**Files:**

**The corpus is composed of 6 excel files.**

**Positive Data:**

**PK DDI IN VIVO**: Pharmacokinetics Drug-drug Interaction in vivo study

**PK DDI VITRO**: Pharmacokinetics Drug-drug Interaction in vitro study

**DDI CLINICAL**: Drug-drug Interaction Clinical study

**Negative Data:**

**Non-Single Drug**: Single Drug studies. Mixed with PK in vivo, PK in vitro, PD in vitro, and Clinical studies.

**Non-DDI\_PD-Nutrition**: Pharmacodynamics DDI in vitro studies and Drug-Nutrition studies.

**Non-Case-NonDrug**: Case reports and Drug irrelevant studies.

**Sheet Tabs:**

**Each excel file contains 9 worksheets.**

**The 1st sheet** with the same name of the excel file displays the pairwise agreement among the annotators. The form on the left is the agreement of the 1st round annotation. While the one on the right is the agreement after the validation by one of the expert annotator.

**The sheets 2 to** 9 are the subsets of the corpus with annotation tags. The code of the sheet name is “File name + subset #\_Annotator #-Annotator #”. For example, in PK DDI IN VITRO file, the sheet IN VITRO 1\_1-6 means this is the 1st subset of IN VITRO study, annotated by Annotator 1 (Shijun) and Annotator 6 (Anvesh).

**Corpus:**

In each sheet of the corpus, the name of the annotator who validated the annotations is listed in the first row. The first column is the PMID of the abstract. The second column is the sequence number of the sentence in the abstract. The third column is the sentence/fragment 1. Then the annotation of fragment 1 is followed (column 4). If the sentence is broken into two, the 2nd fragment and its annotation is put in the column 5 and 6.

When two or more successive sentences have the same annotation tag, we combined them as one fragment. When we are doing the annotation, “**11**” is given for the previous sentence(s), and the annotation tag is only assigned to the last sentence.