

1 **Clinical study type classification, validation, and PubMed filter comparison with**  
2 **natural language processing and active learning**

3

4 David G P van IJzendoorn,<sup>1\*</sup> Philippe C Habets,<sup>1\*</sup> Christiaan H Vinkers,<sup>1#</sup> Willem M Otte <sup>1,2#</sup>

5 *\* Shared first author, # shared last author*

6

7 <sup>1</sup> DeepDoc Academy, Rotterdam, The Netherlands

8 <sup>2</sup> Department of Child Neurology, UMC Utrecht Brain Center, University Medical Center

9 Utrecht, and Utrecht University, Utrecht, The Netherlands

10

11 Willem M Otte, ORCID: <https://orcid.org/0000-0003-1511-6834>

12 David G P van IJzendoorn, ORCID: <https://orcid.org/0000-0002-2249-5919>

13 Philippe C Habets, ORCID: <https://orcid.org/0000-0001-5831-9020>

14 Christiaan H Vinkers, ORCID: <https://orcid.org/0000-0003-3698-0744>

15

16 **Abstract**

17 Each day, many thousands of new studies are published. Identifying specific study types with  
18 high sensitivity and specificity may improve searchability and accelerate updating systematic  
19 reviews and meta-analyses. Machine learning transformer models could facilitate this  
20 identification process if sufficient training data is available.

21 We used an active learning strategy to construct a large training set (n=50,000) and fine-  
22 tuned the PubMedBERT language model to classify PubMed abstracts as randomized  
23 controlled trials, human studies, systematic reviews with and without meta-analyses,  
24 protocols, and rodent studies. In an external dataset (n=5,000), the average sensitivity and  
25 specificity across study types were 0.94 and 0.96, respectively. PubMed's internal filters had  
26 a low sensitivity for both systematic reviews with meta-analysis (0.175, CI: 0.057–0.293) and  
27 randomized controlled trials (0.256, CI: 0.119–0.393). We applied this labeling to all 34  
28 million PubMed abstracts currently available and provide the results within an online meta-  
29 information platform (EvidenceHunt).

30 In conclusion, we show that study type classification in PubMed is opportune, given the  
31 available language models. The high accuracy in this study invites extending these models to  
32 more elaborate and hierarchical identification schemes.

33

34 **Keywords**

35 Study classification, transformer model, eHealth, literature search, PubMed

36

## 37 **Introduction**

38 Systematic reviews and meta-analyses of randomized control trial (RCT) results are the  
39 cornerstones of evidence-based medicine. Identifying the relevant RCTs in databases such as  
40 PubMed is essential to collecting and collaboratively analyzing new results. PubMed  
41 currently offers the largest and most comprehensive collection of abstracts of biomedical  
42 studies, currently spanning over 34 million abstracts. Unfortunately, these abstracts are not  
43 always accompanied by sufficiently reliable and standardized meta-information (Williamson  
44 and Minter 2019). Therefore, to conduct a systematic review, manual screening of all search  
45 results is essential to exclude irrelevant publications. This manual screening includes the  
46 identification of systematic reviews and meta-analyses that have already been published on  
47 the study in question, but also the existing individual RCTs must also be detected. With  
48 insufficiently reliable meta-data of study abstracts, this process can go suboptimal in two  
49 ways; irrelevant studies may be retrieved, or relevant studies may not be retrieved—the  
50 former event results in more reading work, the latter in the unnecessary omission of evidence.  
51 Therefore, current search strategies that scan raw abstracts for keywords are consistently set  
52 up to achieve high sensitivity (i.e., minimal number of false negatives), but this often results  
53 in low specificity (i.e., many false positives) (Beynon et al. 2013; Geersing 2012; Li et al  
54 2019). Therefore, it frequently happens that a systematic review requires thousands of  
55 abstracts to be screened to identify all relevant RCTs. In addition to RCTs, PubMed also  
56 contains many other study types, including narrative reviews, letters, editorials, preclinical  
57 studies, errata, and more fundamental biomedical research. Its low prevalence makes it  
58 challenging to identify RCTs without receiving large numbers of false positives. The time-  
59 consuming task of identifying relevant literature hampers rapid updating of emerging and  
60 often quickly developing evidence. PubMed offers filters for different study types.

61 Unfortunately, these filters' sensitivity and specificity are presumably too low to be useful for  
62 the identification of previously published systematic reviews, meta-analyses, and RCTs.

