Annotation Guidelines PK Relation Extraction

Background

Aim: We aim to develop a relation extraction model to retrieve numerical estimations of PK parameter values from scientific articles. To do so, we focus on relations existing at the sentence level between multiple entities. In this annotation task, we will focus on information around the PK numerical estimations and future studies will complement numerical estimations with contextual data e.g. drugs, diseases, conditions, covariates etc.
In this annotation task, we aim to collect annotation data to train a relation extraction algorithm capable of (1) detecting all the relevant **entities** involved (e.g. PK mentions, numerical values, units, etc.) and (2) their **relations** in scientific text.

Method: To do so, we sample sentences from the abstract, methods, results and discussion sections within our corpus of PK articles. Then, annotators will be asked to label spans of text corresponding to 5 entities and their relations.

Entities and relations

In this annotation task, PK experts will have to annotate **(1) entities** and **(2) relations** for every sentence displayed in the interface. In this section, we describe the types of entities and relations involved.

Entities

In natural language processing (NLP) entities are **spans of text** that correspond to specific **concepts**. For instance, mentions of organisations, persons, countries etc. In this task, we are interested in **five** entities:

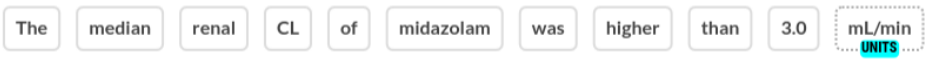
1. **PK**

   Mentions of pharmacokinetic (PK) parameters. This entity refers to spans of text that mention kinetic parameters. Any type of kinetic parameter in the context of PKs will be highlighted. For a list of PK, parameter types see the following tables: in vivo, in vitro. If not sure whether a specific mention should be labelled see but see the Questions & Answers section. Example:

TO ANNOTATE: PK parameters are **pre-highlighted** in the interface by our model. But, there will be cases in which the model **missed** those mentions or **incorrectly predicted** them. Therefore, we need to **check** that PK mentions are well annotated and **correct any mistakes**.

2. **UNITS**

Spans of text corresponding to units of numerical PK estimations. Example:



TO ANNOTATE: In the initial annotation rounds, **UNITS will not be pre-highlighted**. This means that in every example the user needs to **look for any units** and highlight their spans. This is the entity that requires the **most attention** from the annotator. After the initial annotation rounds, UNITS will be pre-highlighted.

3. **VALUE**

Spans of text that refer to numerical values. This includes single numbers, decimals, exponential expressions etc. Example:

TO ANNOTATE: For this, we use a set of **rules** to **pre-highlight** values in the interface. In general, the rules work well and tend to encapsulate numerical values. However, there might be eventual mistakes and the user **might need to correct** them if that happens.

4. **RANGE**

Two numerical values defining a range. On some occasions, numerical estimations might be expressed in the form of ranges:

TO ANNOTATE: We **pre-highlight** those terms using regular expressions. However, the rules are still quite limited and new cases are likely to appear. This means that the user will need to **pay attention** to potential **new RANGE** spans.

5. **COMPARE**

   Comparative terms. The mentions of this entity aim to provide information on whether specific PK estimations refer to the maximum or minimum estimated value. In the literature one might find:

   The COMPARE entity will help us to acknowledge that 3.0 is the minimum value. Some COMPARE terms include: >, <, higher, lower, maximum, minimum, exceeded etc. In essence, COMPARE will give information on whether the extracted number refers to an **estimated boundary** (e.g. max, min, >, <, exceeded). Mentions like **"approximately", "~", "close to" are not considered COMPARE** mentions since they do not indicate whether the value is a minimum or maximum but the confidence of the prediction.

   TO ANNOTATE: We **pre-highlight** those terms using an in-house dictionary. The dictionary is still quite limited and new cases are likely to appear. This means that the user will need to **pay attention** to potential **new COMPARE** mentions.

Relations

Once entities are annotated, the next step is to annotate **relations between entities**. Relations always need to happen between entities and **some relations only happen between specific types of entities**.

