EPPI Reviewer Classifier for PubMed study designs

We have had pre-built classifiers for randomized trials, systematic reviews and economic evaluations available in EPPI Reviewer for many years. However, we are often asked whether there are classifiers available for other types of study. To address this, we put together searches in PubMed to create a training dataset, and built a machine learning classifier to predict the following study designs:

- Case Control Studies
- Case Reports
- Clinical Trial Protocol
- Clinical Trial
- Cohort Studies
- Comment
- Cross Sectional
- Editorial
- Guideline
- Letter
- Meta Analysis
- Qualitative Studies
- Randomized Controlled Trial
- Review
- Systematic Review

For each study type, we developed a search strategy that considered the relevant MeSH publication type to ensure precise retrieval. Otherwise, we used relevant MeSH terms to identify records relevant to that specific type. We excluded records that overlapped with other selected study types. Additionally, we limited our search to studies with MeSH Humans tag, human-indexed records, and having abstracts. Please see below for detail of the searches run to retrieve training records for each of the above study designs (Appendix, Table 4). We retrieved 10,000 random records for each of the specified study design (3,105 for clinical trial protocol), after we downloaded all PMIDs for each study types with the help of a python package (pubmed-api 2.1.2). The dataset split into train and test sets with 85% of records used to train the model and 15% retained for evaluation.

We built and evaluated a BERT-based model using token limits of 256 and 512 tokens over 1-5 epochs, meaning that 10 models were built and evaluated. The classification problem was limited to a single class per record. i.e. even if a study report contained information about e.g. a case report and a qualitative study, only one study design classification could be assigned.

The source code is available here.

The F1 scores for each of the models are shown in Table 1:

Table 1: F1 scores for each model

epoch	256 tokens	512 tokens
1	0.810100922	0.814615101
2	0.81417542	0.820302287
3	0.818030517	**0.823748044**
4	0.813586286	0.817034595
5	0.813344424	0.818087722

We selected the model with the highest F1 score for production and give performance statistics for each study design in Table 2. Accuracy varies from being above 90% for qualitative studies, case reports, clinical trial protocols, guidelines, meta analyses, and systematic reviews; down to accuracy in the 60-80% range for letters, clinical trials, editorials, comments and cohort studies.

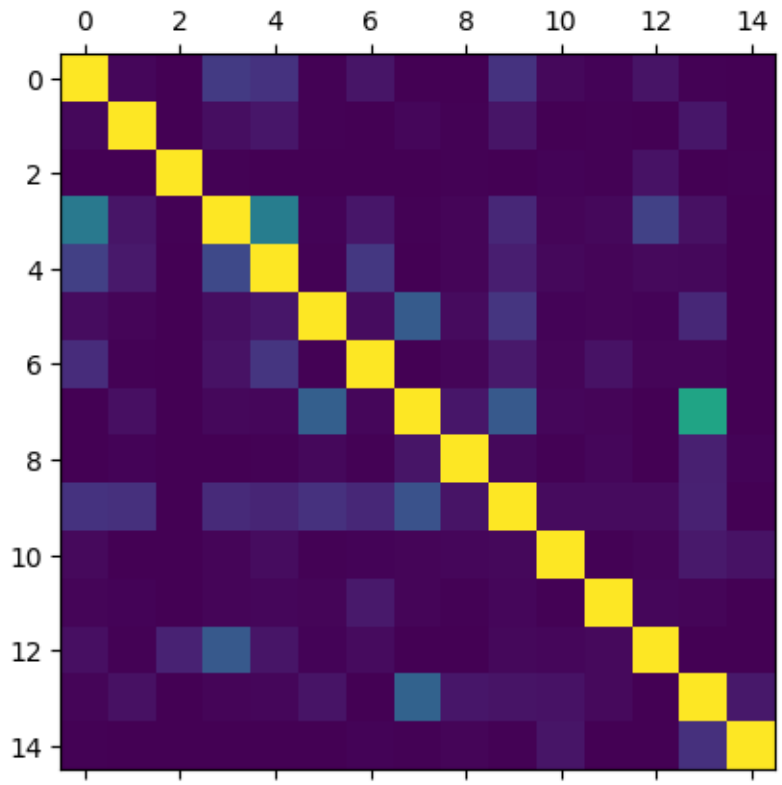
Table 2: performance of the classifier for each study design:

