address the identification of entities within a sentence while questions (c)-(f) address the type and meaning of an entity (all illustrated in figure 10):

(a) *Does this sentence contain one or more relevant entities?*

- ⇒ If this sentence contains text spans that could potentially be relevant to any of the eight entity groups (see 4), proceed to question (b).
- ⇒ If not, continue with the next sentence.

(b) *What are the correct boundaries for each the relevant entities?*

- ⇒ For each of the relevant entities with boundaries, answer questions (c)-(g).

(c) *To which entity type does the identified text span belong?*

- ⇒ If the entity belongs to study object-related data such as *age, age groups, ethnicity and gender*, proceed to question d.
- ⇒ If the entity belongs to diseases-related data such as *signs and symptoms or diseases*, proceed to question e.
- ⇒ If the entity constitutes a potential *cure* for signs, symptoms or diseases, proceed to question f.
- ⇒ If the entity belongs to either measures, measurements, or scale, proceed to question g.
- ⇒ If the entity is demographic information, proceed to question h.

(d) *Does the entity represent a study object-related information such as age, age group, ethnicity or gender?*

- ⇒ If the entity represents the age of an individual, assign an age annotation.
- ⇒ If the entity is an age range/group, assign an age group relationship.
- ⇒ If the entity is an ethnicity, assign an ethnicity annotation.
- ⇒ If the entity is an gender, assign a gender annotation.

(e) Does the entity represent signs, symptoms or diseases?

⇒ If the entity represents a sign or symptom, assign a phenotype annotation.

⇒ If the entity represents a disease, assign a disease annotation.

(f) Does the entity constitute a potential cure for signs, symptoms or diseases?

⇒ If the entity is used for treatment and constitute a drug, assign a drug annotation.

⇒ If the entity is used for treatment and is not listed as a drug, assign a treatment annotation.

(g) Does the entity constitute a measure or scale?

⇒ If the entity is a measure, assign a measure annotation.

⇒ If the entity refers to a scale, assign the corresponding scale annotation(s).

(h) Does the entity constitute demographic data?

⇒ If the entity is used to refer to a country without referring to the study objects, assign a demographics annotation.

⇒ If the entity does not belong into this category, it should not be assigned an annotation as all other options have been ruled out before.

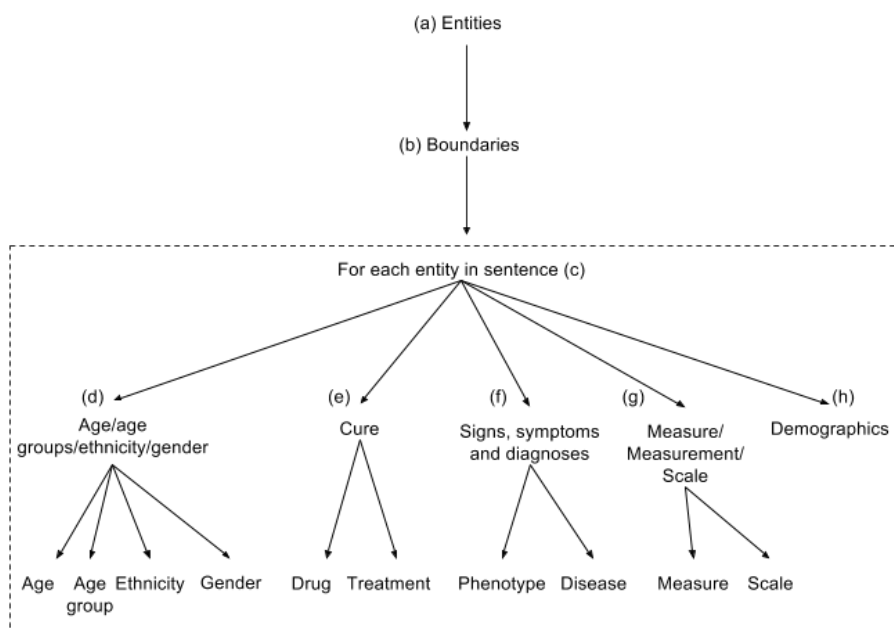


Figure 10: Decision tree to annotate text span of sentences with entity annotations.

DETAILS ABOUT QUESTIONS

The questions illustrated in figure 10 are explained here in further detail to clarify the decision needed to be taken when annotating entities of interest for mental illnesses.

(A) DOES THIS SENTENCE CONTAIN ONE OR MORE RELEVANT ENTITIES?

Each sentence needs to be investigated as to whether it contains one or more entities that are relevant to the entity types identified within these guidelines (see 3). If there are no entities contained within one sentence, then it can be progressed with the next sentence of the paper. If there are one or more entities per sentence, then for each of the entities to correct boundaries need to be identified (see section 5).

(B) WHAT ARE THE CORRECT BOUNDARIES FOR EACH THE RELEVANT ENTITIES?

For each entity in one sentence, the correct boundaries need to be identified. Curators are expected to annotate as little overlapping annotations as possible, though it is necessary in some cases to let the annotations overlap (for more information on overlaps see section 7.1). While concepts in UMLS are mostly reported in singular, if they occur as plural in the text, the boundary of the entity would include the plural instead of only reporting the singular. For example, consider the following sentence:

[...] and second, the extent to which compositional factors (sociodemographic variables and health states) might account for geographical variation in disability scores.

In the context of the paper the example sentence is taken from, “disability scores” are *measurements* of interest. In this case the correct boundaries would include the “s” indicating the plural of the word. Leading and following spaces for entities of interest should not be included within the annotation. Once the correct text span has been identified for annotation, progress to determine the type of the annotation before deciding on the identifier (any of the following CUI, HPO or DrugBank identifier).

Another aspect to consider are abbreviations. Whereever abbreviations are used in conjunction with a term that needs to be annotated

with an entity of interest annotation, the abbreviation should be annotated as part of the annotation. The following example illustrates this:

Segal, Williams and Teasdale [15] adapted the program for patients with recurrent depression, naming it ‘Mindfulness Based Cognitive Therapy’ (MBCT).

In this sentence, the entire text span “Mindfulness Based Cognitive Therapy’ (MBCT)” needs to be annotated with a *treatment* annotation. For more information on *treatments* see sections ?? and ??.