In this task we consider **three** relations:

1. **C_VAL**

   Central Value. This is a relation between a PK parameter mention and its estimated value. This type of relation only happens between the following entity types:

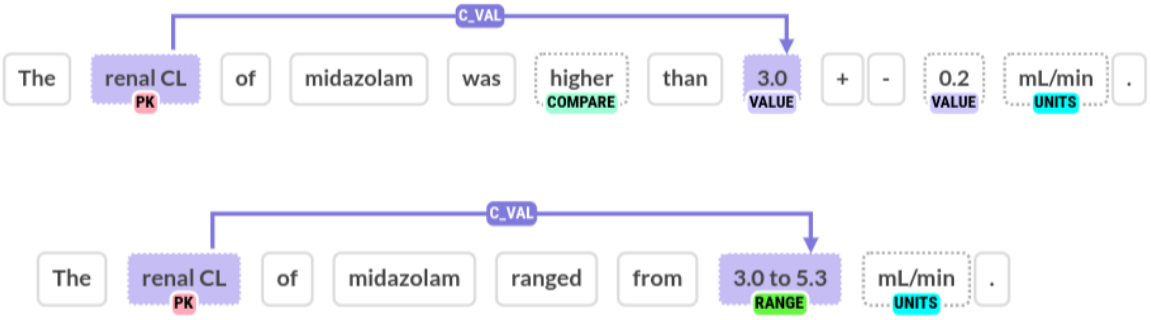
   PK → VALUE
   PK → RANGE

   It's only annotated when the VALUE/RANGE refers to a measurement but NOT A COMPARISON. For instance, there are many comparative sentences that mention:

   *"The **CL** of midazolam increased by **3%** when co-administered with amoxicillin"*

In this case, there is **no C_VAL** relation between CL and 3 since 3 is not a numerical estimate of CL.

Examples:

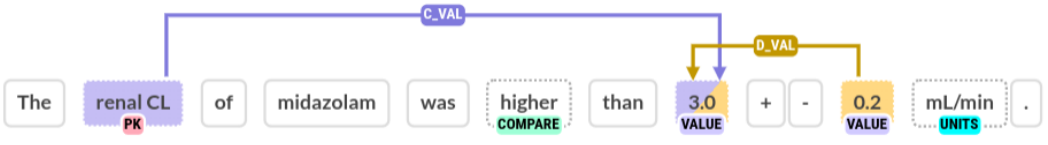


2. D_VAL

Deviation Value. This relation happens between central measurements and their deviation values/ranges. This relation is **only annotated if a C_VAL relation exists** and between deviation values/ranges and central values/ranges. So, this type of relationship only happens between the following entity types:

VALUE → VALUE (previously labelled with C_VAL relation)
VALUE → RANGE (previously labelled with C_VAL relation)
RANGE → VALUE (previously labelled with C_VAL relation)
RANGE → RANGE (previously labelled with C_VAL relation)

Example:

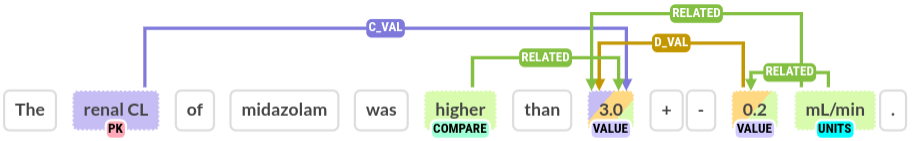


3. RELATED

This type of relation happens between multiple entity types and serves for adding any complementary information for the central or deviation values:

COMPARE → VALUE/RANGE (previously labelled with C_VAL or D_VAL)

UNITS → VALUE/RANGE (most common)

