Guidelines for annotation:

Guidelines for annotation of entities in clinical text. The guidelines are divided into two main parts: guidelines for of Disorders, Findings and Body structures and guidelines for Pharmaceutical drugs.

The intended annotators are primarily domain experts, i.e. clinicians with experience in reading, writing and understanding electronic patient records

These guidelines have been applied on a subset from the Stockholm EPR corpus (Dalianis et al. 2009), consisting of Assessment entries from an emergency ward.

Entity annotation of Disorders, Findings and Body structures in Swedish health records

This part provides guidelines on annotation of clinical findings in the free text section of Swedish electronic patient records. Clinical findings include the classes **Body structure, Finding** and **Disorder**.

Selection of annotation classes

These guidelines describe annotation rules for four classes of clinical entities, inspired from the SNOMED semantic categories Body Structure, Finding and Disorder.

Words and expressions are to be marked as belonging to one of four classes.

- *Body structure*
- *Finding*
- *Disorder*

These classes were chosen for various reasons. The main purpose of this annotation is to identify words and expressions that are significant for the patients' medical history. We wanted to identify expressions for symptoms, medical observations made by the patient or the physician, and the disorders that are discussed in the health record. These are all included in the SNOMED class of Findings. In Snomed, disorder is a subclass included in Findings but is here separated from the wider concept of findings. Often, the diagnosis is the key word in a medical record and holds a higher dignity of meaning. Also, we wanted to include the class Body structure as the part of the body which is ill, damaged or examined is of certain importance.

Principles

Here, the general principles for the annotation are listed. A list of rules that are specific for each annotation class is found below. At the end of the document, there is also a list of examples that was established after the training task.

Marking

Only sequential words are to be annotated in the same marking, no leaps. No nestled markings are allowed, neither any markings of a part of word. Therefore, compound words are to be annotated as a whole word, not as part of a word.

The shortest possible expression is to be marked, but it must still fully describe the finding, disorder or body structure. Modifiers are therefore to be excluded. Example: ”The patient experiences a strong stabbing pain in the knee” *(“Patienten känner kraftigt stickande smärta i knät”)*. The word “strong” is a modifier, and is therefore not to be annotated, “stabbing pain” is a finding, and “knee” is a body part.

Words separated by hyphen are considered as one word, whereas words separated by slash are considered as two separate words. This is in accordance with the annotation tool. Therefore asthma-allergy will be annotated as one word, whereas asthma/allergy will be annotated as two words. (When hyphen is used together with a conjunction, as in the example ”asthma- and allergyproblems”, the words are treated as two separate words, and are annotated as one annotation or two separate, depending on situation.)

Choice of annotation class for identified entity

The primary rule for annotation is to annotate words into the class as they are perceived in clinical reality. The same word/expression can be a finding in one instance and a disorder in another and is to be annotated in accordance with context. So if “angina” in one context is a disorder (angina pectoris) and in another context a finding (patient experiencing chest pain), then it is to be annotated as a disorder in the first case and as a finding in the second. In case of doubt, disorder has priority over finding.

Findings such as “signs of infection” (*tecken till infection*), “no support for diabetes” (*inga hållpunkter för diabetes*) contains words of disorder. To avoid nestled annotations, the annotation class disorder has priority. In these examples, “infection” and “diabetes” are thus to be annotated as disorders.

If it is possible to annotate a finding and a body part separately, as in the example ”pain in knee” (*smärta i knä*), it is to be annotated separately. However, when it is not possible, as in ”CAT scan of head without remarks” (*CT skalle är ua*.), the entire expression is to be annotated as a finding. (”CAT scan” is a procedure, “head” a body structure and ”without remarks” in itself does not make sense, whereas the finding ”pain” makes sense.) Finding in this example has therefore priory over body structure.

Which entities to annotate

Any mention of the four selected classes shall be annotated. It is irrelevant whether, for example, a mention of a disorder is referred to as something that a patient has or had, a patient does not have, a patient might have or might get in the future. It is also irrelevant if it is referred to that the patient or someone else than the patient has the disorder. All these instances shall be annotated as a disorder. The same holds true for all of the four classes, if a part of the body is mentioned it shall be annotated, regardless of whom the body part belongs to. An exception to this rule is when expressions are used as a figure of speech (often with body structures). See examples at the end of the document.

