Automation of Title and Abstract Screening: Can Robots Replace Humans?

Abogunrin S [a], Queiros L [a], Witzmann A [a], Sumner M [b], Wehler P [b], Baehrens D [b] | [a] F. Hoffmann-La Roche Ltd., Basel Switzerland, [b] Averbis GmbH, Freiburg, Germany

BACKGROUND

- Systematic literature reviews (SLRs) are the foundation of evidence based medicine and the volume of SLRs published is increasing annually.
- Given the significant resources required to conduct an SLR, automation techniques have been explored to see how the process can be made faster and more efficient[1].
- One such technique involves support vector machines (SVMs), which use supervised learning methods for the classification of text. Previous work has shown the potential of SVMs for conducting title and abstract screening (TIABS) in select clinical oncology research. However, questions remain on the reproducibility of this method for other types of SLRs[2,3].
- Therefore, we assessed the use of SVMs for automating TIABS in various therapeutic areas and review types.

METHODS

- Ten previously completed human-performed SLRs spanning various therapeutic areas were identified. A description of the eligibility criteria for each of them and the types of SLRs covered is presented in Table 1.

Table 1 - Summary of the research topics considered

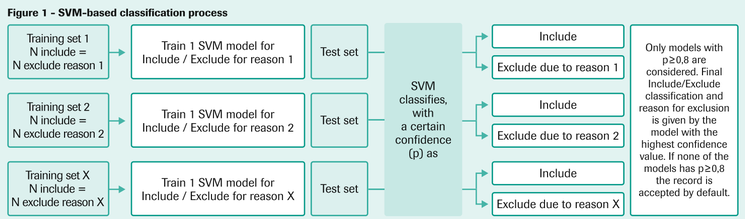
ID	Therapeutic area	Summary of eligibility criteria	Exclusion reasons considered *
Clinical Reviews			
1	Oncology	**P** Adults with advanced/metastatic NSCLC, receiving second- or later lines of treatments **IC** Chemo/immunotherapy, BSC, placebo **O** Efficacy, HRQL and safety **S** RCTs	Wrong Population Wrong Intervention Wrong Outcome Wrong Publication type Wrong Study design
2	Oncology	**P** Adults with metastatic CRPC **IC** Any pharmacological intervention or radiotherapy intervention, placebo, BSC **O** Efficacy, HRQL and safety **S** RCTs, other interventional trials	Animal/In vitro studies Wrong Disease Wrong Publication Type Wrong Study Design
3	Oncology	**P** Adults with resectable early stage NSCLC (stage 1–3B) **IC** Any pharmacological intervention and radiotherapy delivered sequentially in the adjuvant setting, BSC, placebo **O** Efficacy, HRQL and safety **S** RCTs	Wrong Population Wrong Intervention Wrong Outcome Wrong Study Design
4	Infectious diseases	**P** Adults and children with COVID-19 **IC** Any pharmacological treatments **O** Efficacy/effectiveness and safety **S** RCTs, other interventional trials, observational studies	Wrong Population Wrong Intervention Wrong Outcome Wrong Study Design
5	Haematology	**P** Adult patients with R/R DLBCL who are receiving second or third-line (or beyond) therapy **IC** Any pharmaceutical treatment **O** Efficacy/effectiveness, HRQL and safety **S** RCTs, other interventional trials, observation studies	Wrong Population Wrong Intervention Wrong Outcome Wrong Study Design
6	Oncology	**P** Adult patients with histologically or cytologically confirmed, previously untreated, extensive-stage SCLC **IC** Atezolizumab, Carboplatin plus etoposide, other platinum based treatments and immunotherapies **O** Efficacy, HRQL and safety **S** RCTs	Wrong Population Wrong Disease Wrong Intervention Wrong Study Design
7	Oncology	**P** Adult patients with any Stage IV SQ and/or NSQ NSCLC who have not received prior treatment for Stage IV NSCLC **IC** Any pharmacological treatment **O** Efficacy, HRQL and safety **S** RCTs	Animal/In Vitro studies Wrong Population Wrong Intervention Wrong Outcomes Case report Wrong Study design
Surogacy Reviews			
8	Oncology	**P** Adults with resectable early stage NSCLC (stage 1–3B) **IC** All treatment considered part of standard of care and/or treatment used in routine clinical practice, BSC, placebo **O** Effectiveness **S** Non-RCTs, observational studies	Wrong Population Wrong Intervention Wrong Outcome Wrong Study Design
9	Oncology	**P** Adults with resectable early stage NSCLC (stage 1–3B) **IC** All treatment considered part of standard of care and/or treatment used in routine clinical practice, BSC, placebo **O** Efficacy **S** RCTs	Wrong Population Wrong Intervention Wrong Outcome Wrong Study Design
Economic Reviews			
10	Oncology	**P** Adults with metastatic CRPC **IC** Any **O** ICER, utilities **S** Cost-effectiveness analysis, cost-utility analysis, utility studies	Animal/In vitro studies Wrong Disease Wrong Publication Type Wrong Study Design