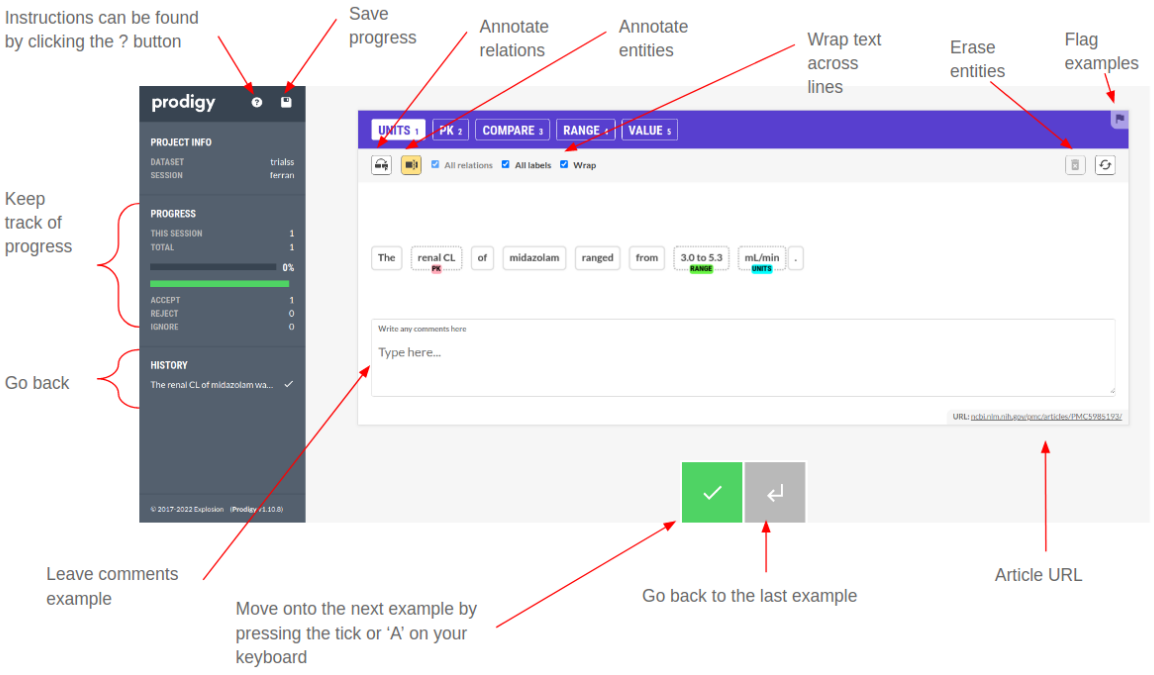


Example:

NOTE: We will **only annotate relations between units and values/ranges if the VALUE/RANGE is part of a C_VAL or D_VAL relation**.

Task description and interface

The interface displays a single sentence with some entities pre-highlighted by the NER model. There are several options available on the interface:

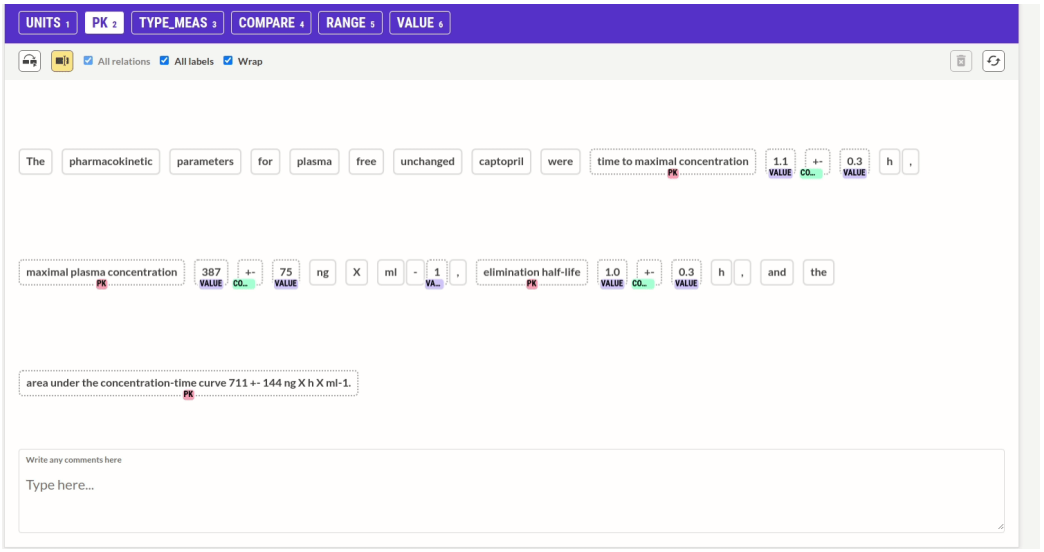


[Click here for a video tutorial on the interface]

Questions, Answers and Common doubts

Q: “What happens if I want to label a span across two lines?”
A: You will need to unselect the “Wrap” tick to be able to see the sentence in a single line. Then select the span and you can click “Wrap” again. For an example see the clip in the next question

Q: “How are we meant to erase spans?”
A: Select one of the available entities, then click on top of the entity that you’d like to erase and select the bin button on the top right. See following clip:



