Machines As a Second Reviewer in Systematic Literature Reviews

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

BACKGROUND

- A systematic literature review (SLR) uses explicit and reproducible methods that allow the identification, selection, critical appraisal and synthesis of the evidence available to answer a specific research question[1].
- SLRs are burdensome especially when it is necessary for two reviewers to screen each record.
- In an effort to make the SLR process more efficient and less time consuming, various artificial intelligence methods such as those involving support vector machines (SVMs) have been studied for the automation of title and abstract screening (TIABS).
- We explored how efficient SVM-based classifiers could be as a second reviewer during TIABS.

METHODS

- Ten retrospective human-performed SLRs addressing different health-related problems were independently assessed.
- A summary of the research topics considered is presented in Table 1.

Table 1 - Summary of the research topics considered

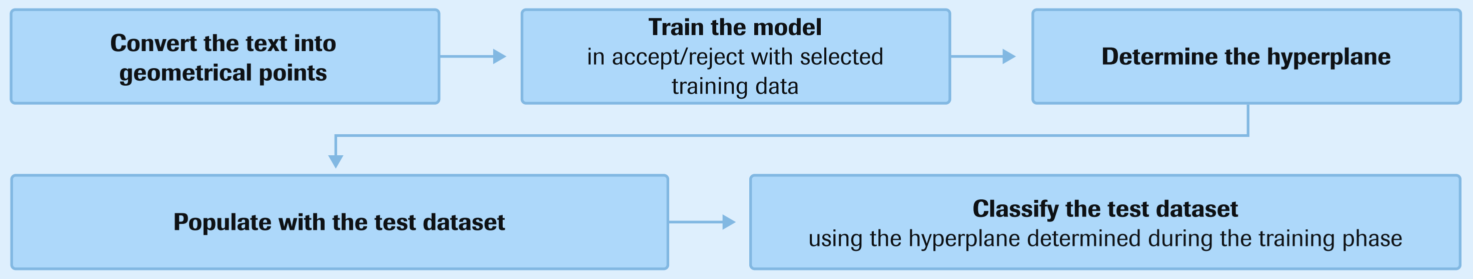
ID	Therapeutic area	Summary of eligibility criteria	Exclusion reasons considered *
Clinical Reviews			
1	Oncology	**P** Adults with advanced/metastatic NSCLC, receiving second- or later lines of treatments **IC** Chemo/immunotherapy, BSC, placebo **O** Efficacy, HRQL and safety **S** RCTs	Wrong Population 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 are based on the idea of dividing a dataset into two classes by determining a linear separation (hyperplane). In this case, SVMs were designed to create a geometrical representation of textual data (title and abstracts records), which were then assessed in relation to the hyperplane created during the training phase. The steps of the process are described in Figure 1.

Figure 1 - Steps for the creation of an SVM model



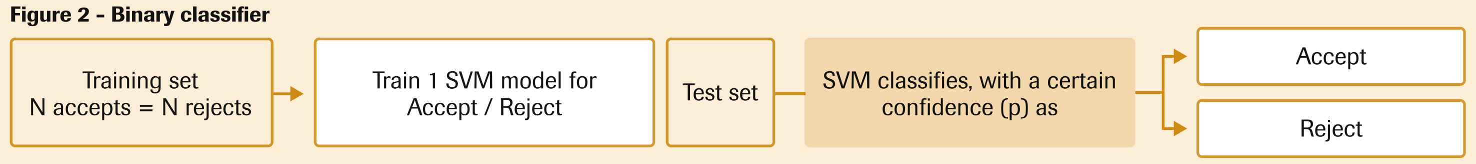
Convert the text into geometrical points → **Train the model** in accept/reject with selected training data → **Determine the hyperplane** → **Populate with the test dataset** → **Classify the test dataset** using the hyperplane determined during the training phase

- To assess how an automated reviewer could be used as second reviewer in TIABS, two SVM-based classifiers were developed for assigning accept or reject statuses to TIAB records:

1. Binary classifier

A binary model that classified the records as accept or reject and attributes a confidence value between 0,5 and 1 to each classification (Figure 2).

