Article annotation scheme

August 8, 2011

1 Overview

This document describes the scheme for annotating the following information in medical research papers.

- Treatment groups: The names of the groups of people who are assigned a particular type of treatment. Group names usually include the name of the treatment assigned to the group (e.g. *quinine group* or *artemether group*). Groups will also have to be marked as *control* or *experiment*.
- Outcomes: The names of outcomes that are measured in the paper. Whether the outcome is *good* (something the treatment should improve such as recovering from an injury or disease) or *bad* (something the treatment should prevent or decrease such as developing an injury, disease or dying) also needs to be annotated.
- Times: These are the follow-up times when outcomes are measured for each treatment group.
- Group sizes: The number of people in a treatment group. Annotations for group sizes also include references to the treatment groups that they describe.
- Outcome numbers: The number of good or bad outcomes measured for a particular group at a given follow-up time. Annotations for outcome numbers include references to the treatment group they are recorded for, time when they are measured, and the name of the outcome.
- Lost to follow-up: The number of people who where originally assigned to a treatment group, but were not available at a particular follow-up time when outcomes were measured. Annotations for the number lost to followup include references to the name of the treatment group they were originally assigned to and the follow-up time when they were lost.

All of this information is needed to calculate the summary statistics *absolute risk reduction* (ARR), which is the percentage of control patients (those with the standard treatment) who would benefit from from taking the new treatment (the experimental treatment), and the *number needed to treat* (NNT) with the new treatment to prevent one bad outcome that would happen with the control. While these statistics sometimes appear papers, often they do not and physicians must calculate them. The annotated data will be used to train and test a machine learning system that can extract all of the necessary information to automatically calculate these statistics.

2 Annotating Abstracts

For now we are only concerned with annotating abstracts, not full papers. This decreases the amount of annotation effort involved and often papers that contain the numeric information that we want, report this

in the abstracts[1] . Furthermore, as all of the numeric data we want are integers, we are only annotating sentences that contain at least one integer. These sentences are also likely to contain treatment group, outcome, and time mentions as this information necessary to disambiguate the numbers the numbers that appear in the sentences. Sometimes a number is mentioned in word form (e.g. “one”, “sixteen”, etc) instead of numeric form. Treat all numbers in word form as if they were in numeric form.

The abstracts are in XML format and have three main sections.

- *annotated*: Sentences elements that should be annotated.
- *fulltext*: The original full text for the abstract. This section is for reference only and does not get annotated.
- *ignored*: Sentence elements that do not need to be annotated, but are included in case we decide to annotated them at some point in the future.

Annotations are XML tags that placed around the segments of a sentence corresponding to a piece of information that we are interested in. These annotations may be added in any text editor, XML editor, or in more sophisticated software packages such as GATE[2].

If you encounter an abstract that does not contain all of the types of information that we are annotating, it is okay. Simply annotate what is there. Times and the number lost to follow-up are not always explicitly mentioned. Some abstracts may not contain any of the information that we want.

2.1 Treatment groups

Groups are noun phrases that denote specific treatment groups in the study, including the control group. They are tagged with the <GROUP> tag, which has the attributes:

- id, which is unique to the particular group in the study.
- role, which is “control” or “experiment” if it is clear from the paper which treatment group is the control group and which has the experimental treatment. This attribute is omitted if the roles are not clear.

In some cases the name of a particular treatment group may seem rather long or the boundaries of the name may seem unclear. In this case try to identify both the minimal and maximal versions of the group name.

- The maximal treatment group is the full noun phrase denoting the treatment group - consider replacing it with the NP “Treatment X group” and seeing if (i) the sentence is still grammatical and meaningful, and (ii) if no other bits of the noun phrase could be deleted and maintain this quality (i.e., it is the maximal such NP). Treatment abbreviations inside parentheses are not tagged separately, but usually included in the full group name.
- The minimal version is the minimal noun phrase that denotes the treatment, uniquely distinguishing it from any other treatment condition in the abstract/paper. Preferably, it should be a base NP (noun

[1] In a random sample of 54 BMJ (British Medical Journal) articles, I found that it was possible to calculate summary statistics for 30 (56%) papers. Of these 30 papers, 13 contained all needed information in the abstract, 11 required the full text to be examined, and for 6 it was necessary to examine tables to find all of the necessary information.

[2] http:gate.ac.uk

preceded by possible determiner and adjectives/adverbs), though this may not always be the case. The words “group”, “arm”, and the like, are considered part of the short group.

The full group name should be annotated with the <GROUP> tag. Inside the full group name, annotate the shortest possible version of the group name with the <SHORT> tag, which has no attributes. A group can have more than one short version.

If a treatment group mention is small enough (as is often the case) that it does not make sense to distinguish between “long” and “short”, (e.g. “phenobarbital” or “didgeridoo playing”). In this case the a short version does not need to be annotated.

The following are some examples:

```
<group id="0" role="control">placebo group</group>

<group id="0" role="control"><short>placebo</short> group
   (<short>control</short>)
</group>

<group id="1" role="experiment">
       home based <short>medication review</short> by pharmacists
 </group>
```

In most sentences, the group names will be relatively short and therefore it often not necessary to identify the “short” version of the group name.

2.2 Outcomes

Outcomes are phrases that denote measured outcomes of the study. The outcome subjects (e.g. “in population X”, “in patients with condition Y”) are not normally included. However, post-modifying prepositional phrases may be included if they further define the outcome (e.g. “injuries to the knee or ankle”, “injuries of the knee”). Adjectives describing the degree of the outcome (e.g. mild, moderate, severe, etc.) are not usually included in the outcome. An outcome is tagged with the <OUTCOME> tag, which has these attributes:

- id, which is unique to the particular outcome in the study, as for Group above.
- type, which is “good” if the outcome is something that we want to increase or “bad” if it is something that we want to decrease. This attribute only applies to the outcome that is annotated. For instance in the clause “33 children in the treatment group did not develop malaria”, the outcome is “develop malaria” and should be considered “bad” even though the number reported is the number of “good” outcomes (i.e. not developing malaria).

As with the group names, longer outcome names may have a short version that can be annotated with the <SHORT> tag.

Examples:

```
<outcome id="0" type="bad">
  <short>kwashiorkor<short> ( defined by the <short>presence of oedema</short>
</outcome>
```

```
<outcome id="1" type="bad">
   admitted for <short>worsening heart failure or to die</short>
</outcome>

<outcome id="2" type="good">stopped smoking</outcome>

<outcome id="3" type="bad">symptomatic venous thromboembolism</outcome>
```

2.3 Follow-up times

Times associated when an outcome is measured (e.g. “six weeks”, “5 months”, “after treatment”, “six week follow up”) are tagged with <TIME> tag, which has the attributes:

- id, which is unique to the particular time in the study, as for groups and outcomes above.
- units, which specifies the units (e.g. “days”, “weeks”, “months”) for the follow-up time if they are not part of the annotated time string. This attribute may be omitted if the units are already part of the annotated text.

2.4 Group sizes

The number of people in a treatment group is tagged with the <GS> tag which has the attributes:

- group, which is the id of the group associated with this value
- time, which is the follow-up time for when the group has this particular size. This attribute is only needed if an outcome is measured at multiple times and the size of the group changes, due to people dropping out. It should not be needed most of the time.

2.5 Outcome numbers

The number of good or bad outcomes for a given treatment group is tagged with the <ON> tag which has the attributes:

- group, which is the id of the group associated with this value.
- outcome, which is the id of the outcome associated with this value.
- time, which is the follow-up time for when this outcome was measured. As with group sizes, this attribute may not always be necessary.

2.6 Lost to follow-up

The number of participants lost to follow-up is not usually explicitly mentioned in an abstract. However, if it is mentioned, the number of participants who were lost to follow-up at a given follow-up time is tagged with the <LOST> tag which has these attributes:

- group, which is the id of the group associated with this value.

- time, which is the follow-up time for when the participants dropped out.