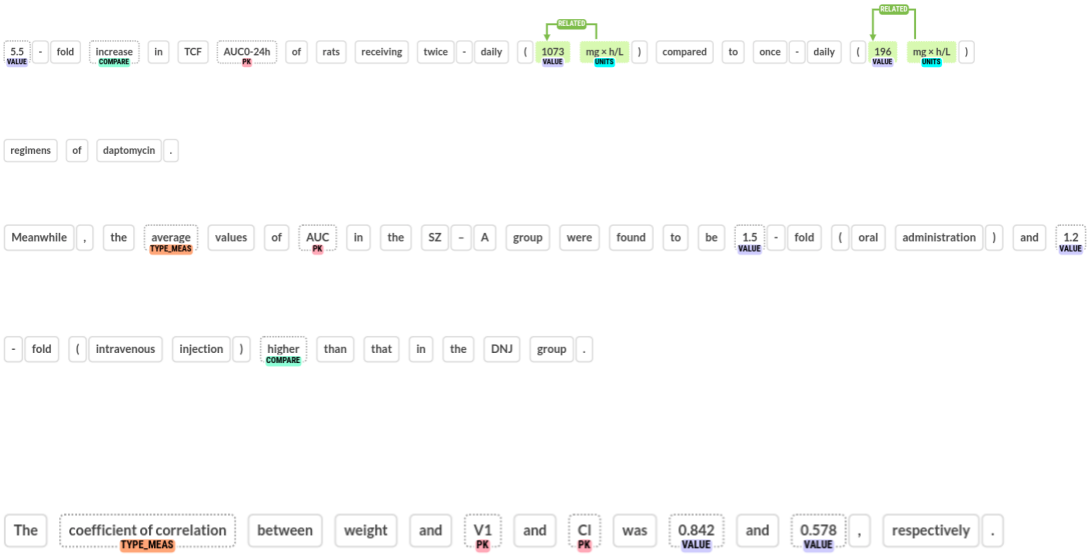
Q: “Do we need to annotate relations between COMPARE and VALUE if the relation doesn’t refer to a central value (C_VAL) or deviation (D_VAL) measurement?”
A: Let’s **not label** it. We only associate COMPARE-VALUE if that VALUE is also part of a C_VAL or D_VAL relation.

In-depth doubts and resolutions from the initial sentences in the test set can be found here: [Comments test set 0-200]

- ## Percentages and fold increases

  Often authors report a comparison of PK parameters between different conditions/drug treatments/patients in the form of % of increase or fold-increase. Some of you had doubts about whether we needed to annotate the relation between PK and those %/fold increases.

The answer is no. In this task, we are trying to extract exact measurements for PK parameters mentioned in the sentence. Unless the PK mention itself is a ratio (e.g. AUCR, AUC/MIC, relative bioavailability (see next point below)) we won't link their relation to the fold/% of increase. So in these examples, there won't be any C_VAL relation between PK and VALUES:



However, if the PK mention is a ratio itself (AUC/MIC, AUC1/AUC2) we will annotate the numerical value associated with it.

- ## Ratios

Single mentions in the form of ratios such as AUC/MIC, AUC1/AUC2, AUCR, relative bioavailability can be considered PK parameters and linked to their estimated values. However, if we cannot understand that the measurement is a ratio given the PK span mention we will not label it. For instance in the example:

"geometric mean ratios for AUC were 1,2,3…

We would need an additional entity type potentially "type of measurement" (instead, relative bioavailability, AUC/MIC etc, are known parameters that people might search for) to understand that 1,2,3 are ratios of AUCs and not AUC so we won't annotate any C_VAL in that case.

On the other hand, if the sentence says:
"The AUC1/AUC2 were 1,2,3"
From the PK mention "AUC1/AUC2" we can understand that this is a ratio without additional information, so we will label the C_VAL and related entities.

