TREC 2020 Precision Medicine Track
Relevance Judgment Guidelines

Version *2020-07-22*

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.

In addition to finding relevant articles, the judges will assess the level of evidence provided by the article.

Topic Structure

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

Disease: melanoma
Variant: BRAF (V600E)
Treatment: Dabrafenib

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.

Treatment specifies the proposed treatment for the patient.

The assessor should spent at least 20-30 minutes up front investigating this topic on her/his own to gain an understanding of the background of the disease/gene/treatment. This can be done by using Google to search for the disease and gene, disease and treatment, or gene and treatment. Ideally, the assessor will have a reasonable level of understanding of how the genetic variant relates to the type of cancer, as well as the role the treatment plays. This initial conception may be changed as the assessment process continues (in which the assessor will be exposed to a sizable amount of science on the subject), so an open mind is necessary throughout the project. However, some initial understanding of the topic will be important for consistent judging.

Result Type

The search results are scientific abstracts. 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?*

Result Assessment

The first judging step is Result Assessment, which itself proceeds in a cascaded manner. There is an initial pass to ensure the abstract is broadly relevant to precision medicine, after which the assessor must categorize the abstract according to the fields above.

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

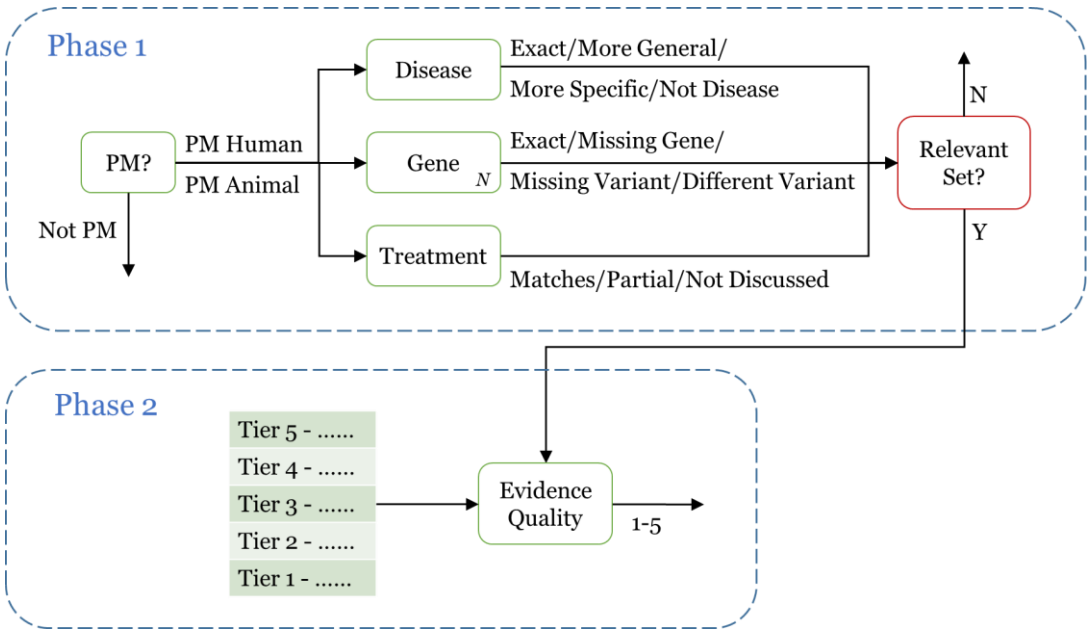

Figure 1: Overview of Judgment Process

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

- **Human PM**: The abstract (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 is identical to the one in the topic.
    - **More General**: The form of cancer in the abstract is more general than the one in the topic (e.g., blood cancer vs. leukemia).
    - **More Specific**: The form of cancer in the abstract is more specific than the one in the topic (e.g., squamous cell lung carcinoma vs. lung cancer).
    - **Not Disease**: The abstract 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 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 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 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 focuses on the particular gene in the topic, but on a different variant than the one in the topic.

3. *Treatment*
    - **Matches**: The abstract directly evaluates the proposed treatment.
    - **Partial**: The abstract evaluates the proposed treatment as part of a drug combination.
    - **Not Discussed**: The abstract does not evaluate the proposed treatment.