(C) TO WHICH ENTITY TYPE DOES THE IDENTIFIED TEXT SPAN BELONG?

In order to determine the type of the entity, it is necessary to start to distinguish the major groups of entities: disease-related information (disease names and signs and symptoms that can be exhibited by a patient), a treatment for a disease (whereby this could either be a drug name, psychotherapy approach or even alternative treatments), characteristics of individuals other than sign or symptoms (e.g. age or age groups, ethnicity, gender), measures/measurements relating to diagnosis and determination of severity of disease or demographic data.

(D) DOES THE ENTITY REPRESENT A STUDY OBJECT-RELATED INFORMATION SUCH AS AGE, AGE GROUP, ETHNICITY OR GENDER?

If a piece of study object-related information has been identified within a sentence, this limits the options to *age or age group*, *ethnicity* or *gender*. These entities are most often contained in methods and material sections of papers, but may also be mentioned in other sections including the abstract.

(E) DOES THE ENTITY REPRESENT SIGNS, SYMPTOMS OR DISEASES?

If an entity has been identified as disease-relevant data, then the entity types are limited to *diseases* or *phenotypes*. Disease-relevant information is equally likely in all part of the paper. For example, this data can either provide background information to the disease studied, provided in-/exclusion criteria for study objects (usually materials and methods), or serve as disintguisher in discussions. Independent from where this information occurs within the paper, an annotation needs to be assigned and a classification needs to occur. *Diseases* and *phenotypes* are distinct in that *phenotypes* are characteristics of *diseases*

and as well as patients, based on which a diagnosis can be made or the progression of a *disease* can be assessed. For more information on the distinction of these two entities see sections 6.3, 6.8 or ??.

(F) DOES THE ENTITY CONSTITUTE A POTENTIAL CURE FOR SIGNS, SYMPTOMS OR DISEASES?

If a text span refers to a treatment, then two options are available for annotations: *drugs* and *treatments*. The distinction between both is that one is chemically synthesised whereas the other group could be anything else, such as psychotherapies or alternative treatments. Further information on both types of entities is provided in sections 6.4 and 6.9. If the focus of the paper is a treatment or (group of) medication(s) of a mental illness then this particular treatment or (group of) medication(s) will be mentioned throughout the entire paper. Otherwise treatments are likely to occur through the entirety of a paper as it can be background information as well as form part of the discussion. Treatment information can also be shortly referenced in materials and methods when describing in-/exclusion criteria for study objects.

(G) DOES THE ENTITY CONSTITUTE A MEASURE OR SCALE?

If the entity refers to measures, a decision needs to be taken as to whether the entity is a *measure* or refers to the *scale* of a measure. For both *measure* and *scale*, there are different annotations and *scale* annotations may even have to be constructed out of multiple different annotations (see sections 3.6 and 6.7 for more information on this). If measures are part of the investigation, measures are likely to occur in the entire paper including the abstract. If a paper uses scores only to assess study objects, then a score may be mentioned only shortly in methods and/or results.

(H) DOES THE ENTITY CONSTITUTE DEMOGRAPHIC DATA?

If all the other entity types have been excluded, the entity last option left is demographic data. These entities usually corresponds to countries, e.g. when providing country-wide summarised research results or providing environmental factors of study objects. For example:

“Pension coverage was especially low in the Dominican Republic, rural Peru, rural Mexico, rural China, and India; food insecurity was prevalent in these sites.”

In this sentence, all country names should be annotated as a *demographic* entity as the sentence does not directly relate to the origins

of the study objects but instead relates to environmental factors (e.g. “low pension coverage”).

THE RELEVANT ENTITIES IN DETAIL

This chapter provides information on how annotations for the different entity types are expected to be assigned and illustrates these explanations with examples and figures. While most of the relevant text spans are expected to be contained in the chosen reference resources ([UMLS](#), [HPO](#), and [DrugBank](#)), there may be a small number of entities referenced in the literature that is not contained in any of the resources. If this should happen, an annotation should be assigned without a referencing identifier. Section [7.2](#) provides more information on how these annotations should be assigned. Thereafter,

6.1 DETAILS

In the context of *scales* (see more information in [6.7](#)), additional details may be provided to aid the interpretation of results/findings or specify in-/exclusion criteria of study objects. As these additional details can change the interpretation of results/findings, they need to be annotated in the context of these guidelines. However, we note here that only details for *scales* are annotated in a first instance. However, we kept this separate from the scale annotation, as it can be used to record other details in later stages.

Details are to be assigned with the same rules (boundaries, see section [5](#), and missing information, see section [7.2](#)) used for the identification of entities of interest. Once a detail has been marked, it needs to be linked to the entity of interest it belongs to (in a first instance either a *scale* minimum or maximum, or a categorical *scale*). Information on how to link to annotations with a relation is provided in section [2.1](#). If there is no entity this detail belongs to, then this annotation needs to be removed.

Each domain is covered by two questions, with scores ranging from 0 (no difficulty) to 4 (extreme difficulty or cannot do).

In the previous example, there are two different *details* provided “(no difficulty)” and “(extreme difficulty or cannot do)” that are used to explain the extremes of a numeric *scale*. Both *details* need to be annotated and a relationship between the scale extreme and the *detail* needs to be added. Figure [11](#) illustrates the annotation of the example sentence with Brat. More information on *measures* and *scales* is provided in section [6.7](#).

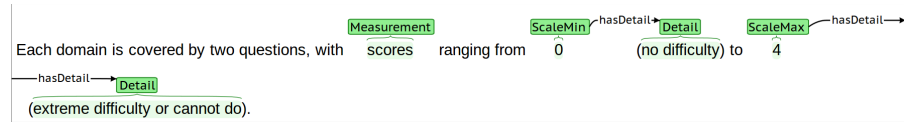


Figure 11: Illustrates how details to relevant entities should be annotated in accordance with these guidelines.

Further examples

TO BE ADDED LATER ...

6.2 AGE AND AGE GROUPS

As outlined in section 3.1, one entity type of interest are *ages* and *age groups*. While an *age* belongs to one individual, an *age group* usually refers to a group of individuals. Age groups can be given as an age range, e.g. “older than 65 years”, or with a textual description, e.g. “elderly”. Both textual description and age ranges are expected to be annotated.