- ## P-values

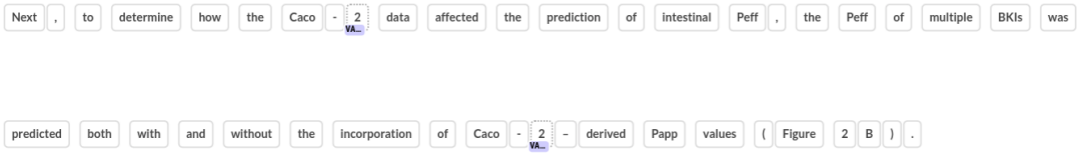
  Since p-values are the result of a comparison to a null hypothesis and not an estimated PK value **we won't annotate them**. When finding p-values you can leave them as they are without the need to correct entities or annotate any relations

- ## PD parameters

  Yes, we are including PD parameters if they are mentioned, so please label/correct them if not detected by the algorithm

- ## Remove irrelevant VALUE entities?

  We often find numerical values that are part of chemicals, enzymes, genes, tables, equations and don't have anything to do with real estimations of values that have units associated with them.

  In these cases, we could either leave those VALUE entities or remove them. We ask annotators to **not modify irrelevant VALUE entities unless they are part of a C_VAL or D_VAL relation**. So, in general, we can leave those values as they appear and only modify them if they need to be modified in order to make a C_VAL or D_VAL relation properly (rare case). For instance, if in the following sentence we only find this highlighted:

  "The clearance of MDZ was 3*10-2"

  The annotator would need to modify this span in order to annotate the C_VAL relation between clearance and the estimated value:

  "The clearance of MDZ was 3*10-2"

  Otherwise, no need to touch VALUE entities since those VALUES with no associated relations will be removed automatically.

- ## Confusing COMPARE entity

  COMPARE is only used to identify those values that are maximum, minimums or lower than or higher than the reported value. For instance:

  “The CL was higher than 3”
  “The CL was > 3”
  “The CL was < 3”

  In these cases, we want to relate those COMPARE to the value since it will tell us that 3 is not the exact estimate for CL. However, on some occasions, it was not clear whether they should be linked to the value. Consider the following reduced

  In this case, reduced does not affect 8.68% since the estimated value for “absolute bioavailability” is not higher or lower than 8.68% but 8.68%. For this reason, we will leave it as it is and not annotate any relation to these cases.

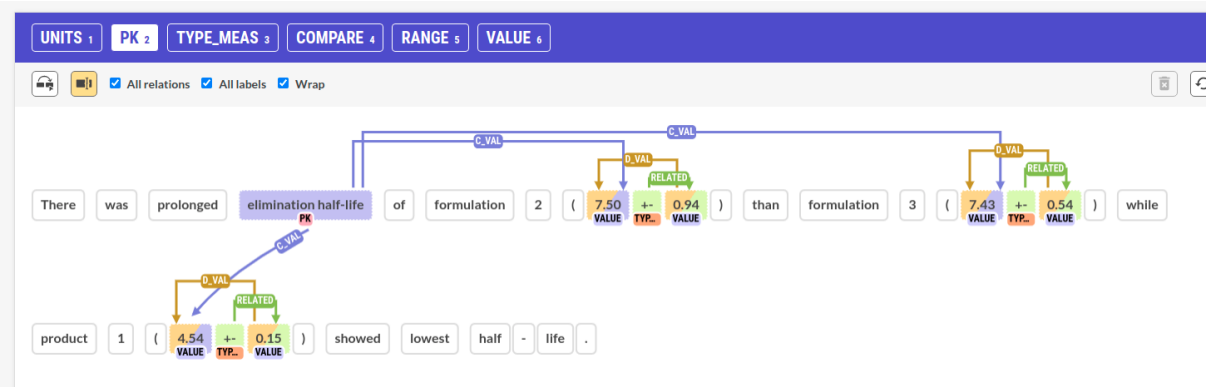
- ## What about all the missing context? Drugs, doses, species, study design etc are not annotated?

  We saw many comments about how to annotate the drugs, doses, and species related to the PK measurements. We certainly care about this context to disambiguate values and also to filter for the desired parameters in a specific drug/population/etc.

  However, to simplify the labelling process, we split the task into 2 parts and **won’t annotate this complementary information in this task**. When we finish the annotations of this task, then we will complement the central values with all the relevant context, but this will come at a **later stage**.

