# CONSORT

CONsolidated Standards Of Reporting Trials aka CONSORT, is a set of guidelines for parallel group randomized controlled trials. There are 25 checklist items in the CONSORT statement The aim of this repository is to classify each of the sentences in the critical trial journals.

## Notebooks

Notebooks folder contains mainly 3 important jupyter notebooks 1. **BERT**.ipynb - notebook to run BERT based models on the consort

\_\_NOTES\_\_ - Use this information while running BERT.ipynb

* **use\_top\_section** - Set this flag to **TRUE** if you want to concat BERT embeddding with **embedding\_column**
* **embedding\_column** in BERT.ipynb is name of the column in train and test which would be used if **use\_top\_section** is true

1. **data**.ipynb - convert manually annotated dataset from xml format to csv format with the following features

* **PMCID** - ID of the file/document from which sentence is taken
* **sentence\_id** - unique identifier of the sentence
* **sentence\_text** - text of the sentence
* **start\_char\_pos** - start position of the sentence in the entire document/file
* **end\_char\_pos** - end position of the sentence in the entire document/file
* **section** - section header of the sentence, e.g. **METHODS**
* **CONSORT\_Item** - category/label to which sentence belongs
* **top\_section** - highest order section header of the sentence

1. **xml\_data\_generator**.ipynb - extract **methods** related data from raw **.ann** files and convert them to csv format. Items extracted from these files include

* **file\_id** - file from which the sentence is taken
* **section** - section of the sentence (related to methods section)
* **sentence** - text of the sentence

## Files 1. **BERT**.py - python version of the **BERT**.ipynb notebook for end-to-end execution for multi-label classification 2. **run**.py - end to end execution of the BERT model created for multi-label classification to be used on new data. It accepts input in the form a list of sentences for which **CONSORT** labels are required. Currently it only supports **Methods** related labels

## Models and Data

We used **BioBERT** to train multi-label classification model on the CONSORT data

* Data path - */efs/sahilw2/data*
* methods\_data\_new.csv - output of **xml\_data\_generator**.ipynb containing sentences. Input to **run**.py
* manual\_train\_data\_FINAL.csv and manual\_test\_data\_FINAL.csv are methods section files used for training and testing
* Model path
  + **/efs/sahilw2/model/manual\_linh** - Trained model on methods data without section information
  + **/efs/sahilw2/model/manual\_linh\_top\_section\_concat** - Model trained on methods data concatenated with section information
  + **/efs/sahilw2/model/BioBert** - Pytorch format of **biobert** used as a base model for all the experiments in CONSORT