Training Models for Machine-Enabled Systematic Literature Reviews: Do Large Datasets Always Give Better Results?

Seye Abogunrin[1], Elena Batanova[1], Siva Karthick[2], Nicole LeDrew[3], Maria Rosario Mestre[4], Gaugarin Oliver[2], Luísa Queirós[1], Andreas Witzmann[1]

1 - F. Hoffmann-La Roche Ltd., Basel, BS, Switzerland; 2 - CapeStart, Inc, Massachusetts, United States of America; 3 - DistillerSR Inc, Ontario, Canada; 4 - DataQA, London, United Kingdom

BACKGROUND

Artificial intelligence (AI) model fitting by supervised learning suggests that larger training datasets tend to produce more accurate predictions. This can be demonstrated by evaluating the final tuned model on a labeled test dataset. However, it is not clear what the minimum number of records is for training such models and whether larger training datasets will produce significantly better results. We investigated these issues using an example of randomized controlled trial data of oncology patients.

METHODS

Data from a retrospective human-led systematic literature review (SLR) investigating the efficacy and safety benefit of second- and later lines of treatment for adults with advanced/metastatic non-small cell lung cancer were processed at title and abstract review level in three SLR tools with AI-capabilities. A summary of the AI methodology used for each tool is presented in Table 1.

Table 1 - Summary of AI methodology used in the different tools

Tool 1

Type of classifier
Pretrained model BioMed-RoBERTa

Description on how classifier works
The training dataset is split into 70% for training and 30% for validation. Then the data is tokenized using BioMed-RoBERTa. The model consisting of pre-trained RoBERTa weights with a classification head on top was fine-tuned until a high recall value for the included articles was achieved.

Tool 2

Type of classifier
Approach 50TR: Random forest
Approaches 816TR and 1631TR : BERT model

Description on how classifier works
The classification of the records into relevant/irrelevant was achieved by fitting a multiclass classifier on the whole data where the labels were the inclusion or exclusion reason. The classifier was then turned into a binary inclusion/exclusion classifier by selecting a threshold on the available training data such that the false positive rate was kept to a maximum of 5%. When applying the model on unseen data, the rows were first classified into excluded and included categories using the first-stage binary classifier.

Tool 3

Type of classifier
Complement Naïve Bayes

Description on how classifier works
This AI system was trained using tagged training sets of references that either contain the attribute(s) of interest or do not. The system allowed the tuning of scoring thresholds for inclusion and included a process for testing its classifiers against a presorted set of references to determine the recall and precision of the classifier.

For each tool, three binary classification models were trained with 50 records (Approach 50TR), 816 records, which represents approximately 10% of the dataset sample (Approach 816TR), and 1631 records, which represents approximately 20% of dataset sample (Approach 1631TR). In total 9 different models were evaluated. The records used to train were selected randomly. The same training dataset was used for the three tools. Each model classified records not used for the model training as relevant or irrelevant. The following parameters were used to compare the automatic classifications to the human classifications:

- **Confusion matrices** (not presented here) that summarize the performance of a classifier and used to derive the precision and recall values.
- **Precision**: [True Positives/(True Positives + False Positives)]; high precision suggests that the retrieved documents would be highly relevant; range 0 - 1.
- **Recall:** [True Positives/(True Positives + False Negatives)]; high recall suggests that most, if not all, relevant documents would be retrieved; range 0 - 1.
- **F1 score**: 2 * Precision * Recall /(Precision + Recall); a high F1 score suggests an acceptable balance between specificity and relevance; range 0 - 1.

		Human Classification		Totals
		Relevant	Irrelevant	
Automatic Classification	Relevant	True Positives	False Positives	Total predictive relevant labels
	Irrelevant	False Negatives	True Negatives	Total predictive irrelevant labels
Totals		Total true relevant labels	Total true irrelevant labels	Total number of documents

RESULTS

The dataset sample included 8816 records. Approach 50TR was tested in 8766 records, Approach 816TR was tested in 8000 records and Approach 1631TR was tested in 7185 records. The results of the different models were generally consistent irrespective of the sample size of the training dataset. For Tool 3, the Approach 50TR did not generalize sufficiently and its results were excluded from the analysis. See Tables 2 to 4 and Figure 1 for additional details.