We undertook cross-sectional surveys of residents aged older than 65 years (n=15 022) in 11 sites in seven countries with low and middle incomes (China, India, Cuba, Dominican Republic, Venezuela, Mexico, and Peru).

The text span “aged older than 65” refers to the age of the study participants. It is therefore to be annotated with at least one *AgeGroup* annotation. However, looking at table 1 shows that this overlaps with two age groups of interest: “aged adult (elderly)” and “adult, aged 80 and over”. In this particular case, two *AgeGroup* annotations have to be assigned, one for each age group that is covered in the table. Figure 12 exemplifies the use of *age group* annotations for the example sentence.



Figure 12: Illustrates how an age group should be annotated in accordance with these guidelines.

To speed up referencing age groups to UMLS concepts, all age groups defined as relevant in these guidelines have been provided as part of

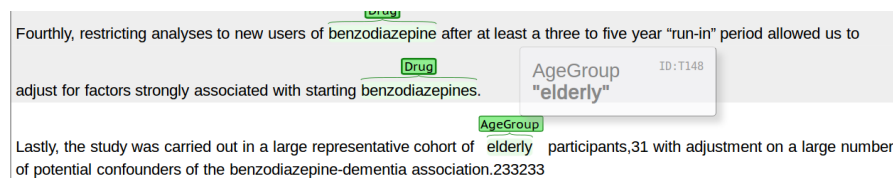
the annotation tool Brat (see 2.1 for more information on assigning identifiers); select “AgeGroups” in the drop down menu.

Age group annotations expect a strict correspondence between age range and the CUI assigned, by means that if a paper refers to the age group “14-78 year old”, all age groups covering this age range have to be assigned simultaneously. Note that an age range can also be smaller than those given in table 1, in which case it still needs to be referenced to the group the age range falls into.

6.2.1 Age groups as textual descriptions

In a number of cases, an *age group* is provided through a textual description, such as “elderly” or “middle-aged”. More often than not, the textual description is not further defined, in which case we cannot make any assumptions as to which age groups are covered by the textual description. In this case, an *age group* annotation needs to be assigned, but the identifier is to be left open. See the following example for illustration.

Lastly, the study was carried out in a large representative cohort of elderly participants, with adjustment on a large number of potential confounders of the benzodiazepine-dementia association.



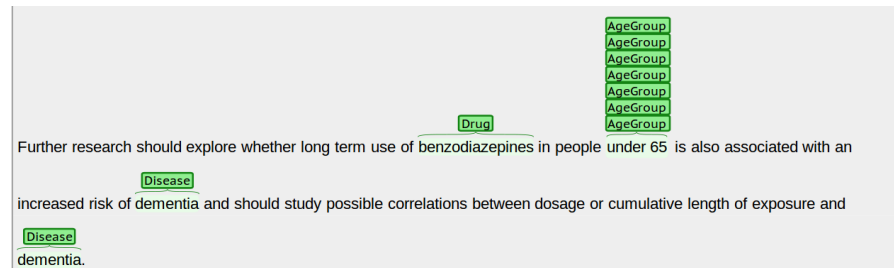
If both details are provided, the textual description and the age range, annotate both individually: the textual description with a missing identifier and the age range with identifiers of all the relevant *age groups* covered.

Further examples

Further research should explore whether long term use of benzodiazepines in people under 65 is also associated with an increased risk of dementia and should study possible correlations between dosage or cumulative length of exposure and dementia.

In this example sentence, the text span “under 65” refers to an *age group*. Similarly to previous examples, this text span would have to

be annotated with seven *age group* annotations to reference all *age groups* that cover the age ranges up to 65.



6.3 DISEASES

Another group of relevant entities are *diseases*. We are not only interested in recording mental illnesses, but also any other physical ailment that is reported in the scientific literature, e.g. “stroke” or “diabetes”. As described in section 3.2, *diseases* fall under the semantic types “Disease and syndromes (To47)” and “Congenital abnormality (To19)” in UMLS and should only be referenced if an identifier can be found belonging to these semantic types. If no concept is available falling under these semantic types, then an annotation still needs to be assigned but the identifier field should be left blank (see missing information section 7.2). To illustrate the use of *disease* annotations, consider the following example sentence.

Although the prevalence and incidence of most chronic diseases are strongly age dependent, only 23% of the disability burden caused by chronic disease in countries with low and middle incomes occurs in people aged 60 years and older, compared with 36% for high-income countries, where demographic ageing is much more advanced.

In this example, both text spans referring to “chronic disease(s)” need to be annotated with a *disease* annotation and the corresponding CUI retrieved from UMLS. Figure 14 shows how the annotations are expected to be assigned to the text.

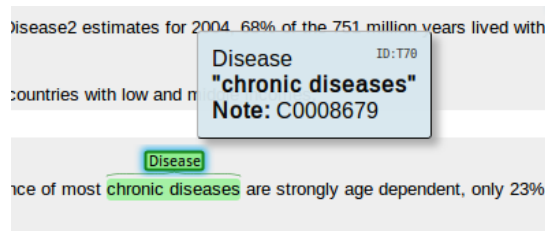


Figure 13: Illustrates how the text spans referring to “chronic disease(s)” should be annotated in accordance with these guidelines.

In some cases, diseases are reported as list and there are two different types of lists. List items are expected to be distinguished by the curator and annotated with multiple overlapping annotations that cover all the list items. To illustrate this further, consider the following example.

Secondly, we could not adjust separately for anxiety and sleep disorders, both putative dementia prodromes, owing to the lack of specific measurement of these symptoms in the PAQUID programme.

This example contains a list of diseases, namely “anxiety and sleep disorders”, which is expected to be annotated with two overlapping disease annotations, one for “anxiety disorders” and one for “sleep disorders”. Note that this sentence would also require a third *disease* annotation for “dementia”.

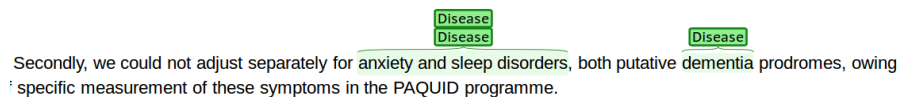


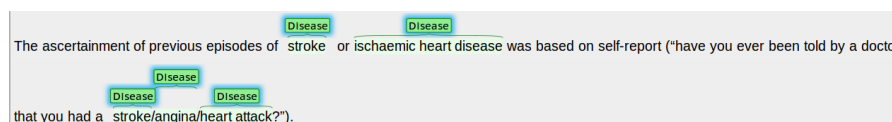
Figure 14: Illustrates how the text spans referring to “anxiety and sleep disorders” should be annotated with multiple *disease* annotations.

We note here that we slightly diverge from previous disease annotation schemes in that we allow abbreviations to be included in the annotation when they are first introduced. This means that a text span such as “Major Depression Disorder (MDD)” would yield only one annotation for the entire text span instead of two, one for “Major Depression Disorder” and one for “MDD”.

Further examples

The ascertainment of previous episodes of stroke or ischaemic heart disease was based on self-report (“have you ever been told by a doctor that you had a stroke/angina/heart attack?”).

In this example, “stroke”, “ischaemic heart disease”, “agina” and “heart attack” would all be single *disease* annotations. Note that some of the diseases occurring could constitute a brief medical event while others may be long-lasting. However, the duration of the disease does not have any influence on the annotations by means all disease need to be annotated, no matter how short or long their duration.



The diagram illustrates the annotation of the sentence: "The ascertainment of previous episodes of stroke or ischaemic heart disease was based on self-report (“have you ever been told by a doctor that you had a stroke/angina/heart attack?”)." Green boxes labeled "Disease" are placed above the words "stroke" and "ischaemic heart disease" in the first part of the sentence, and above "stroke/angina/heart attack" in the second part.

6.4 DRUGS

As highlighted earlier in chapter 3, *drugs* are another group of entities of interest. Drugs are chemically synthesised compounds with active ingredients that aim to treat the symptoms of illnesses. In the frame of these guidelines, *drugs* are distinguished from *treatments*, which are organic active compounds such as herbs and plant extracts and other treatments such as psychotherapy or mindfulness. Drugs should be mainly annotated with identifiers obtained from DrugBank¹, or in exceptional cases, from UMLS if it contains an entry for this drug or group of drugs.

We considered all benzodiazepines and similar drugs available in France between 1988 and 2006 (alprazolam, bromazepam, chlordiazepoxide, clobazam, clonazepam, clorazepate, clotiazepam, diazepam, estazolam, flunitrazepam, loflazepate, lorazepam, lormetazepam, nitrazepam, nordazepam, prazepam, oxazepam, temazepam, tetrazepam, tofizopam, triazolam, zolpidem, and zopiclone).

In this example, all the drugs named in parentheses should be annotated with a *drug* annotation holding the corresponding DrugBank identifier. For example, the annotation for “alprazolam” should hold the identifier “DB00404” as identified by searching the web interface

¹ <http://www.drugbank.ca/>

of the database. Furthermore, the example sentence also references a group of drugs, “benzodiazepines”. This drug group should also be annotated, though not with a DrugBank identifier. As DrugBank does not contain the term “benzodiazepines”, a [UMLS](#) lookup is required. [UMLS](#) provides the identifier “C0005064” for this specific drug group. Figure 15 illustrates this example sentence. Note that only in the case of the identifier not being present in DrugBank, should [UMLS](#) be used. Relevant semantic types are *Organic Chemical* (*T109*) and *Pharmacologic Substance* (*T121*). Further note that “loprazolam” (C0077013), “nordazepam” (C0011279), and “tetrazepam” (C0076341) are the exceptions in that they cannot be found in DrugBank and need to be annotated with an [UMLS](#) identifier.

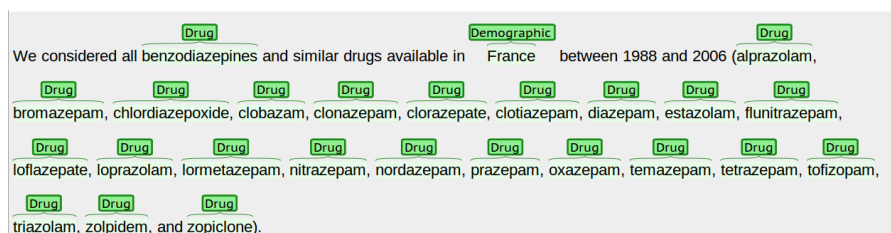
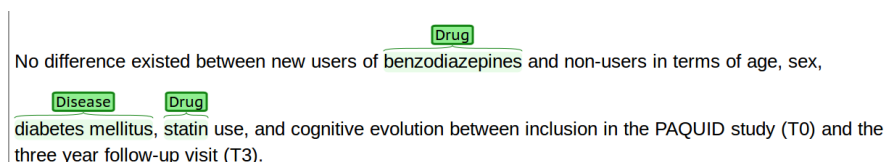


Figure 15: Illustrates how the all text spans referring to specific drug names should be annotated in accordance with these guidelines.

Further examples

No difference existed between new users of benzodiazepines and non-users in terms of age, sex, diabetes mellitus, statin use, and cognitive evolution between inclusion in the PAQUID study (T_0) and the three year follow-up visit (T_3)

In this sentence, the text span “statin” refers to a group of drugs. Similar to the previous example the drug group cannot be found in DrugBank and searching for it only delivers a long list of drugs falling into this group. So again, [UMLS](#) needs to be queried whether an entry exists for this drug group. [UMLS](#) uses the identifier “C0360714”, which should be used for annotating this text span.



A regimen was defined as all antidepressant and antipsychotic/antimanic drugs (ADAP) that were valid concurrently during an MDD episode.

In this sentence, the text span “antidepressant and antipsychotic/antimanic drugs (ADAP)” needs to be annotated with one *drug* annotation referencing three different identifiers, as illustrated in figure 16. While this example cover drug groups, rather than individual drug names/brands, the same rules for lists need to be applied to drug lists.