63 We propose a solution based on the rapidly developing field of natural language processing.  
64 Instead of creating a filter based on individual words, the overall coherence of the abstract is  
65 mapped and linked to a specific study type. General patterns from descriptions of different  
66 types of studies – including systematic reviews and RCTs – could then assign abstracts to a  
67 specific study type with high sensitivity and specificity. This approach requires a training set  
68 of enough abstracts assigned with study-type labels by human reviewers; it also requires  
69 proper validation to know whether the classification is sufficiently generalizable to serve the  
70 public interest.

71 In this study, we created a large training set (N=50,000), modeled the overarching patterns,  
72 and validated the resulting model on an independent dataset (N=5,000). We compared this  
73 new approach to existing PubMed filters. Because the prevalence of systematic reviews and  
74 RCTs is relatively low, we tested a new method of creating a training set applying active  
75 learning to present all potentially relevant studies to the rater rather than manually classifying  
76 all studies individually. Using this approach diminishes raters' efforts as it allows stopping  
77 when these potentially relevant studies are identified rather than labeling all studies in the  
78 dataset.

79 We hypothesized that it would be feasible be possible to generate a classification model, for  
80 different types of studies, with high sensitivity and specificity. We also showed a relatively  
81 low performance of the internal PubMed filters and an overall efficiency gain in generating a  
82 training data set using an active learning approach.

83

84 **Methods**

85

86 *Data collection*

87 In September 2022, we selected two batches of PubMed abstracts. The first dataset was for  
88 interrater agreement characterization, model training, and internal validation, and the second  
89 dataset was for external validation.

90 The first batch consists of the first 50,000 abstracts (out of 199,563) in the first week of  
91 January 2018 (query: ALL[SB] AND 2018/01/01:2018/01/07[DP] AND HASABSTRACT).

92 This batch was split into an interrater set of 5,000 and a training set of 45,000 abstracts. The  
93 second batch consists of the first 5,000 abstracts (out of 255,506) taken four years later  
94 (query: ALL[SB] AND 2022/01/01:2022/01/07[DP] AND HASABSTRACT).

95 We deliberately selected our external validation data several years later – after the COVID-19  
96 pandemic – as this would allow us to test the generalizability of our model if the literature  
97 gives rise to previously unseen topics and shifts in research trends.

98

99 *Data labeling*

100 All abstracts in the interrater set were independently labeled using a pre-specified protocol by  
101 two authors with at least fifteen years of PubMed experience (Vinkers and Otte). Abstracts  
102 were assigned one of the following study labels: 1) a randomized controlled trial, 2) a human  
103 study, 3) a systematic review without meta-analysis, 4) a systematic review with meta-  
104 analysis, 5) a study protocol, 6) a rodent study or 7) any other abstract type. Study labels are  
105 provided in **Table 1**. Not all abstracts contained valuable texts: in some cases, abstracts  
106 referred to a duplicated study. We removed these studies and ended up with a total dataset  
107 size of 4,867 abstracts.

108 Labeling was done assisted by active learning. For each study label, the two independent  
109 raters selected five relevant and irrelevant abstracts and used this to initiate a Naïve Bayes  
110 classifier. The classifier assigned label probabilities to all abstracts and presented the raters  
111 with the abstracts that had the highest probabilities for a given class. After each rating, the  
112 classifier was updated to re-assign the probabilities. The rating was continued until all studies  
113 were identified (operationally decided after 200 presented non-relevant abstracts). Because of  
114 the unbalanced distribution of the study types, we chose to use active learning. We used  
115 open-source ASReview LAB v1.0 for labeling.<sup>1</sup>

116 After pooling the active learning results and inventorying all disagreements, we determined  
117 the interrater labeling agreement. We relabeled the disagreements independently and resolved  
118 the latest remaining label disagreements in a final discussion round. The inter-rater agreement  
119 is expressed as Cohen's kappa coefficient ( $\kappa$ ). This qualitative statistic is more robust than  
120 simple percent agreement calculation, as it considers the possibility of the agreement  
121 occurring by chance. We used the kappa function from the R package 'psych' v2.2.9.

122 The training set of 45,000 abstracts was similarly labeled with active learning, except for the  
123 prevalent 'human study' type. Given the larger dataset and excellent interrater agreement (see  
124 Results), the data was split and labeled by one of the two rating authors. We combined this  
125 additional dataset with the first 4,867 abstracts and used it for model training.