* Note: Exclusions are not presented in any hierarchical order

Abbreviations: BSC - best supportive care; COVID-19 - coronavirus disease 2019; CRPC - castration-resistant prostate cancer; HRQL - health related quality of life; IC - intervention and comparators; ICER - incremental cost effectiveness ratio; NSCLC - non-small cell lung cancer; O - outcomes; P - population; R/R DLBCL - relapse refractory diffuse large b-cell lymphoma; RCT - randomized clinical trial; S - study design; SCLC - small cell lung cancer;

- SVMs consist of supervised machine learning algorithms frequently used for classification. They divide the datasets into classes by determining a visual linear separation (hyperplane).
- The classification of documents with a SVM algorithm consist of the following steps:
  - **Converting the text into geometrical points**
  - **Training the model** to recognise records that should be accepted or rejected with selected training data
  - **Determining the hyperplane**
  - **Populating with the test dataset**
  - **Classifying the test dataset** using the hyperplane determined during the training phase
- By constructing more than one SVM and applying advanced analytical methods, data can also be classified into three or more categories simultaneously. An example of this is the classification into different exclusion reasons.
- In the experiment presented, for every SLR, multiple SVMs were independently trained following which they were used to assign an include or exclude status plus exclusion reason to each record. The number of models used per SLR equaled the number of exclusion reasons defined for that SLR.
- A subset of the human classifications was used to train the automatic classifiers. Each model was trained using an evenly distributed dataset for each class considered. Generally, a set of 20 or 40 records per class (include/exclude) was used depending on the size of the data set and the prevalence of accepts in the original data set. For the different questions, the number of exclusion reasons considered varied.
- For each classification a confidence estimate ranging between 0.5 and 1 was calculated. To ensure relevant records were not missed, records with a confidence estimate of <0.8 were included by default.
- An overview of the classification process is presented in Figure 1.

Figure 1 - SVM-based classification process



Abbreviations: SVM - support vector machine; p - confidence value

- The automatic classifications were compared to the human classifications, using:
  - **Confusion matrices** that summarize the performance of a classifier. Columns represent the totals of the manual results and rows the totals of the automated results for each class.
    - **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.
- where true negatives are the number of correctly classified irrelevant records, false negatives are the number of incorrectly classified records that are relevant, true positives are the number of correctly classified relevant records, false positives are the number of incorrectly classified records that are irrelevant.

RESULTS

- The research questions included clinical and economic SLRs in oncology, infectious diseases and haematology.
- The search hits for the ten research questions ranged from 519 and 17,242, while the test dataset varied between 319 and 16,962 records.
- The recall, precision, and F1 scores for include versus exclude classification ranged between 0.90 and 1.00, 0.02 and 0.37, and 0.05 and 0.53, respectively. Details on the results obtained for each of the questions are presented in Table 2 and Table 3.

Table 2 - Results of the automatic classification Include/Exclude

ID	SLR	Recall	Precision	F1 score
1	mNSCLC (2L+)	1.00	0.07	0.13
2	mCRPC (cl)	1.00	0.03	0.06
3	eNSCLC	0.90	0.07	0.13
4	COVID-19	1.00	0.13	0.23
5	DLBCL	0.97	0.11	0.20
6	SCLC	0.99	0.02	0.05
7	mNSCLC (1L)	0.99	0.12	0.22
8	eNSCLC (non-RCT)	0.95	0.19	0.31
9	eNSCLC (RCT)	0.95	0.37	0.53
10	mCRPC (eco)	1.00	0.03	0.05

Abbreviations: cl - clinical review; COVID-19 - coronavirus disease 2019; DLBCL - diffuse large b-cell lymphoma; eco - economic review; eNSCLC - early non-small cell lung cancer; mCRPC - metastatic castration resistant prostate cancer; mNSCLC - metastatic non-small cell lung cancer; RCT randomised clinical trial; SCLC - small cell lung cancer; SVM - support vector machine; WSS@95% - work saved over sampling at 95% recall;

Legend: - maximum values; - minimum values

Table 3 - Results of the automatic classification attributing reasons for exclusion