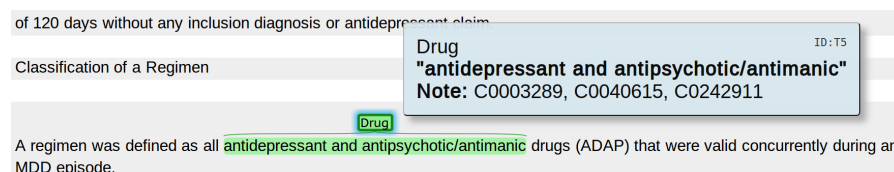
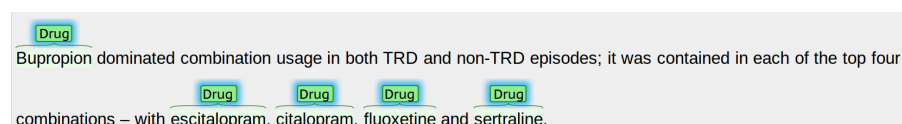


Figure 16: Illustrates how the text span referring to a list of drug groups should be annotated. Also highlights how a different resource (here UMLS) is used for a drug name.

Bupropion, sertraline, escitalopram, fluoxetine and citalopram were the most commonly used drugs and were used almost equally in both TRD and non-TRD episodes when used as monotherapy.

This example illustrates a list of drugs used for depression. All of the drugs possess entries in DrugBank and are therefore to be referenced with DrugBank identifiers.



Compared with non-users (n=968), new users of benzodiazepines (n=95) were more likely to have shorter school duration (66% v 77% with duration ≥ 7 years), to be single or widowed (52% v 41%), to have more significant depressive symptoms (16% v 4%), to use antihypertensive drugs (74% v 58%), and to use platelet inhibitors or oral anticoagulants (15% v 6%) and consumed wine less regularly (63% v 73%).

In this sentence, four drug groups have been mentioned: “benzodiazepines”, “antihypertensive drugs”, “platelet inhibitors” and “oral

anticoagulants”. Each of these drug groups needs to be annotated individually with the corresponding identifiers from [UMLS](#).

Compared with non-users (n=968), new users of benzodiazepines (n=95) were more likely to have shorter school duration (66% v 77% with duration ≥ 7 years), to be single or widowed (52% v 41%), to have more significant depressive symptoms (16% v 4%), to use antihypertensive drugs (74% v 58%), and to use platelet inhibitors or oral anticoagulants (15% v 6%) and consumed wine less regularly (63% v 73%).

6.5 ETHNICITY AND DEMOGRAPHICS

Another two entity types of interest are *ethnicity* and *demographics*, which are reported here together as they are closely related. Nonetheless, both require separate annotations within scientific papers and the difference between both is illustrated here.

Ethnicities

Ethnicities refer to the cultural, societal or ancestral roots of individuals. Mostly this information is provided by naming the country a study object is from. This means that a sentence referring to the origin of one or multiple study objects with a country, an *ethnicity* annotation is to be assigned. Searching [UMLS](#) requires the transformation of country names (e.g. “India”) into ethnicities (e.g. “Indian”) to be provided with an identifier. All ethnicity annotations in [UMLS](#) fall under the semantic type *Population Group* (*T098*). The following example illustrates the use of country names as ethnicities.

We undertook cross-sectional surveys of residents aged older than 65 years (n=15022) in 11 sites in seven countries with low and middle incomes (China, India, Cuba, Dominican Republic, Venezuela, Mexico, and Peru).

This sentence provides country names in parentheses to indicate the origins of study objects recruited for the represented investigation (“China”, “India”, “Cuba”, “Dominican Republic”, “Venezuela”, “Mexico”, and “Peru”). Note that this is a special case and ethnicities provided as such in the text should be also annotated, not only country names referring to the origins of study objects.

Demographics

While in some cases a country name refers to the ethnicity of study objects, in others the intended use is in fact as a country name. In those cases where the country name is used to refer to the country instead of people belonging to the country, a *demographics* annotation

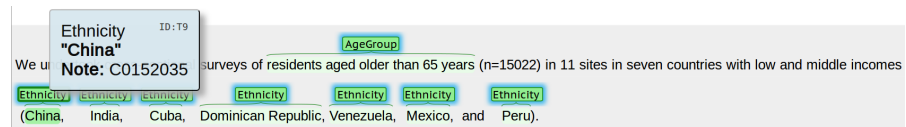


Figure 17: Illustrates how the text span “China” should be annotated with an *ethnicity* annotation in accordance with these guidelines (similarly, “India”, “Dominican Republic”, “Venezuela”, “Mexico”, and “Peru”).

should be assigned. *Demographic* data in general falls under the UMLS semantic type “Geographic Area (To83)”.

Mean age of the participants varied between 71.3 and 75.2 years, and was higher in Latin America and urban China than in rural China and India.

In this example sentence, the perspective changes. While it still tangentially refers to the origins of participants, the main aim is to communicate a finding, i.e. the variation of the mean age of participants for the different countries. In this example, all country names should obtain a *demographics* annotation containing the corresponding UMLS identifier (here “Latin America”, “China”, “China” and “India”).

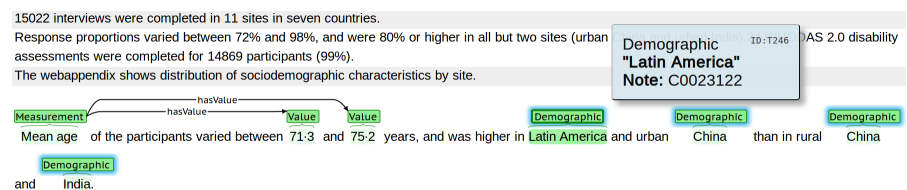


Figure 18: Illustrates how the text spans “Latin America” should be annotated with a *demographics* annotation in accordance with these guidelines.

Further examples

Two of them were done in Taiwan using health insurance data of people aged 45 years and older and found an increased risk of dementia in chronic users (>6 months) (adjusted odds ratio 1.24, 1.01 to 1.53)¹⁸ and current users (adjusted odds ratio 2.71, 2.46 to 2.99).¹

This example refers in the context of the paper to background information and related studies that have been executed elsewhere. The way the sentence is phrased (also taking the context of the paper

into consideration), does not provide additional information as to whether the study was executed in Taiwan and the study objects were recruited there or not. Given that in this case, it is unclear whether “Taiwan” also refers to the origins of the study participants, a *demographics* annotation needs to be assigned, referencing a [UMLS](#) identifier with the semantic type “Geographic Area (To83)”.