126

### 127 *Model fine-tuning*

128 We fine-tuned PubMedBERT, a state-of-the-art language model pruned to the PubMed  
129 database. PubMedBERT has a transformer architecture to leverage unsupervised pre-training  
130 on a large multi-domain corpus of millions of PubMed abstracts by capturing the used  
131 scientific and clinical language. The transformer model has learned the textual context by

---

<sup>1</sup> <https://asreview.nl/>

132 mapping subword relationships in example corpora into a neural network. The model applies  
133 an evolving set of mathematical techniques, called attention or self-attention, to detect subtle  
134 ways, even distant data elements, in paragraphs up to a few hundred words. This attention  
135 mechanism has a substantial advantage compared to previous machine-learning designs. The  
136 textual pattern mapping could be done once, as a pre-training, to capture the general structure  
137 of a specific text genre effectively. Pre-training is computationally expensive and requires a  
138 vast amount of example data. PubMedBERT captures the textual structure of fourteen million  
139 PubMed abstracts containing 3.2 billion words (i.e., 21 Gb of raw text data).

140 We only need to fine-tune this pre-existent language model to our training set for our  
141 classification task. To that aim, we added one additional neural network layer to  
142 PubMedBERT and trained the entire network on the classification to predict which abstracts  
143 belong to which study type. The pre-trained PubMedBERT transformer was downloaded  
144 using the HuggingFace Python library (V4.17, python V3.9). The additional network layer  
145 had a dropout of 0.2, a reshape of 7, and a SoftMax layer. All abstract texts were tokenized  
146 and truncated to a maximum sequence length of 512.

147 As our set of labels is not larger enough to cover all abstracts – excluding, for example,  
148 studies on non-human primates, cell studies, classical reviews, or diagnostic and prognostic  
149 studies – we also added an additional study type label: ‘remaining’ to the training set,  
150 covering the unlabeled abstracts.

151 We split the training dataset into train, test, and eval datasets using an 80%, 15%, and 5%  
152 split, respectively, and trained the model until the loss did not improve for two epochs (12  
153 epochs total) at a learning rate of  $1e^{-5}$  in batches of 8 abstracts on a V100 NVIDIA graphics  
154 processing unit. We tested the model after each epoch (average accuracy on the eval samples)  
155 and selected the model instance with the best accuracy for external validation. The model was  
156 finally tested on the withheld test dataset (accuracy of 92.45%).

157

158 *External validation*

159 The model was trained on 2018 data only. External model validation was done in an  
160 independent dataset of 5,000 abstracts sampled in 2022. The two authors (Otte and Vinkers)  
161 raters independently labeled a random subset (i.e., 2,500 each) of these abstracts. We  
162 characterized the model's performance with sensitivity and specificity based on the true  
163 positive, true negative, false positive, and false negative numbers (human labels considered  
164 the 'gold standard'). Due to the relatively large number of studies not belonging to one of the  
165 study type labels, we subsampled this 'remaining' category such that we evaluated 1,000  
166 abstracts per rater to the model's predictions.

167

168 *Comparison to PubMed's internal filter types*

169 PubMed provides filters to select various study types, including 'Books and Documents',  
170 'Clinical Trial', 'Meta-Analysis', 'Randomized Controlled Trial', 'Review', and 'Systematic  
171 Review'. These labels are assigned to the studies by librarians; however, the exact criteria  
172 and overall labeling procedure are not publicly documented. Based on our experience, we  
173 notice various false positives and negatives, for example, the 'Randomized Controlled Trial'  
174 label assigned to an observational cohort study and a systematic review with meta-analysis  
175 lacking a 'Meta-Analysis' label. Therefore, we quantified the agreement between our human  
176 rater dataset (N 4,867) and PubMed's labels for a subset of labels. We restricted the  
177 comparison to 'Randomized Controlled Trial', 'Meta-analysis' given the lack of definition  
178 details on the other categories.

179



180 **Results**

181

182 *Interrater agreement*

183 The active learning approach accelerated the labeling of low-prevalent study types. For  
184 example, the prevalence of randomized controlled trials among PubMed abstracts is 4.4%.  
185 Systematic reviews with meta-analysis are even scarcer, at 1.7%. The approximate time to  
186 read and classify an abstract was 5-10 seconds. The kappa coefficients of the first and second  
187 labeling round are presented in **Table 2** and **Table 3**. We found high interrater agreement,  
188 particularly if mismatches were reevaluated in an independent second round.

189

190 *External validation*

191 We found excellent sensitivity and specificity on the external dataset with abstracts published  
192 four years later compared to the training data (**Table 4**). This indicates that the model is  
193 generalizable for texts written in different periods (i.e., post-COVID-19 pandemic), as the test  
194 dataset dated before the COVID-19 pandemic (2018), whereas the test dataset was four years  
195 later (2022).

