Can Artificial Intelligence (AI) Replace a Human Reviewer in Systematic Literature Review (SLR)? Validation of the LiveSTART™ Tool.

Rozee (Junhan) Liu[1], Reza Jafar[2], Lee Ann Girard[3], Kristian Thorlund[4], Anna Forsythe[5]
[1]Cytel Inc., Toronto, ON, Canada, [2]Cytel Inc., Vancouver, BC, Canada, [3]Cytel Inc., Montreal, QC, Canada,
[4]McMaster University, Hamilton, ON, Canada, [5]Cytel Inc., Waltham, MA, USA

MSR74

Introduction

- Living HTA has been suggested as an innovative approach to address challenges in the current Health Technology Assessment (HTA) processes.
- One of the key challenges is how to systematically review an increasingly higher volume of evidence while ensuring unbiased and timely decisions are made for the assessment of new technologies.[1]
- It has been suggested that the HTA processes should be enhanced using technological advances. Meanwhile, the new PRISMA guidelines[2] do not prohibit the inclusion of automated tools in screening.

Objective

- **To address this challenge in HTA process and to adhere to the living HTA methodology, we developed an AI tool, LiveSTART™, utilizing transfer learning to perform the title and abstract (TiAb) review stage of a systematic literature review (SLR).**

Methods

- LiveSTART™ utilizes a biomedical language model to identify texts relevant to population, intervention/comparator, outcome, and study design (PICOS).
- Publication acceptance is then hierarchically predicted based on the given inclusion/exclusion criteria.
- LiveSTART comprises 4 functions:
  1. de-duplicate by grouping abstracts with the same or similar content;
  2. provide probability of inclusion for each PICOS criteria;
  3. predict the inclusion of each publication by comparing its abstract to the inclusion/exclusion criteria; and
  4. predict the reason of rejection based on PICOS with the pre-specified hierarchy.
- LiveSTART was trained on 59 SLR datasets with 65,328 publications, all of which were manually annotated by two independent reviewers and the discrepancies were verified by a third senior reviewer.
- **Figure 1** shows a visual illustration of the training process.
- Among the 59 datasets used for training:
  - 51 were oncology in 17 unique indications. **Figure 2** shows the distribution of these oncology indications.
  - 8 were non-oncology in 6 indications. **Figure 3** shows the distribution of the non-oncology indications.
  - 47 contained clinical datasets, while 6 were economic datasets, and 6 quality of life (QOL). **Figure 4** shows the distribution of the evidence types in these datasets.

Figure 1. Training LiveSTART with annotated SLR datasets prepared by two separate reviewers and a third independent review to reconcile their discrepancies