Two of them were done in **Taiwan** using health insurance data of people aged 45 years and older and found an increased risk of **dementia** in chronic users (>6 months) (adjusted odds ratio 1.24, 1.01 to 1.53)¹⁸ and current users (adjusted odds ratio 2.71, 2.46 to 2.99).¹⁹ A nested case-control study among

6.6 GENDER

As described in section 3.5, gender-specific information encountered while curating a paper needs to be annotated with a *gender* annotation. Table 2 introduces all genders that should be annotated when found in scientific text. Similar to *ethnicity* information, *gender* information is more likely to occur in method sections or in the summary of results, but may also be present in any other part of the paper and needs to be annotated in every instance. The following sentence illustrates how gender-specific could be used in a paper.

Depressive symptoms were assessed by using the Center for Epidemiologic Studies depression scale,³⁷ categorised according to the threshold validated for the French population with a cut-off score of ≥ 17 for men and ≥ 23 for women,³⁸ defining the presence of significant depressive symptoms.

This example provides information about gender-specific thresholds for the “Center for Epidemiologic Studies depression scale” measure employed in the study. Both the occurrence of “women” and “men” would have to be annotated with a *gender* annotation.

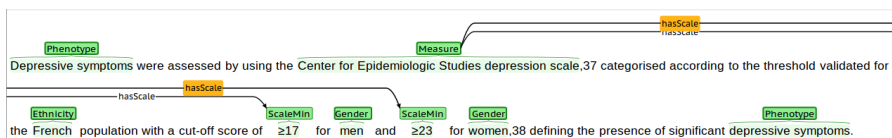


Figure 19: Illustrates how the text spans “women” and “men” should be annotated in accordance with these guidelines.

Further examples

MDD is more common among women than men and often begins in young adulthood [3].

As illustrated with the previous example, in this sentence also both words “women” and “men” would need to be annotated with a *gender* annotation.

Gender Gender AgeGroup

MDD is more common among women than men and often begins in young adulthood [3].

6.7 MEASURES AND SCALES

Section 3.6 details several existing and commonly applied measures with relevance to dementia. However, the measures provided in this table are only aimed at speeding up curation as they are the most common scores. Thus, this list is incomplete and other scoring methods may have been used in the literature that may or may not be referenced in UMLS.

Measures

“For cognitive decline, three different tests were considered: the mini-mental state examination,³⁴ the Benton visual retention test,³⁵ and the Isaacs set test.”

This sentence references three different measures that all need to be annotated with a *measure* annotation. For the first two (“mini-mental state examination” and “Benton visual retention test”), an identifier can be retrieved from UMLS. For the third measure, no UMLS entry can be found so that an annotation with a missing identifier needs to be assigned (see 7.2). Note that when searching for “Benton visual” in UMLS, four different items are retrieved out of which “Psychologic test, Benton visual retention test (C0204461)” needs to be chosen, because this term falls under the semantic type *Diagnostic procedure (To60)* and is indicated to be relevant for the mental illness domain. The annotations for this sentence are illustrated in figure 20.

Scales

Scales provide additional information for *measures* in that they aid the interpretation of a *measurement*. *Scales* can be either continuous

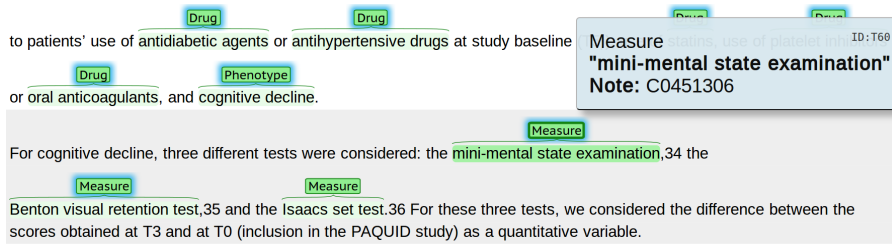


Figure 20: Highlights the annotations of *measures* in accordance with these guidelines.

numerical scales or categorical scales. In the frame of this study, we aim to report both types of scales, continuous and categorical. Continuous scales are special in that they are reported mostly with their minimum and maximum value and the interpretation for this value. See the following example for a continuous scale.

Each domain is covered by two questions, with scores ranging from 0 (no difficulty) to 4 (extreme difficulty or cannot do).

In this example, a continuous *scale* for the measured “score” is provided with a minimal value of “0” and a maximum value of “4”. An additional two annotations are required for the *details* which need to be highlighted with *detail* annotations (see section 6.1).

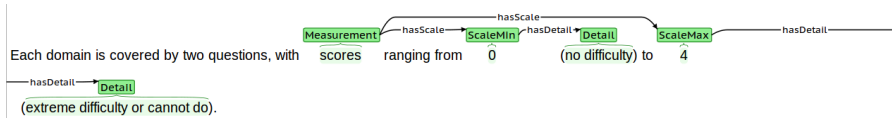


Figure 21: Illustrates the annotations of a continuous *scale* in accordance with these guidelines.

The following example highlights the use of *categorical scales* that need to be annotated within a publication.

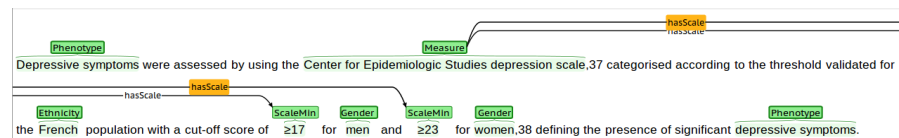
TO BE ADDED LATER: principle same: assign measure annotation, assign categorical scale for the text span explaining entire scale and then connect both via “hasCategoricalScale” relationship.

Further examples

Depressive symptoms were assessed by using the Center for Epidemiologic Studies depression scale, categorised according to the threshold validated for the French population with a cut-off

score of ≥ 17 for men and ≥ 23 for women, defining the presence of significant depressive symptoms.