Indirect entities shall not be annotated

Only things that are explicitly mentioned shall be annotated. For example, if it is mentioned that the patient takes insulin, it can be concluded that the patient has diabetes. However, such conclusions require medical knowledge. Also, if it is stated that a patient visits his nurse at the Diabetes center

(compound word in Swedish; *diabetesklinik)*, it is implicated that he has the diagnosis diabetes, but the implication is of such a level that the word diabetesclinic shall not be annotated as a disorder.

Very fuzzy expressions are not be annotated. Examples: ”It is probably a matter concerning psychology more than cardiology" (*Det rör sig troligen mer om något psykologiskt än om något kardiellt*). ”Pain in abdomen possibly of gynecological origin” (*Magsmärtor möjligen av gynekologiskt ursprung*).

Compound words

The Swedish language produces compound words in an unpredictable, creative way. If a constructed compound word still is a finding, disorder or body structure, the compound shall be annotated as such. However, if the compounding changes its class, special care must be taken when annotated.

- In the sentence: "The patient receives treatment for his diabetes", it is clear that it is the disorder diabetes that is discussed. The word “diabetes” is therefore annotated as a disorder. If the writer instead chooses to use a compound word as in “The patient receives diabetic-treatment” (*diabetesbehandling*) it shall not be marked as a disorder as this is a procedure or treatment.
- Likewise, in the sentence "The patient receives diabetic-diet (*diabeteskost*)". In this case, the word is no longer a disorder, but the compounding has transformed it into a kind of diet. The same is true with the word “diabetes-ward” (*diabetesavdelning*), which is a location.
- The compound word "diabetesdisease" (*diabetessjukdom*), on the other hand, is still a disorder. Likewise, the word "diabeticsymptom" (*diabetessymptom*) is a finding. However, in the expression "symptom of diabetes", the word “diabetes” is a disorder, and since a disorder is present, it should have priority when annotating.

Expressions containing combined findings or disorders

Some expressions contain more than one finding (or disorder), but are conceptually to be seen as one item. For example “Stable in heart rate and blood pressure” (S*tabil i puls och blodtryck*) is an expression of the state of circulation for the reading physician. Another example of such an expression is “ECG and troponin without comment“ (*EKG och troponin u a*). These are to be annotated as one joint finding.

There are also some expressions that contain more than one finding that are conceptually not one item, but instead findings that are unrelated to each other. For example in a sentence with the structure: ”Symptom 1, symptom 2, symptom 3 without comment.”. These are not possible to annotate separately without creating nested or split annotations. They are therefore to be annotated as one finding.

The finding “ECG normal” (*EKG normalt*) is in the following sentence split up with a lot of unrelated words. “ECG on the other hand appears in large parts normal, but…” (*EKG ser å andra sidan till stora delar normalt ut, men…*). Such instances are to be omitted since there are no practical possibilities to annotate them, without also capturing the unrelated words. (Split annotations are not used.)

Disorders and findings according to SNOMED CT

As mentioned above, the basic principle is to annotate words into the class as they are perceived in clinical reality, and therefore the same word/expression can be a finding in one instance and a disorder in another. The basis for distinguishing between a finding and a disorder is taken from "SNOMED CT Style Guide: Clinical Findings" and is described in the following passage:

" Clinical findings may be simply defined as observations, judgments or assessments about patients. The problem with the terms "finding" and "observation" is that they seem to refer to the judgment of the observer rather than to the actual state of the body. "Organism state" has been suggested as a more neutral name, but it would need to be delimited from a "course of disease." Examples of clinical findings include: difficulty swallowing, nose bleed, diabetes, headache, and so forth. More precise and reproducible definitions of clinical findings, and the precise boundaries between findings and events, between findings and observables, between findings and situations, and the distinction between "finding" and "disorder", remain ongoing challenges at the margins. The distinction between a disorder and an observation has proven to be difficult to define in a reproducible manner across the tens of thousands of concepts included under clinical findings. Nevertheless, there are several reliable characteristics of each sub-category (disorders and findings):

1.1 Disorders

1) Disorders necessarily are abnormal.

2) They have temporal persistence, with the (at least theoretical) possibility of their manifestations being treated, in remission, or quiescent even though the disorder itself still present.