ID	Disease	Total number of records	N records used to train	N records used to test	Excluded: True negatives	Excluded: False negatives	Included (would move to the next step): True positives	Included (would move to the next step): False positives
1	mNSCLC (2L+)	5285	80	5045	40.46%	0.02%	4.06%	55.46%
2	mCRPC (cl)	1025	40	925	31.24%	0.00%	1.95%	66.81%
3	eNSCLC	2338	80	2138	41.86%	0.47%	4.07%	53.60%
4	COVID-19	5721	80	5521	17.61%	0.04%	10.52%	71.83%
5	DLBCL	3386	80	3186	9.98%	0.31%	9.73%	79.97%
6	SCLC	10044	80	9844	46.35%	0.01%	1.34%	52.30%
7	mNSCLC (1L)	17242	82	16962	32.26%	0.12%	8.27%	59.35%
8	eNSCLC (non-RCT)	702	80	532	22.37%	0.75%	14.47%	62.41%
9	eNSCLC (RCT)	519	80	319	24.45%	1.57%	27.59%	46.39%
10	mCRPC (eco)	1126	40	926	24.30%	0.00%	2.05%	73.65%

Abbreviations: cl - clinical review; COVID-19 - coronavirus disease 2019; DLBCL - diffuse large b-cell lymphoma; eco - economic review; eNSCLC early non-small cell lung cancer; mCRPC - metastatic castration resistant prostate cancer; mNSCLC - metastatic non-small cell lung cancer; RCT randomised clinical trial; SCLC - small cell lung cancer; SVM support vector machine; WSS@95% work saved over sampling at 95% recall;

- When looking into all the exclusion reasons models, the recall, precision, and F1 scores varied between 0.00 and 0.97, 0.00 and 0.96, and 0.00 and 0.97, respectively.
- The confusion matrices for all the analyses can be found in the Appendix.
- Regarding the ability of the classifier to correctly assign exclusion reasons, from the correctly excluded records (true negatives), the percentage of correctly assigned reasons for exclusion varied between 32.73% and 87.54%. Full details presented in Table 4.

Table 4 - Assessment of reasons for exclusion attributed to the true negatives

ID	Disease	N records used to test	True negatives	Correct reason for exclusion	% Correct reason for exclusion
1	mNSCLC (2L+)	5045	2041	668	32.73%
2	mCRPC (cl)	925	289	253	87.54%
3	eNSCLC	2138	895	417	46.59%
4	COVID-19	5521	972	468	48.15%
5	DLBCL	3186	318	132	41.51%
6	SCLC	9844	4563	2435	53.36%
7	mNSCLC (1L)	16962	5472	2703	49.40%
8	eNSCLC (MPR)	532	119	74	62.18%
9	eNSCLC (non-RCT)	319	78	39	50.00%
10	mCRPC (eco)	926	225	178	79.11%

Abbreviations: cl - clinical review, eNSCLC - early non-small cell lung cancer; COVID-19 - coronavirus disease 2019; DLBCL - diffuse large b-cell lymphoma; eco - economic review; mCRPC - metastatic castration resistant prostate cancer; mNSCLC - metastatic non-small cell lung cancer; RCT randomised clinical trial; SCLC - small cell lung cancer;

DISCUSSION

- When using automatic classification, a trade-off between precision and recall is always necessary, making it challenging to achieve results with both high precision and high recall.
- During TIABS, it is important that all relevant records are retained. As such, the models used in this experiment were tuned to prioritize recall over precision during the automatic classification.
- The results suggest that this approach alone may not be able to significantly alleviate the human effort needed to complete literature reviews.
- A key limitation of this work is that the manual results against which the automatic results were compared had only one final exclusion reason stated when in fact multiple reasons could have been applicable.
- Overall, the results across the different SLRs were consistent suggesting that when used, SVM-based classifiers tend to be agnostic to the indication and type of review.

CONCLUSION

- The analysis consistently found a high recall for all investigated SLR questions, resulting in little or no relevant record being missed. However, given the observed high number of false positives, SVMs alone may not be sufficient for TIABS automation and should be investigated in combination with other artificial intelligence methods with text mining capabilities.

References:

1. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019 Jul 11;8(1):163. doi: 10.1186/s13643-019-1074-9. PMID: 31296265; PMCID: PMC6621996.
2. Queiros L, Witzmann A, Bednarski M, Sumner M, Baehrens D, Abogunrin S. A Systematic Review of Non-Small Cell Lung Cancer Clinical Trial Literature: Robots Versus Humans. Value in Health 2020 23;2 (S677), doi: https://doi.org/10.1016/j.jval.2020.08.1662
3. Abogunrin S, Queiros L, Witzmann A, Bednarski M, Sumner M, Baehrens D. Do Machines Perform Better Than Humans at Systematic Review of Published Literature? a Case Study of Prostate Cancer Clinical Evidence . Value in Health 2020 23;2 (S404), doi: https://doi.org/10.1016/j.jval.2020.08.041