	Study design	Correct (N)	Total (N)	Recall (%)
1	Case Control Studies	1243	1500	82.87
2	Case Reports	1389	1500	92.60
3	Clinical Trial Protocol	430	466	92.27
4	Clinical Trial	955	1500	63.67
5	Cohort Studies	1166	1500	77.73
6	Comment	1180	1500	78.67
7	Cross Sectional	1293	1500	86.20
8	Editorial	956	1500	63.73
9	Guideline	1403	1500	93.53
10	Letter	998	1500	66.53
11	Meta Analysis	1390	1500	92.67
12	Qualitative Studies	1413	1500	94.20
13	Randomized Controlled Trial	1265	1500	84.33
14	Review	1214	1500	80.93
15	Systematic Review	1404	1500	93.60

Table 3: which study designs are most likely to be misclassified as what?

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1243	7	1	68	59	3	23	0	0	59	9	4	21	3	0
2	8	1389	0	15	25	2	0	7	2	23	1	3	1	24	0
3	0	0	430	3	0	1	0	0	2	1	4	2	20	0	3
4	161	23	2	955	169	4	24	2	5	45	6	9	77	18	0
5	76	27	1	89	1166	3	65	1	5	34	8	6	10	9	0
6	14	5	0	15	24	1180	13	115	12	62	4	6	4	45	1
7	51	2	0	20	62	3	1293	0	5	27	6	20	5	6	0
8	0	16	0	9	7	121	7	956	24	112	7	6	1	233	1
9	0	4	0	1	3	9	2	22	1403	8	2	7	0	35	4
10	59	56	1	48	42	57	45	101	21	998	11	11	12	37	1
11	10	0	0	6	13	2	4	5	7	8	1390	3	5	28	19
12	5	4	1	5	7	5	28	6	2	7	3	1413	7	6	1
13	17	2	38	112	23	4	12	0	0	9	7	10	1265	1	0
14	6	18	0	5	7	21	1	126	25	21	19	10	1	1214	26
15	2	0	0	0	0	0	4	2	6	3	22	1	0	56	1404

Figure 1: which study designs are most likely to be misclassified as what?



The classifier should not be considered sufficiently accurate for automated identification (or exclusion) of studies in systematic reviews, but can give an estimate of the likely distribution of study designs in a given set of records (titles and abstracts).

Appendix

Table 4: Search strategy for creating the training dataset (10 April 2024)

No.	Search string	Number of hits
1	("Case-Control Studies"[Mesh]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	187,940
2	("Case Reports" [Publication Type]) NOT ("Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	943,889
3	(Clinical Trial Protocol [Publication Type]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	3,105
4	("Clinical Trial" [Publication Type]) NOT ("Case Reports" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	210,699
5	("Cohort Studies"[Mesh]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Case-Control	696,448

	Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	
6	(Comment [Publication Type]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	10,069
7	("Cross-Sectional Studies"[Mesh]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	227,649
8	(Editorial [Publication Type]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	15,715
9	(Guideline [Publication Type]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	11,924
10	(Letter [Publication Type]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative	10,287

	Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	
11	(Meta-Analysis[Publication Type]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	31,580
12	("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh]) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	60,882
13	("Randomized Controlled Trial" [Publication Type]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial"[Publication Type:NoExp] OR "Adaptive Clinical Trial"[Publication Type:NoExp] OR "Clinical Trial, Phase I"[Publication Type:NoExp] OR "Clinical Trial, Phase II"[Publication Type:NoExp] OR "Clinical Trial, Phase III"[Publication Type:NoExp] OR "Clinical Trial, Phase IV"[Publication Type:NoExp] OR "Controlled Clinical Trial"[Publication Type:NoExp] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	175,357
14	(Review [Publication Type]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial" [Publication Type] OR Clinical Trial Protocol [Publication Type] OR Systematic Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	1,515,564
15	(Systematic Review [Publication Type]) NOT ("Case Reports" [Publication Type] OR "Clinical Trial" [Publication Type] OR "Randomized Controlled Trial"	20,942

	[Publication Type] OR Clinical Trial Protocol [Publication Type] OR Review [Publication Type] OR Meta-Analysis[Publication Type] OR Guideline [Publication Type] OR Editorial [Publication Type] OR Comment [Publication Type] OR Letter [Publication Type] OR "Cohort Studies"[Mesh] OR "Case-Control Studies"[Mesh] OR "Cross-Sectional Studies"[Mesh] OR ("Qualitative Research"[Mesh] OR "Focus Groups"[Mesh])) AND ((medline[sb]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND ("humans"[MeSH Terms]) AND ("hasabstract"[All Fields])	