3) They necessarily have an underlying pathological process.

1.2 Findings

1) Findings may be normal (but not necessarily); no disorders may.

2) Some findings may exist only at a single point in time (e.g. a serum sodium level); no disorders may.

3) Findings cannot be temporally separate from the observing of them (you can't observe them and say they are absent, nor can you have the finding present when it is not capable of being observed).

4) They cannot be defined in terms of an underlying pathological process that is present even when the observation itself is not present.

Disorders may be present as a propensity for certain abnormal states to occur, even when treatment mitigates or resolves those abnormal states. In some cases the disease process is irrefutable, e.g. meningococcal meningitis. In others an underlying disease process is assumed based on the temporal and causal association of the disorder and its manifestation, e.g. nystagmus disorder is different from the finding/observation of nystagmus, which can be a normal physiological response to rotation of the head. If you spin around and around and then have nystagmus (the finding) you still do not have nystagmus disorder. And someone can have a nystagmus disorder without currently manifesting nystagmus. Similarly, deafness disorder is different from the symptom (observation) of reduced hearing, which can be due to a number of temporary causes such as excessive ear wax. "

Body structure

Relation to SNOMED CT

Corresponds to the semantic category Body Structure

Guidelines

The annotation class includes all body structures that can be anatomically defined. Sometimes a structure is named in a wider meaning, such as “pain in her muscles”, but as muscles are definite anatomical structures, the word “muscles” are to be annotated. Body fluids such as blood, sweat, or urine are not to be annotated, as they are not identifiable structures.

If a constructed compound word still is a body structure, it shall be annotated as such. If the compounding changes the meaning of the word so it no longer denotes a body structure, it is not to be annotated. Example: In the sentence “The patient is recommended a high leg-posture” (*Patienten rekommenderas högt benläge*), leg in leg-posture is no longer a body structure and shall not be annotated.

In the name of consequence, there will be “unfair” instances when the annotation is dependent on the writers’ choice of words for a certain concept. For example “x-ray of the lungs” can in Swedish be expressed either as a compound word “lungxray” (*lungröntgen*) or as separate words “x-ray of the lungs” (*röntgen av lungor*). Both are expressions for the same procedure. We have chosen to annotate “lungs” (*lungor*), and other similar words for body structures in this kind of expressions as the annotation class body structure, whereas the compound word “lungxray” (*lungröntgen*) is a procedure and shall not be annotated, nor are any half word-markings of “lung-“ to be used.

Exemple

- Chest pain can in Swedish be expressed either as a compound word “chestpain” (*bröstsmärta*) or as separate words “the patient has pain in her chest” (*patienten har ont i bröstet*). Both are expressions of the same symptom and should as such be annotated as findings. However, in the latter case the word “pain” (*ont*) is the finding and “chest” (*bröstet*) a body structure.
- The left ear can be referred to as a compound word “leftear” (*vänsteröra*), and shall be annotated as a body structure. If it instead is referred to in two words “left ear” (*vänster öra*), only the body part “ear” shall be annotated as “left” is a modifier.
- The compound word “earclinic” (öronklinik) shall not be marked as a body structure, as the meaning of the word is an institution or a place, a clinic.
- ”the vertebral column at chest level” (*kotpelaren i brösthöjd*); the vertebral column shall be annotated as body structure but not “chest”, as “chest level” (*brösthöjd*) indicates a position, not a body structure.
- Likewise, “subcutaneously” (*subcutant*) and “subcostaly” (*subkostalt*) are positions and not body structures.

- In the more diffuse expression “pain in the body” (*ont i kroppen*) the word “pain” (*ont*) is a finding and “body” (*kroppen*) shall be annotated as body structure as the body is an anatomically identifiable structure.
- Respiratory tract (andningsvägarna) =body part
- The veins (venerna)=body part
- “Thorax” is sometimes a body structure but can also, in this certain hospital, be the nickname of the clinic “Thoracic surgery” (*Thoraxkliniken*) or the building where the department of Thoracic surgery is located. Thorax shall only be annotated as a body structure when it refers to the body structure and not when it refers to the clinic.
- “svalg- och nasopharynxodling” (pharyngeal and nasopharyngeal culture) both are procedures and shall not be annotated, not even the word “throat” (svalg-) since it conceptually is a procedure svalgodling.
- blod (blood) is a body fluid and shall not be marked as a body structure.
- vad, (calf) can be a body structure, but also a word meaning ”what”. It shall not be annotated as a body structure when it is present in the text with the meaning “what”.