Table 2 - Results for Tool 1

Approach	Recall	Precision	F1 score
Approach 50TR	0.58	0.31	0.40
Approach 816TR	0.81	0.22	0.35
Approach 1631TR	0.78	0.28	0.41

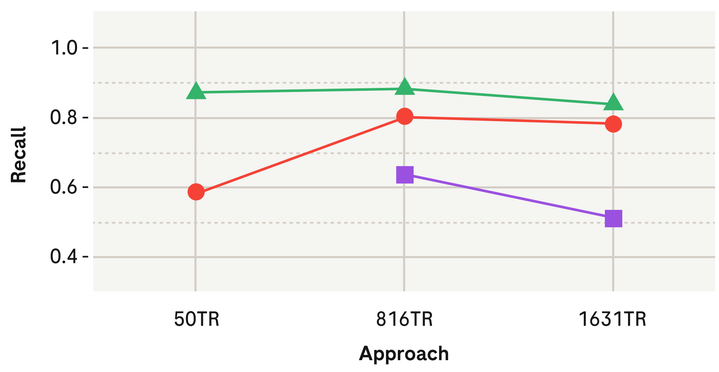
Table 3 - Results for Tool 2

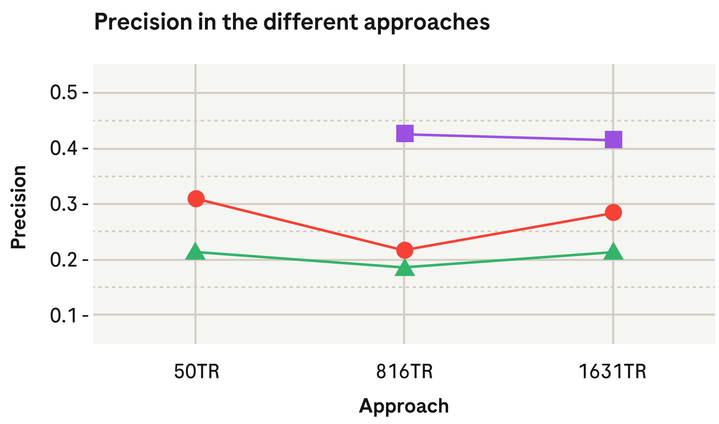
Approach	Recall	Precision	F1 score
Approach 50TR	0.87	0.22	0.35
Approach 816TR	0.88	0.18	0.30
Approach 1631TR	0.83	0.22	0.35

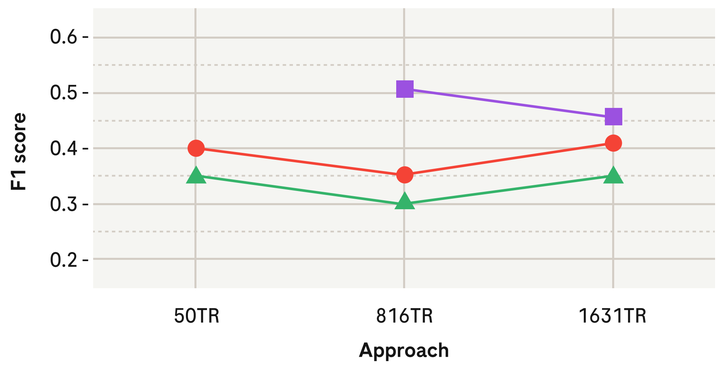
Table 4 - Results for Tool 3

Approach	Recall	Precision	F1 score
Approach 50TR	-	-	-
Approach 816TR	0.63	0.43	0.51
Approach 1631TR	0.51	0.42	0.46

Figure 1 - Summary of recall, precision and F1 score in the different approaches







DISCUSSION

- The amount of data needed to train an AI model depends on the model used and on the nature of the problem that is being solved. However, it is generally assumed that the more data used to train, the better the model performs. In addition, AI models benefit from being trained with balanced datasets.
- When it comes to **automation of SLRs** there are **different challenges** related with **preparing the AI model**:
  - Training data may be specific for each SLR and identifying the training information could be a **time consuming** task. For this reason, the more training data that is needed, the more time and human resources are required;
  - SLRs vary in their **size**. For **small SLRs** the number needed to train might represent a big percentage of the total dataset and the **automation process might become obsolete**;
  - Given the usual **low prevalence of the accepts** in SLRs, there is a threshold of maximum training data set size that can be considered in order to ensure a balance between the classification categories. In some cases, more data might not translate in better AI performance;
- Having all these in mind, our goal for this specific task of automating SLRs was to have a model that is able to return **good results with as few training records as possible**. For that reason we focused the assessment on records ranging between 50 records and 20% of the total dataset.
- Our results show that, regardless of the tool used, **there is no significant steep improvement in the results** (recall, precision and F1 score) when the number of **records used to train is increased**. Additionally, it seems that there is a threshold from which additional data does not improve the results of the algorithm (in our experiment around 10%).
- One limitation of this research is that **we did not evaluate the impact of incremental training sets** between 50 records and 10%, or between 10% and 20%. This was because we believed that due to the varying sizes of abstracts reviewed in SLRs, there may not always be certain absolute numbers for preparing a training dataset.
- In addition, **we did not evaluate the impact of using a larger proportion of data to train the models** because in a prospective setting we may need to process the complete datasets to identify a balanced training dataset comprising more than 20% of the complete datasets.
- The **records used to train were randomly selected** and we do not know if using a different selection of records would affect the performance of the models. Further research should assess how best to select training datasets.

CONCLUSION

Results were similar irrespective of the number of records used to train the AI models. Using a smaller number of training records could be advantageous in particular for SLRs based on a limited number of articles.