Relevance Assessment

This step is an entirely automatic process (the top-right box in Figure 1). Assessors will inform the coordinator that they are finished with Phase 1 of the assessment. The results above will be automatically converted to a 3-point relevance scale as follows.

In order to be **Definitely Relevant**, a result should have *Disease* in {*Exact*, *More General*, *More Specific*}, at least one *Gene* is *Exact*, and *Treatment* is *Matches*.

In order to be **Partially Relevant**, a result should largely be the same as Definitely Relevant, but with the exception that *Disease* can also be *More General*; *Gene* can also be *Missing Variant* or *Different Variant*; and *Treatment* can also be *Partial*.

All other results are **Not Relevant**.

Evidence Assessment

After conducting Phase 1 and informing the coordinator that this portion of the assessment is complete for the topic, Phase 2 begins (the bottom of Figure 1). The assessor will develop a 5-point rubric for grading the quality of evidence of a study. A separate document provides further guidance on this matter, but essentially the goal is to help identify the most important relevant studies so that these can be prioritized. Based on the understanding the assessor gains through all the reading in Phase 1, a 5-point scale will be developed. As one possible example, the top tier (Tier 5) may only include randomized controlled trials for the specific drug (not in combination), while the bottom tier may only include animal studies. Another example could be placing RCTs on two different tiers based on how strong a result the study demonstrated (e.g., very positive or negative results on a higher tier, less conclusive results on a lower tier). Different scales are possible: precision medicine is a flexible paradigm and the scale that the assessor develops should be relative to the studies encountered during Phase 1. For instance, if no animal studies were encountered, there is no need to include this as an evidence tier and more fine-grained tiers can be utilized. An important point to note, however, is that a positive result (the drug definitely worked) and a negative result (the drug definitely did not) are considered equal in terms of evidence—both of these are preferable to weaker or inconclusive results.

Ideally, the relevant documents from Phase 1 will be equally distributed throughout the tiers, but this does not need to be the case. It is much more important to distinguish between different levels of evidence than to develop a rubric that equally distributes the abstracts.

The assessor will document the proposed tiers based on a provided template, then submit this to the coordinator for approval. After approval, the assessor will proceed to re-judge a sample of the relevant abstracts according to this tier.

This re-judging according to the topic-customized rubric will enable search engines to not only retrieve what is “relevant”, but to properly rank scientific articles by quality of evidence.

To give the assessor a sense for how evidence-based medicine can separate articles into tiers, consider the GRADE system:

Initial quality of evidence	Study design	Lower if	Higher if
High	**RCT, systematic review, meta-analysis**	Study limitations: **1↓ Serious** **2↓ Very serious** Inconsistency: **1↓ Serious** **2↓ Very serious** Indirectness: **1↓Serious** **2↓Very serious** Impression: **1↓Serious** **2↓Very serious** Publication bias 1 **likely** **2↓Very likely**	Magnitude of effect: **2↑ Very strong** **1↑ Strong** Dose-response gradient **1↑** All plausible confounders would have reduced the effect **1↑**
Moderate			
Low	**Observational study (cohort study, case control study**		
Very low	**Any other evidence (case series, case study)**		
Definition: Overall quality of evidence across studies for the outcome level A：「High」 level B：「Moderate」 level C：「Low」 level D：「Very low」			

Note that this system is in no way specific to precision oncology, so some translation to the individual topic will be necessary. This is why this step happens after the prior phase completes: only the assessor that has viewed all the articles can have a sense of the diversity of studies that are available for placing items into tiers.

Factors that can impact the quality of evidence:

- The type of study (RCT vs. observational vs. pre-clinical/case study) typically the largest impact on the level of evidence. The tiers of study type are well-established in evidence-based medicine. See the GRADE scale above.
- The size of the trial (how many patients).
- The diversity/inclusiveness of the trial (multi-center, multi-ethnic trials are superior to single-site trials that exclude major sub-populations based on race, ethnicity, sex, age, etc.)
- The impact of the treatment in the population, which can be both positive (the drug has a clear beneficial effect) or negative (the drug is either harmful or clearly shown to not be beneficial). This roughly corresponds to the statistical notions of *p*-value and effect size, but the assessor is encouraged to think more broadly than simply grouping by p-values. In short: how likely are the study results to change the priority level of this treatment (e.g., do the results indicate this treatment should this treatment be tried prior to other treatments)