This example shows the application of a measure, “Center for Epidemiologic Studies depression scale”, to determine mental illness (here “depression”). This sentence further defines cut-off values, “ ≥ 17 for men” and “ ≥ 23 for women”, which means that scores for individuals must exceed these cut-off values depending on their gender. Thus, the cut-off values are considered to be minima for the applied measure and have to be annotated as such and connected to the measure with a relationship.



Information about age, sex, marital status, educational attainment (defined as: none; some, not completed primary; completed primary; and completed secondary or tertiary) and living circumstances were assessed with a standard sociodemographic questionnaire.

In this sentence, no categorical *scale* annotation needs to be assigned as the scale provided is not pertaining to a measure related to mental illness. In fact, this example sentence does not require any annotations.

6.8 PHENOTYPES (SIGNS AND SYMPTOMS)

Phenotypes are observable characteristics of an organism, mostly used to describe abnormal states of an individual, e.g. in the context of disease. *Phenotypes* imply on observation/measurement by means that a certain characteristic is further described with a value (comparable to attribute value representations). For example, in the sentence “Molly has brown hair”, “brown hair” would need to be annotated as *phenotype*. However, if a phrase implies no value for a characteristic, the phrase refers to a trait and should not be annotated. For example, in the sentence “We recorded the gender of study participants”, “gender” should not be annotated as it constitutes a trait and not a phenotype in the context of these guidelines. In the context of these guidelines, phenotypes are to be distinguished from diseases (see sections 6.3 for further details) and traits (which are not to be annotated). The following sentence exemplifies the usage of phenotype expressions in the scientific literature.

However, heterogeneity was severe ($I^2 > 56\%$) for dementia, paralysis or weakness of a limb, depression, skin disorders, eyesight problems, and myocardial infarction or angina.

This example sentence contains a list of several *diseases* and *phenotypes* and every one instance would have to be annotated. The phrases “paralysis [of a limb]”, “weakness of a limb” and “eyesight problems” are considered to be *phenotypes* in the preceding example sentence.

However, heterogeneity was severe ($I^2 > 56\%$) for dementia, paralysis or weakness of a limb, depression, skin disorders, eyesight problems, and myocardial infarction or angina.

Annotations:
 - **Disease**: dementia, depression, skin disorders, myocardial infarction or angina.
 - **Phenotype**: paralysis or weakness of a limb, eyesight problems.

Further examples

Physical impairments were assessed on the basis of self-reported paralysis, weakness, or loss of a limb; eyesight problems; stomach or intestine problems; arthritis or rheumatism; heart problems; hearing difficulties or deafness; breathlessness; difficulty in breathing or asthma; fainting or blackouts; skin disorders, such as pressure sores, leg ulcers, or severe burns; or persistent cough.

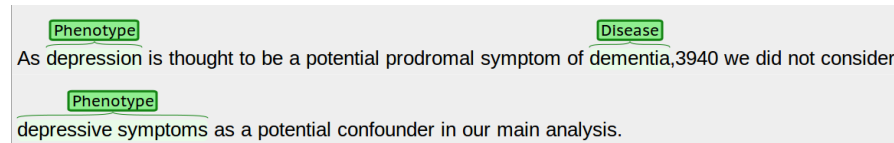
In the previous sentence, several phenotypes (separated by “;”) are listed whereby some of the list items itself could be considered to be listings, e.g. “arthritis or rheumatism”. In those cases where list items constitute listings, several annotations per list item are necessary (one for “arthritis” and one for “rheumatism” in the given example). Note, that some list items may be summarised in a similar fashion in [UMLS](#) in which case the corresponding list items should be summarised in accordance to [UMLS](#). For example, “fainting or blackouts” exists as one concept in [UMLS](#) and should be annotated as such (with ID “C3842979”).

Physical impairments were assessed on the basis of self-reported paralysis, weakness, or loss of a limb; eyesight problems; stomach or intestine problems; arthritis or rheumatism; heart problems; hearing difficulties or deafness; breathlessness; difficulty in breathing or asthma; fainting or blackouts; skin disorders, such as pressure sores, leg ulcers, or severe burns; or persistent cough.

Annotations:
 - **Phenotype**: paralysis, weakness, or loss of a limb; eyesight problems; stomach or intestine problems; heart problems; difficulty in breathing or asthma; fainting or blackouts; or persistent cough.
 - **Disease**: arthritis or rheumatism; hearing difficulties or deafness; breathlessness; skin disorders, such as pressure sores, leg ulcers, or severe burns.

As depression is thought to be a potential prodromal symptom of dementia,³⁹⁴⁰ we did not consider depressive symptoms as a potential confounder in our main analysis.

While “depression” can refer to a *disease*, in this case the word refers to a “prodromal symptom of dementia” and is therefore to be annotated as a *phenotype*. Note that in this example the additional information in the sentence was sufficient to distinguish between a *phenotype* and *disease* annotation, but in other cases (such as the previous example) the entire content of the article needs to be taken into consideration.



As **Phenotype** depression is thought to be a potential prodromal symptom of **Disease** dementia,³⁹⁴⁰ we did not consider **Phenotype** depressive symptoms as a potential confounder in our main analysis.

6.9 TREATMENTS

Any method or substance that is used to alleviate the symptoms experienced by a patient are considered to be relevant for the purpose of annotation. However, a distinction needs to be made between *drugs* and *treatments*: anything that is used for symptom improvement which is registered as an official drug and can be prescribed to patients requires a *drug* annotation; anything else used to alleviate symptoms requires a *treatment* annotation.

Mindfulness is defined as non-judgmental awareness of the present moment [13] and was introduced by Kabat-Zinn [14] as a treatment option for patients with chronic somatic conditions.

In this particular example, “[m]indfulness” is used as an treatment in the context of the paper. Thus, “mindfulness” would have to be assigned a *treatment* annotation (as opposed to a *drug* annotation) with the [UMLS](#) identifier “C3542996”. Figure ?? illustrates how this example should be annotated.

Interestingly, this definition does not take response to psychological treatment into account, while cognitive behavior therapy (CBT) or interpersonal therapy (IPT) are part of international treatment guidelines for recurrent and chronic depression [9].