Examples when body structure is used in a figurative sense, and therefore shall not be annotated, are the three expressions “in one hand you have…” (i första hand…), “on the other hand…” or “secondly” (i andra hand…), and “He recieves the receipt in hand when leaving...” (Han får med sig receptet i handen när han går…).

Finding

Relation to SNOMED CT

Corresponds to the SNOMED CT semantic category finding

Guidelines

See also the definition of finding in the passage from the documentation of differences between findings and disorders.

Results that are interpreted, such as ”low blood pressure”, are to be annotated as a finding. If results are expressed in numbers and need to be interpreted e.g. “bloodpressure 100/40”, they are not be annotated, as these numbers require medical knowledge to be understood.

Only findings that have a medical relevance are to be annotated. There is for examples the SNOMED finding “writes Cantonese”, and these kinds of findings are not to be annotated unless they have a medical relevance.

Not only pathological findings are to be annotated. Also absence of pathological findings, such as “normal heart rhythm”, is to be marked. These non-pathological findings must however be the result of an observation, for example that the patient does no longer complain of a finding or that the patient is now well.

Mentions of procedures (colectomi) or devices (pacemaker, ileostomi) are not to be confused as findings and are not to be annotated. These are considered as procedures and devices and thereby potential separate classes.

Example

The finding ”vilovärk” ”resting-pain” is a clinical concept, and therefore in the expression ”värk i vila” (”pain while resting”), ”while resting” is not a modifier, but a part of the finding.

Disorder

Relation to Snomed CT

Corresponds to the Snomed category disorder:

Guidelines

See also the definition in the passage from the documentation of differences between findings and disorders.

General states such as cardiovascular disease (*kärlsjuk*) or lung disease (*lungsjuk*) are annotated as disorder as they are defined as the disorders in SNOMED. There are underlying pathological processes. However, expressions that are even more general, such as ”probably something gynecological” (*troligen något gynekologiskt*) or ”of cardiac origin” (*av kardiell genes*) are not be annotated as disorders as they are too diffuse and there may not be an underling pathological process.

Disorders expressed as an identity of a person as “a diabetic”, “an allergic (patient)” (diabetiker, allergiker) or as an attribute of a person are not to be annotated. For instance, the patient appears to be manic, is pregnant, is overweight (*patienten ter sig manisk, är gravid, är överviktig*).

Example

(malignancy) malignitet is to be annotated as a disorder, whereas “something gynaecological” (*något gynekologiskt*) is not.

Entity annotation of Pharmaceutical Drugs in Swedish health records

This part provides guidelines on annotation of the clinical entity class **Pharmaceutical drug** in the free text section of Swedish electronic patient records.

Relation to SNOMED CT

Includes the SNOMED CT categories:
410942007 Drug or medicament (substance*) [läkemedel]*
373873005 Pharmaceutical / biologic product (product) *[farmaceutiskt eller biologiskt medel]*

Guidelines

All mentions of a drug or a group of drugs shall be annotated with the class drug. It is irrelevant if it is named under its commercial name or its generic name, or as a class of drugs such as "antibiotics" (antibiotika). Also, fluids given intravenously shall be annotated as a drug since they are pharmaceutical products.

The Swedish compound word "sleepingpill" (sömntablett) shall be annotated as drug since it denotes a group of medications and the compounding does not change the meaning of it as a drug. "Sleepingaid" (sömnhjälpmedel) shall not be annotated as a Pharmaceutical drug, since aid can be given in many ways, not necessarily as a drug, but as relaxation programs, acupuncture and other methods.

Illegal drugs and other drugs not taken for medical purposes, such as cocaine or alcohol, shall not be annotated.

Compound words referring to a treatment with a drug should also be annotated as a drug.

Example

- Blood transfusion shall be annotated as a drug.
- Oxygen or other gases (e.g. anesthetic gases) shall be annotated as drugs.
- Lexinortreatment (Lexinorbehandling) shall be annotated as a drug.