Arizona Disease Corpus
Annotation Guidelines

Draft version 0.1.3, Sept 14, 2011

The goal of this project is to create an annotated corpus of disease mentions in sentences taken from PUBMED abstracts. This document provides a definition of disease for our purposes, the details of how to annotate a disease, and examples of each.

A disease:

- Is an abnormal state of a living organism.
  - May not be human. In particular, an animal model of disease is a disease.
  - Pregnancy is a normal condition and is therefore not a disease.
- Is observable in some way, though this may require specialized equipment or testing.
- Causes some sort of dysfunction.
  - Exception: A condition is a disease if it is currently in a latent state or in remission.
  - Examples: infections in the incubation period (e.g. HIV), cancer in remission, irritable bowel syndrome in remission
- May have direct causes (an infectious agent, toxicities) or contributory causes (genetic, environmental).
  - Diseases are typically not caused by another disease, conditions that are caused by another disease are typically symptoms. The only case of a disease clearly causing another disease is when one is a latent condition, such as HIV infection (a disease) causes AIDS (also a disease).
  - Is not primarily caused by an acute external physical force. Conditions of this type (including fractures, lacerations, trauma) are injuries.
  - The cause may be unknown.
- Scope: can affect any body part, organ, system or function. May be based in physical processes (e.g. genetic or anatomic) or mental processes (cognitive, behavioral, emotional).
- Typically will require and respond to some kind of treatment.
- Can be termed a *disease*, a *syndrome*, an *illness*, or a *disorder*.

Specific examples:

- Diseases:
  - Alzheimer's disease
  - Creutzfeldt–Jakob disease
  - Down syndrome
  - Breast cancer
  - Diabetes
  - Depression
  - Carpal tunnel syndrome
  - Obesity
  - Radiation poisoning
  - Malaria
  - Asthma
  - Schizophrenia

    - Hypochondriasis
    - Chronic fatigue syndrome
- Not diseases:
    - Pregnancy; it is a normal condition
    - Organ transplantation, smoking; are not a state of the organism
    - Nausea, hyperglycemia, anemia, adrenal suppression, pain and dementia; are all symptoms
    - Lacerations; are caused by external physical forces

Related entities:

- Symptoms: are caused by disease.
- Treatments: can be used to cure or relieve both diseases and symptoms.

Annotate mentions of disease entities:

- Each mention of a disease should be annotated exactly once. Each annotation should refer to exactly one mention of a disease. Disease entities should be annotated each time they are mentioned – a disease that is mentioned multiple times should be annotated multiple times.
- Mentions are a span of text where an entity is named. Therefore, the disease must be used in the text as a name. It does not have to be a proper noun.
- Mentions should be identifiable: annotator must be able to identify a specific disease using only the text of the mention itself.
    - Example:
        - Not a disease: "infections caused by the important pathogen Streptococcus pyogenes"
- Mentions should be specific: must refer to a single, specific disease, not a class of diseases or a set with a common trait
    - Include specific subtypes
    - Infections must indicate the infectious agent responsible.
    - Mentions of *cancer*, *tumor*, *neoplasm*, or *infection*, without additional information, are not sufficiently specific to annotate as a disease.
    - Two diseases are different entities when they have substantially different symptoms, treatments or causes
- Mentions are a single contiguous span of tokens. Tokens are delimited by both whitespace and punctuation. The annotation span:
    - Should only consist of entire tokens.
    - Should not start or end with whitespace.
    - Should not both start and end with parenthesis.
    - Should only include the disease mention, which is a noun phrase used as a name
- A mention could be an acronym.
    - A long form, short form pair should be annotated as two mentions. Example: "congenitally dislocated hips (CDHs)"
- The mention span should include terms such as *disease*, *syndrome*, *disorder* or *infection*, if present.
- Disease mentions may be embedded within a mention of a non-disease entity if the entity the outermost mention refers to has the disease.
    - Do not include terms from the outermost mention in the annotation span (*tissue*, *patient*) unless it is required by the tokenization.
    - Disease: Alzheimer's patient, asthmatic, breast cancer tissue, HIV-1 infected patient, diabetic

    - Not disease: cancer care center
- Mentions may be negated or hedged (indicated as not completely certain).
    - Do not include the negation or hedging in the annotation span unless required by the tokenization.
    - Example: probable chronic fatigue syndrome (only annotate “chronic fatigue syndrome”), nondiabetics (annotate the entire word)
- Lists and coordinations are phrases which mention multiple entities in a complex way. A simple illustrative example is “breast and ovarian cancer”, which refers to the diseases “breast cancer” and “ovarian cancer”. These constructs often overlap or do not explicitly mention some terms, and it is often not possible to specify a single, non-overlapping sequence of tokens for each concept. Therefore, the entire list or coordination should be annotated one time for each disease mentioned. Some lists and coordinations will mention both diseases and non-disease entities. Examples:
    - “Breast and ovarian cancer” would be annotated twice. One time the text “breast and ovarian cancer” would be designated the disease “breast cancer” and the second time the same text would be designated the disease “ovarian cancer”.