196

197 *Comparison to PubMed's internal filter types*

198 PubMed's internal 'Randomized Controlled Trial' and 'Meta-Analysis' filters only partially  
199 captured the 'gold standard' manual labels (**Table 5**). The number of false negatives  
200 exceeded the number of false positives in both categories. The sensitivity was very low for  
201 the 'Meta-Analysis' (0.17) and low for the 'Randomized Controlled Trial' (0.26), indicating  
202 limited usability of PubMed's internal filters to select, for example, randomized trials for a  
203 systematic review.

204

205 *Implementation*

206 We implemented the model in an online cloud platform (<https://evidencehunt.com/>) after  
207 applying the study type prediction to all PubMed abstracts available (34.7 million). Users  
208 may query, select and export study types. The database is updated weekly with new abstracts  
209 to facilitate scholars and clinicians in need of robust study-type selection and precise search  
210 queries of the vast scientific literature.

211

## 212 **Discussion**

213 We show that it is possible to classify study types in PubMed abstracts with high sensitivity  
214 and specificity (>95%) based on a transformer model that is fine-tuned on an efficiently  
215 constructed training set. Creating a suitable training set requires a lot of time, mainly if the  
216 labels include low-prevalence categories. Because RCTs and systematic reviews are  
217 relatively infrequent in PubMed, a conventional method of training set construction requires  
218 labeling thousands of abstracts. By effectively using active learning, we were able to label  
219 only those abstracts that were marked as potentially interesting. This approach confirms a  
220 trend in the research field of systematic reviews where active learning has previously been  
221 very efficient. The current active learning environment is user-friendly and freely accessible.  
222 Because labeling via active learning mostly works with binary classifications (i.e., relevant or  
223 irrelevant), we repeated the process for each study type and later combined these labels to  
224 create the appropriate dataset for the multiclass classifier. The added value of active learning  
225 is interesting for constructing more detailed classification schemas. The added value may be s  
226 limited if the prevalence of positive entries is high (as in our case with ‘human study’ type).

227 One limitation of the current classification algorithm is that it does not cover all existing  
228 PubMed abstracts. For example, classical reviews, cell line studies, and prediction studies are  
229 missing. Also, our classification only assigns a single label to an abstract. In practice, there  
230 are situations where an abstract describes both human and animal results, for example. Open  
231 trials are now not separately classified in a separate category, and there is currently no  
232 distinction between phase 1, 2, 3, and 4 clinical trials. However, this may be possible in the  
233 future with additional training. Hierarchical classification is desirable as meta-analyses may  
234 or may not be nested in rodent-oriented or human-oriented systematic reviews, but also  
235 necessary to establish a classification scheme separating children and adults. RCTs are

236 possible in children and adults, so assigning multiple labels to a single abstract is a necessary  
237 next step.

238 Machine learning models require exceedingly larger training data sets, we show that using  
239 active learning can accelerate the labeling process. With the exponential increase of  
240 publications each year, finding relevant publications is becoming more difficult. Here we  
241 show that transformer models can aid in selecting studies, potentially preventing wasting  
242 precious time and energy to find relevant publications.

243

244 **Reference list**

245

246 Beynon R, Leeflang MM, McDonald S, Eisinga A, Mitchell RL, Whiting P, Glanville JM.

247 “Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE”,

248 Cochrane Database of Systematic Reviews 2013, Issue 9. Art. No.: MR000022. DOI:

249 10.1002/14651858.MR000022.pub3.

250

251 Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. “Search filters

252 for finding prognostic and diagnostic prediction studies in Medline to enhance systematic

253 reviews”, PLoS One. 2012;7(2):e32844. DOI: 10.1371/journal.pone.0032844.

254

255 Li L, Smith HE, Atun R, Car LT. “Search strategies to identify observational studies in

256 MEDLINE and Embase”, Cochrane Database Syst Rev. 2019 Mar 12;3(3):MR000041. DOI:

257 10.1002/14651858.MR000041.pub2.

258

259 Williamson PO, Minter CI, “Exploring PubMed as a reliable resource for scholarly

260 communications services”, J Med Libr Assoc. 2019 Jan; 107(1): 16–29. DOI:

261 10.5195/jmla.2019.433.

262

263 **Tables**

264

<b>Study type label</b>	<b>Used working definition</b>
1. Randomized controlled trial.	An explicit mention of randomization. Excluding non-human primate and rodent trials.
2. Human study.	Human studies other than randomized controlled trials, observational and intervention (e.g., cohorts, open trials, or case-control studies).
3. Systematic review without meta-analysis.	Excluding non-systematic reviews, non-human studies, protocols, or systematic reviews with a meta-analysis.
4. Systematic review with meta-analysis.	An explicit mention of a meta-analysis.
5. Study protocol.	Study protocol for trials or systematic reviews.
6. Rodent study.	Only primary research studies on rats or mice, excluding rodent cell line research.
7. Misc.	All other studies.