This example sentence contains two *treatments* to be annotated: “cognitive behaviour therapy (CBT)” and “interpersonal therapy (IPT)”. As both are contained in [UMLS](#), they need to be characterised with [UMLS](#) identifiers (“C0009244” and “C0204548 ” respectively). Note that for both *treatments* abbreviations are introduced. When annotating, include the abbreviation in the *treatment* annotation and also assign *treatment* annotations to any subsequent use of this abbreviation.

Further examples

Segal, Williams and Teasdale [15] adapted the program for patients with recurrent depression, naming it ‘Mindfulness Based Cognitive Therapy’ (MBCT).

In addition to containing a *treatment* (“Mindfulness Based Cognitive Therapy (MBCT)”), this sentence also contains a *disease* (“recurrent depression”, [CUI](#) “C0221480”). Both need to be annotated and the following figure provides an illustration of the required annotations. Note that there is no suitable entry for “Mindfulness Based Cognitive Therapy (MBCT)” in [UMLS](#) and an identifier for this annotation cannot be provided (further information see section [7.2](#)).

An additional analysis revealed that patients had an average of 16 psychotherapy sessions in each TRD episode as compared to only 4 sessions in a non-TRD episode.

In this particular example, only one *treatment* annotation is required for “psychotherapy” ([CUI](#) “C0033968”). While the length of the treatment is not relevant, the *disease* referred to by “TRD” (treatment resistant depression, [CUI](#) “C2063866”) needs further annotating.

RESOLVING CONFLICTS DURING THE ANNOTATION PROCESS

In some cases, there may be conflicts arising as to whether and what annotations to assign to a particular text span. This chapter aims to offer some conflict resolutions for the most common problem encountered while annotating entities relevant to mental illnesses.

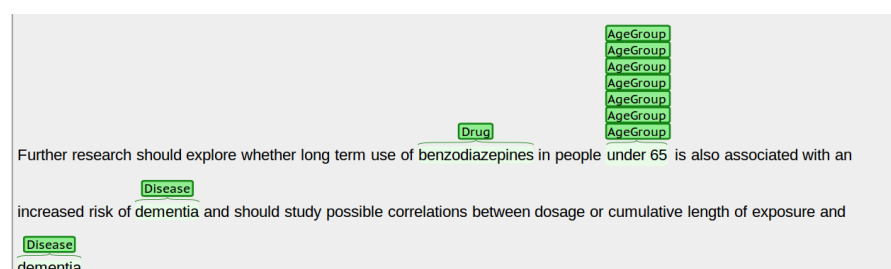
7.1 OVERLAP OF CONCEPTS

While in some cases it is required to assign multiple, overlapping annotations to one text span, overlaps should be avoided where possible. In no circumstance should overlapping annotations be used to indicate uncertainty about the type/identifier of an entity. In case of doubt, decide for the one that makes most sense to you and assign this annotation. In case of truly overlapping annotations, assign multiple annotations to a text span, e.g. for *AgeGroups* when an age range covers multiple groups referenced in these guidelines.

7.1.1 Examples

Further research should explore whether long term use of benzodiazepines in people under 65 is also associated with an increased risk of dementia and should study possible correlations between dosage or cumulative length of exposure and dementia.

In this example sentence, the text span “under 65” refers to an *age group*. Similarly to previous examples, this text span would have to be annotated with seven *age group* annotations to reference all *age groups* that cover the age ranges up to 65.



Further research should explore whether long term use of benzodiazepines in people under 65 is also associated with an increased risk of dementia and should study possible correlations between dosage or cumulative length of exposure and dementia.

The image shows the sentence with several green boxes containing labels above specific text spans. Above "benzodiazepines" is a box labeled "Drug". Above "under 65" is a box labeled "AgeGroup". Above "increased risk of dementia" is a box labeled "Disease". Above "dementia." is a box labeled "Disease". On the right side of the image, there is a vertical stack of seven "AgeGroup" labels, indicating that multiple overlapping annotations are used for the "under 65" span.

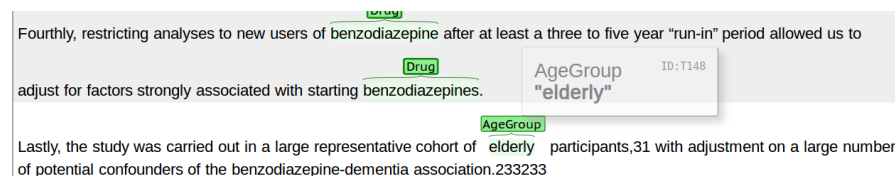
7.2 MISSING IDENTIFIERS IN REFERENCE RESOURCES

In some cases, the reference resources as outline above do not contain an entity that is referenced in the text and relevant to mental illnesses. It is crucial that these text spans are still annotated even though they cannot be referenced to a resource identifier. In the case of missing identifiers, the procedure is similar to assigning an annotation with identifier: the text span needs to be highlighted to activate the annotation dialogue from which the entity type can be chosen. However, the note field that in our use case holds the corresponding resource identifier is expected to be left blank. To illustrate this further, please see the following example.

7.2.1 Examples

Lastly, the study was carried out in a large representative cohort of elderly participants, with adjustment on a large number of potential confounders of the benzodiazepine-dementia association.

In this example, the text span “elderly” refers to an *Age group*. However, there is no detail provided what the age range is, the authors see covered by “elderly”. Therefore, an *Age group* annotation without identifier should be assigned to this text span.



The annotations without referenced identifier can be distinguished in the annotation process by the way the annotations are represented in the web interface. Annotations that are referenced to a resource contain a blue surrounding around the annotation type whereas this blue surrounding is absent for unreferenced annotations. Figure 22 shows the difference between a referenced (“benzodiazepines”) and an unreferenced annotation (“elderly”).

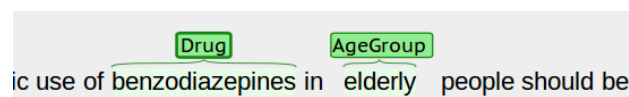


Figure 22: Illustration to help distinguishing annotations that are referenced (“benzodiazepines”) and un-referenced (“elderly”) while annotating with Brat. Referenced annotations contain a thicker outline around the annotation type that is absent in un-referenced annotations.