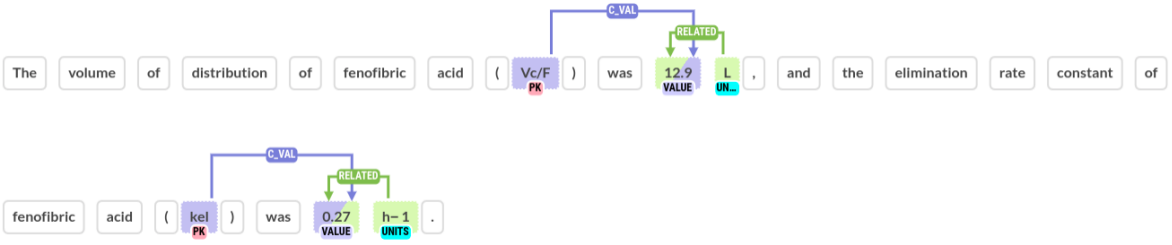
- ## Two mentions referring to the same parameter

  On some occasions, we observed more than one PK mention referring to the same PK value and parameter. For instance, in the following sentence we can see that half-life is mentioned twice, at the beginning, and end of the sentence:



On these occasions, we will prefer to label only the **closest left-side PK mention** as part of the relation

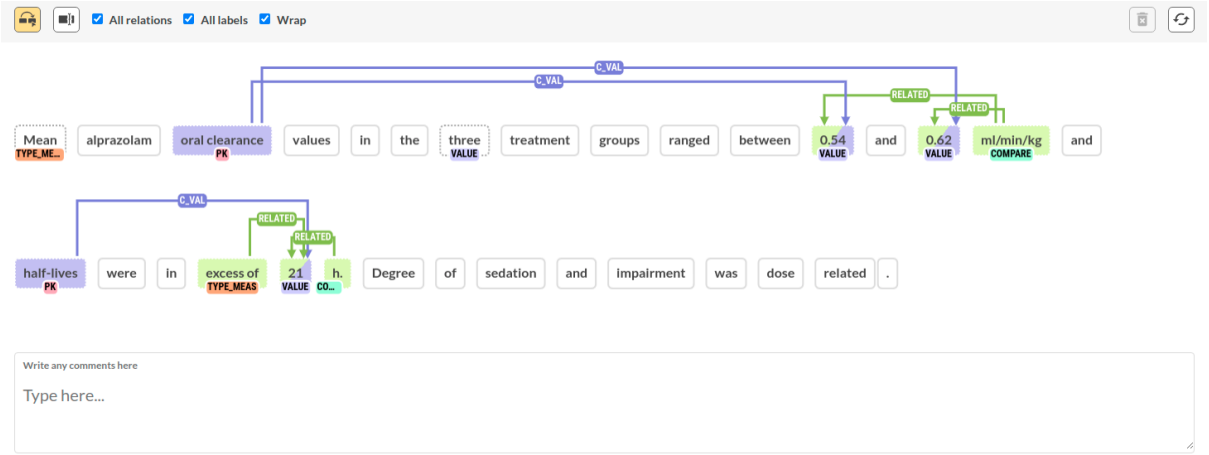
Similarly:



> In this case, labelling volume of distribution of fenofibric acid (Vc/F) as a single PK mention would have the issue of including the compound inside the PK entity, so the actual PK entities would be volume of distribution of fenofibric acid (Vc/F). In these cases, we will only consider the **second mention (abbreviated form)** as part of the relation and associate it with the central value

- Range vs Values

Sometimes we might be confused on whether labelling 2 values or 1 range:



In general, if the values refer to different experiment designs (different doses, drugs, patient cohorts etc) we will label 2 values. However, if the range is for a single condition and is the result of variability in the measurement we will label it as a range:

“The AUC for midazolam ranged from 3 to 5”
“The AUCs for midazolam and amoxicillin were 3 and 5”

If units are within the range, then we label 2 separate values. For instance:

“Clearance ranged from 0.54 ml/min/kg to 0.62 ml/min/kg”

If we decide to make 0.54 ml/min/kg to 0.62 ml/min/kg a range span, that would overlap with units, which can be an issue. So we will label two separate values and C_VAL relations.

- Value +- Range

  This refers to the cases wherein “VALUE (RANGE)” the value is the central value of a PK parameter and RANGE is the measured range of the parameter. You could either consider the RANGE a central value of the parameter or a D_VAL of the VALUE. We decided to label these cases as the latter since then we know both, VALUE and RANGE come from the same estimation and there is no variation between them given by the experimental context:

- IIV

  We considered inter-individual variability as a form of deviation: