TREC 2017 Precision Medicine Track
Relevance Judgment Guidelines

Version *2017-06-30*

In the previous years of the TREC Clinical Decision Support Track, relevance assessors have judged results on a simple scale: *definitely relevant*, *partially relevant*, and *not relevant*. Due to the particular challenges involved with precision medicine, however, this is not necessarily appropriate. Not only is precision medicine a highly specialized field (and thus difficult to get true experts to act as assessors), but the notion of relevance is far more flexible and case-specific. As such, relevance judgements will follow more specific guidelines.

Topic Structure

Each topic has four primary fields: *Disease*, *Variant*, *Demographic*, and *Other*. For instance:

Disease:	Acute lymphoblastic leukemia
Variant:	ABL1, PTPN11
Demographic:	12-year-old male
Other:	None

Disease specifies the particular disease for the patient (in this case, a form of cancer).

Variant specifies the genetic variants the patient has. This includes the "abnormal" gene and optionally the particular variant. Patients may have more than one genetic variant as well.

Demographic specifies the basic demographics (age and sex) of the patient.

Other specifies other potentially relevant factors the oncologist thought important, or "None".

Result Type

There are two types of results to consider: *Scientific Abstracts* and *Clinical Trials*.

A *Scientific Abstract* is a short text summarizing a scientific publication, presentation, or other research endeavor. The primary utility of abstracts for precision medicine is: *will this abstract provide information relevant to the treatment of the patient's cancer?*

A *Clinical Trial* describes research trial involving the recruitment and testing of human research subjects. The primary utility of clinical trials for precision medicine is: *is the patient eligible for this clinical trial*?

Result Assessment

Judging an individual result, whether an abstract or trial, will proceed in a cascaded manner. There is an initial pass to ensure the abstract/trial is broadly relevant to precision medicine, after which the assessor must categorize the abstract/trial according to the four fields above.

See Figure 1 below for a flow-chart style overview of this process. The first step is designed to save assessor time by filtering out unrelated abstracts/trials, since the second step can take more time (and possibly more detailed reading of the abstract/trial). The assessor is free to quickly skim the abstract/trial in order to make the initial decision. Then, if the abstract/trial is relevant to precision medicine (by the standard outlined below), a more detailed reading may be necessary in order to accurately assess all fields.

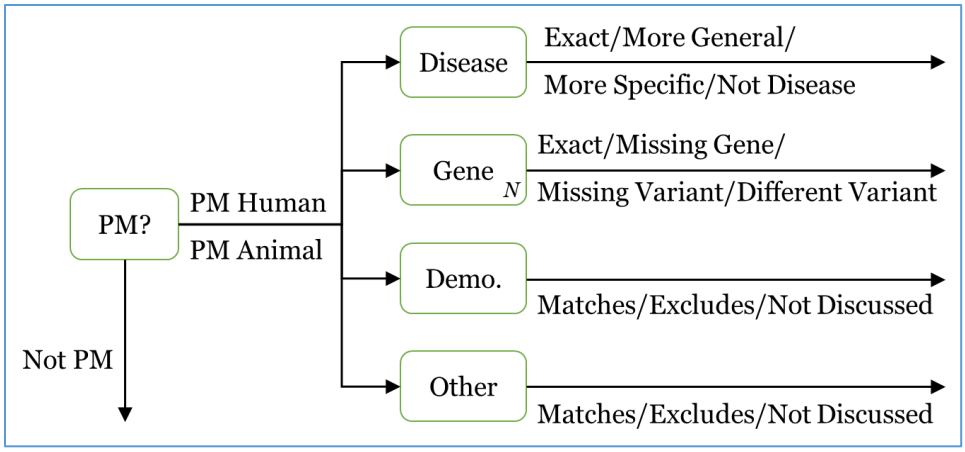

Figure 1: Overview of Result Assessment

Step 1 is to determine whether the abstract/trial is related to precision medicine. There are three options:

- **Human PM**: The abstract/trial (1) relates to humans, (2) involves some form of cancer, (3) focuses on treatment, prevention, or prognosis of cancer, and (4) relates in some way to at least one of the genes in the topic.
- **Animal PM**: Identical to Human PM requirements (2)-(4), except for animal research
- **Not PM**: Everything else. This includes "basic science" that focuses on understanding underlying genomic principles (e.g., pathways), but provides no evidence for treatment.

Step 2 is to determine the appropriate categorization for each of the four fields:

1. *Disease*:
    - **Exact**: The form of cancer in the abstract/trial is identical to the one in the topic.
    - **More General**: The form of cancer in the abstract/trial is more general than the one in the topic (e.g., blood cancer vs. leukemia).
    - **More Specific**: The form of cancer in the abstract/trial is more specific than the one in the topic (e.g., squamous cell lung carcinoma vs. lung cancer).
    - **Not Disease**: The abstract/trial is not about a disease, or is about a different disease (or type of cancer) than the one in the topic.

2. *Gene* [Note: this should be done for each particular gene in the topic]
    - **Exact**: The abstract/trial focuses on the exact gene and variant as the one in the topic. If the topic does not contain a specific variant, then this holds as long as the gene is included. By "focus", this means the gene/variant needs to be part of the scientific experiment of the abstract/trial, as opposed to discussing related work.
    - **Missing Gene**: The abstract/trial does not focus the particular gene in the topic. If the gene is referenced but not part of the study, then it is considered missing.
    - **Missing Variant**: The abstract/trial focuses on the particular gene in the topic, but not the particular variant in the topic. If no variant is provided in the topic, this category should not be assigned.
    - **Different Variant**: The abstract/trial focuses on the particular gene in the topic, but on a different variant than the one in the topic.

3. *Demographic*
    - **Matches**: The abstract/trial demographic population matches the one in the topic.
    - **Excludes**: The abstract/trial demographic population specifically excludes the one in the topic.
    - **Not Discussed**: The abstract/trial does not discuss a particular demographic population.

4. *Other*
    - **Matches**: The abstract/trial population matches the one in the topic. If the other field is "None", this category should also be assigned.
    - **Excludes**: The abstract/trial population specifically excludes the one in the topic.
    - **Not Discussed**: The abstract/trial does not discuss a population relating to the provided factors.

Relevance Assessment

*Note that this part of the guideline is **DRAFT**. However, this part is not relevant to the human assessment process. Human assessors simply assign the categories above. Further, without an existing benchmark dataset to train/tune systems, the exact mapping of categorical assessments to the relevance score is only marginally useful to TREC participants.*

In order to go from the items above to an information retrieval style notion of relevance, an abstract/trial result is judged according to all four of the categories. In order to calculate a relevance score (i.e., what is used for metrics like P@10, infNDCG), the human-assigned categories are automatically converted to a relevance score.

In order to be **Definitely Relevant**, a result should have *Disease* in {*Exact*, *More General*, *More Specific*}, at least one *Gene* is *Exact*, and both *Demographic* and *Other* are in {*Matches*, *Not Discussed*}.

In order to be **Partially Relevant**, … [*this has not been decided yet, it is actually possible there will be no notion of partial relevance, in which case some of the retrieval metrics might need to be altered*]