265

266 **Table 1.** Study labels with pre-specified working definitions used for labeling.

267

268

<b>Study type label</b>	<b>Total (N)</b>	<b>Mismatch (N)</b>	<b><math>\kappa</math> (CI)</b>
Randomized controlled trial	218	14	0.77 (0.65–0.88)
Human study	2267	27	0.97 (0.96–0.98)
Systematic review without meta-analysis	354	32	0.78 (0.71–0.85)
Systematic review with meta-analysis	83	7	0.83 (0.71–0.95)
Study protocol	130	4	0.90 (0.80–1.00)
Rodent study	403	46	0.77 (0.71–0.83)

269

270 **Table 2.** After the initial active learning round, the interrater agreement is expressed as a  
271 kappa coefficient with 95% confidence interval. The total N is based on the finalized version.

272

273

<b>Study type label</b>	<b>Total (N)</b>	<b>Mismatch (N)</b>	<b><math>\kappa</math> (CI)</b>
Randomized controlled trial	218	2	0.97 (0.92–1.00)
Human study	2267	9	0.99 (0.98–1.00)
Systematic review without meta-analysis	354	6	0.96 (0.92–0.99)
Systematic review with meta-analysis	83	2	0.95 (0.88–0.99)
Study protocol	130	0	1.00 (1.00–1.00)
Rodent study	403	7	0.97 (0.94–0.99)

274

275 **Table 3.** After the second active learning round, the interrater agreement is expressed as  
276 kappa coefficient with 95% confidence interval. The total N is based on the finalized version.

277



278

Rater	Study type label	N	TP	TN	FP	FN	Sensitivity	CI (lower)	CI (upper)	Specificity	CI (lower)	CI (upper)
1	Randomized controlled trial	1000	31	956	5	8	0.795	0.668	0.922	0.995	0.99	0.999
1	Human study	1000	32	796	170	2	0.941	0.862	1.020	0.824	0.800	0.848
1	Systematic review without meta-analysis	1000	25	966	7	2	0.926	0.827	1.025	0.993	0.987	0.998
1	Systematic review with meta-analysis	1000	40	956	4	0	1.000	1.000	1.000	0.996	0.992	1.000
1	Study protocol	1000	18	959	23	0	1.000	1.000	1.000	0.977	0.967	0.986
1	Rodent study	1000	89	893	9	9	0.908	0.851	0.965	0.990	0.984	0.997
1	(Remaining, non-labeled) *	1000	542	251	5	202	0.728	0.697	0.760	0.980	0.964	0.997
2	Randomized controlled trial	1000	20	970	8	2	0.909	0.789	1.029	0.992	0.986	0.997
2	Human study	1000	4	799	196	1	0.800	0.449	1.151	0.803	0.778	0.828
2	Systematic review without meta-analysis	1000	11	976	13	0	1.000	1.000	1.000	0.987	0.980	0.994
2	Systematic review with meta-analysis	1000	60	930	9	1	0.984	0.952	1.015	0.990	0.984	0.997
2	Study protocol	1000	9	985	6	0	1.000	1.000	1.000	0.994	0.989	0.999
2	Rodent study	1000	76	887	35	2	0.974	0.939	1.009	0.962	0.95	0.974
2	(Remaining, non-labeled) *	1000	551	184	2	263	0.677	0.645	0.709	0.989	0.974	1.004

279

280 **Table 4.** The fine-tuned PubMedBERT model’s performance on unseen external data. TP,  
 281 true positive, TN, true negative, FP, false positive, FN, false negative. CI, 95% confidence  
 282 interval. We prevented skewed contingency tables and artificially high performances, by  
 283 pruning the ‘remaining’ category of unlabeled abstracts to a random subset, such that the total  
 284 sample size per rater equals N 1,000 (originally n=2,500).

285

286

<b>Study type label</b>	<b>N</b>	<b>TP</b>	<b>TN</b>	<b>FP</b>	<b>FN</b>	<b>Sensitivity</b>	<b>CI (lower)</b>	<b>CI (upper)</b>	<b>Specificity</b>	<b>CI (lower)</b>	<b>CI (upper)</b>
Systematic review with meta-analysis	4867	7	4824	3	33	0.175	0.057	0.293	0.999	0.999	1.000
Randomized controlled trial	4867	10	4820	8	29	0.256	0.119	0.393	0.998	0.997	0.999

287

288 **Table 5.** PubMed's internal filter comparison. TP, true positive, TN, true negative, FP, false

289 positive, FN, false negative. CI, 95% confidence interval.