Figure 2 - Binary classifier



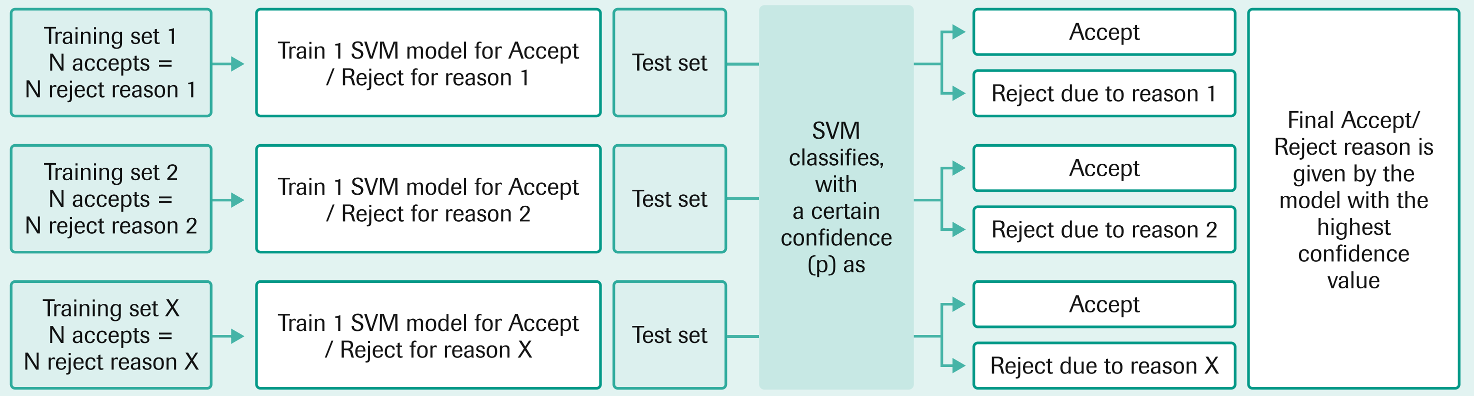
Training set N accepts = N rejects → Train 1 SVM model for Accept / Reject — Test set — SVM classifies, with a certain confidence (p) as → Accept / Reject

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

2. Ensemble classifier

This approach involved the use of multiple binary models (one for each exclusion reason considered e.g., accept - correct population, reject - wrong population). With this classifier, a record was rejected if the model with the highest confidence value labelled it as "Reject", and accepted if the model with the highest confidence value labelled it as "Accept".

Figure 3 - Ensemble classifier



Training set 1 N accepts = N reject reason 1 → Train 1 SVM model for Accept / Reject for reason 1 — Test set

Training set 2 N accepts = N reject reason 2 → Train 1 SVM model for Accept / Reject for reason 2 — Test set

Training set X N accepts = N reject reason X → Train 1 SVM model for Accept / Reject for reason X — Test set

SVM classifies, with a certain confidence (p) as → Accept / Reject due to reason 1; Accept / Reject due to reason 2; Accept / Reject due to reason X

Final Accept/ Reject reason is given by the model with the highest confidence value

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

- 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. For example, the binary classifier was trained with the same amount of accepts and rejects. Generally, a set of 20 or 40 records per class (accept/reject) 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.
- The results of the two automatic classifiers were then compared to the human results and presented 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)] and
    - **Recall** [True Positives/(True positives + False Negatives)] were also computed
    - where true negatives are the number of negative (non-relevant) abstracts correctly classified, false negatives are the number of positive (relevant) abstracts incorrectly classified as negatives (non-relevant), true positives are the number of positive (relevant) abstracts correctly classified, false positives are the number of negative (non-relevant) abstracts incorrectly classified as positives (relevant).
    - Both values vary between 0 and 1. A high precision suggests that the retrieved documents are highly relevant, while a high recall suggests that most, if not all, relevant documents were retrieved.
- **Work-saved-over-sampling at 95%-recall (WSS@95)**, defined as the percentage of papers that meet the original search criteria with a recall of 95%, was computed to determine the human effort averted when using either classifier.
- The **time to complete automated screening (TCAS)** was also calculated; this accounts for the time needed by a human reviewer to prepare the training datasets, the time to train the models (10 minutes per model) and the time spent by the automatic classifiers to process the test dataset (5 minutes per model). A human rate of 60 records reviewed per hour was used in the calculations[2].

RESULTS

- The search hits for the ten research questions ranged from 519 and 17,242. The test sample sizes varied between 319 and 16,962 records. The proportion of data used to train varied between 0.5% and 38.5% across the different questions. A detailed summary of the sample size per question is presented in Table 2 and 3 below. The therapeutic areas included were haematology, infectious diseases and oncology.

Binary classifier

- For the binary classifier, the recall, precision and WSS@95 varied between 0.53 and 1.00, 0.07 and 0.65, 0.57 and 0.81, respectively. The proportion of conflicts ranged from 12.3% to 33.1%. Details on the results obtained for each question are presented in Table 3. The confusion matrices for each of the analyses conducted can be found in the Appendix.

Ensemble classifier

- The ensemble classifier showed a recall between 0.47 and 0.95, precision between 0.17 and 0.83, and WSS between0.71 and 0.90. The proportion of conflicts ranged from 4.5% to 20.6%. Details on the results obtained for each question are presented in Table 4. The confusion matrices for each of the analyses conducted can be found in the Appendix.

Table 2 - Results of the binary classifier

ID	Disease	Total number of records	N records used to train	N records used to test	Precision	Recall	WSS@95	%conflicts (human vs SVM)	Time to complete automated Screening	Time to complete human TIABS	Δ time spent for automated vs human TIABS
1	mNSCLC (2L+)	5285	80	5045	0.13	0.90	0.67	24.4	1.6	85.4	83.8
2	mCRPC (cl)	1025	40	925	0.12	0.89	0.81	12.3	0.9	16.1	15.2
3	eNSCLC	2338	80	2138	0.10	0.82	0.59	33.1	1.6	37.0	35.4
4	COVID-19	5721	80	5521	0.20	0.72	0.58	32.7	1.6	93.4	91.8
5	DLBCL	3386	80	3186	0.24	0.53	0.71	23.0	1.6	54.4	52.9
6	SCLC	10044	80	9844	0.07	0.92	0.78	15.5	1.6	165.4	163.8
7	mNSCLC (1L)	17242	82	16962	0.17	0.89	0.63	27.0	1.6	284.1	282.5
8	eNSCLC (non-RCT)	702	80	532	0.34	0.83	0.58	26.7	1.6	10.2	8.6
9	eNSCLC (RCT)	519	80	319	0.65	0.84	0.57	17.9	1.6	6.7	5.1
10	mCRPC (eco)	1126	40	926	0.12	1.00	0.77	15.8	0.9	16.1	15.2

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; h - hours; 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; TIABS - title and abstract screening; WSS@95% - work saved over sampling at 95% recall

Table 3 - Results of the ensemble classifier

ID	Disease	Total number of records	N records used to train	N records used to test	Precision	Recall	WSS@95	%conflicts (human vs SVM)	Time to complete automated Screening	Time to complete human TIABS	Δ time spent for automated vs human TIABS
1	mNSCLC (2L+)	5285	240	5045	0.27	0.58	0.86	8,2	9.6	88.1	78.5
2	mCRPC (cl)	1025	100	925	0.25	0.67	0.90	4,5	6.2	17.1	10.9
3	eNSCLC	2338	200	2138	0.26	0.57	0.85	9,4	7.8	39.0	31.1
4	COVID-19	5721	200	5521	0.36	0.47	0.81	14,4	7.8	95.4	87.5
5	DLBCL	3386	200	3186	0.27	0.61	0.72	20,6	7.8	56.4	48.6
6	SCLC	10044	200	9844	0.17	0.84	0.88	5,9	7.8	167.4	159.6
7	mNSCLC (1L)	17242	280	16962	0.32	0.72	0.76	15,4	11.3	287.4	276.0
8	eNSCLC (non-RCT)	702	170	532	0.43	0.68	0.71	18,6	7.3	11.7	4.4
9	eNSCLC (RCT)	519	200	319	0.83	0.67	0.71	13,8	7.8	8.7	0.8
10	mCRPC (eco)	1126	200	926	0.18	0.95	0.84	9,0	7.8	18.8	10.9

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; h - hours; 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; TIABS - title and abstract screening; WSS@95% - work saved over sampling at 95% recall

DISCUSSION

- Overall, the two approaches performed well at reproducing title and abstract screening in the context of the SLR process.
- As the main goal of the analysis was to assess the use of SVMs as second reviewer in SLRs, both classifiers were refined to optimize the number of conflicts (humans vs automatic classification), and thus reduce the time needed to solve the conflicts.
- The results across the different SLRs varied but generally the ensemble classifier tended to show better results than the binary classifier, indicating that using an approach that considers reasons for exclusion may result in better performance. Further research would be needed to better understand the differences observed in the performance of each of the two classifiers across the different SLRs considered.
- The percentage of conflicts between humans and the two automatic classifiers was relatively minimal, implying that a machine could be employed as a second reviewer when reviewing TIAB records.
- The time needed to complete the automated TIAB screening is significantly lower when compared to the time needed to complete the human TIAB screening, with a difference of up to 283 hours, depending on the size of the dataset and methods used. This time difference considers the initial screening activity only and does not take into account the entire workflow including conflict resolution.

CONCLUSION

- The findings show that using SVM-based machines as a second reviewer during title and abstract screening may shorten the time to complete SLRs and thus, enable swifter decision-making. Further work should assess other automatic methods that can be used for optimal single screening of title and abstract records.

References:

1. Cochrane Library. What is a Systematic Review?. available at: https://www.cochranelibrary.com/about/about-cochrane-reviews (accessed on 2 November 2021).
2. Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH (2020) Error rates of human reviewers during abstract screening in systematic reviews. PLoS ONE 15(1): e0227742. https://doi.org/10.1371/journal.pone.0227742