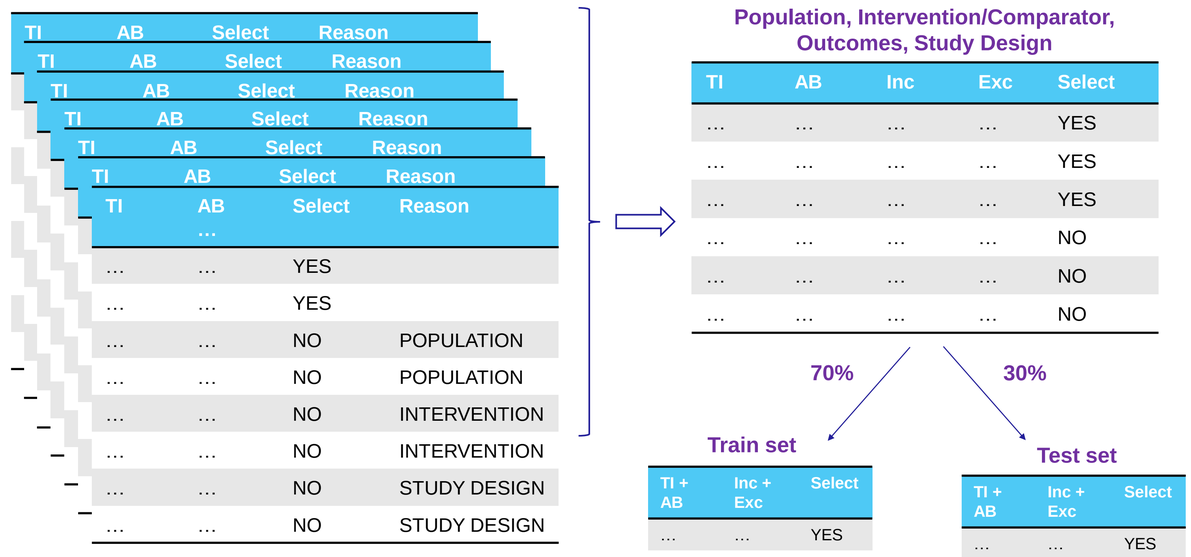


Figure 2. Oncology Indications Used to Train LiveSTART

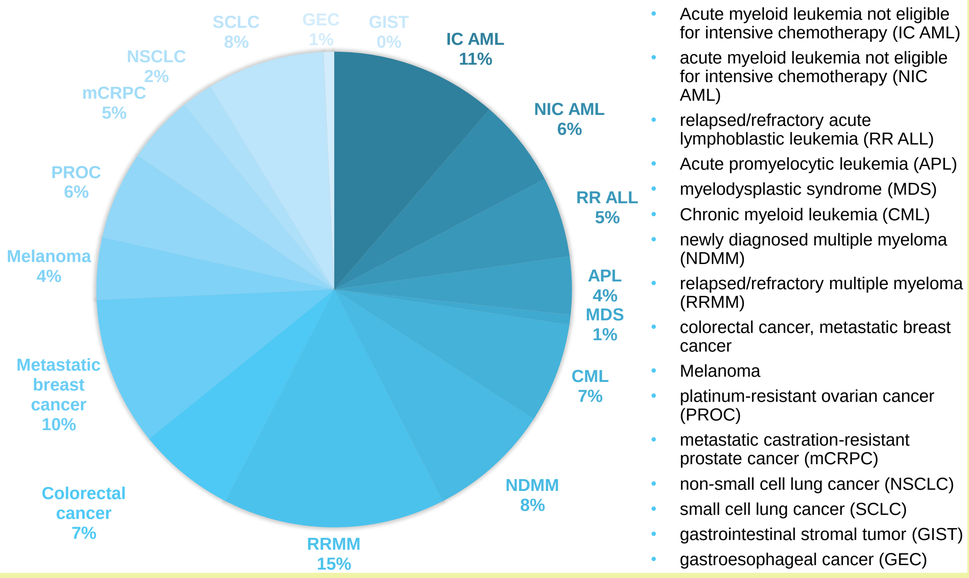


Figure 3. Non-Oncology Indications Used to Train LiveSTART

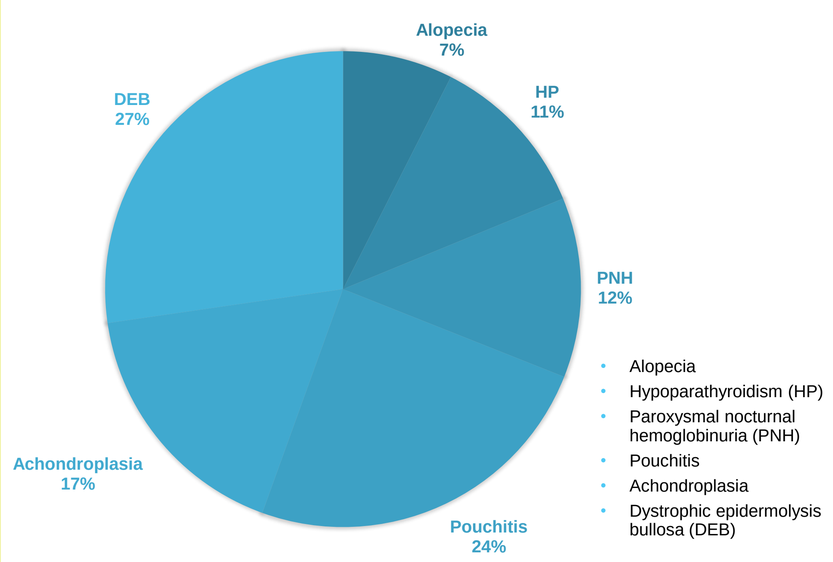
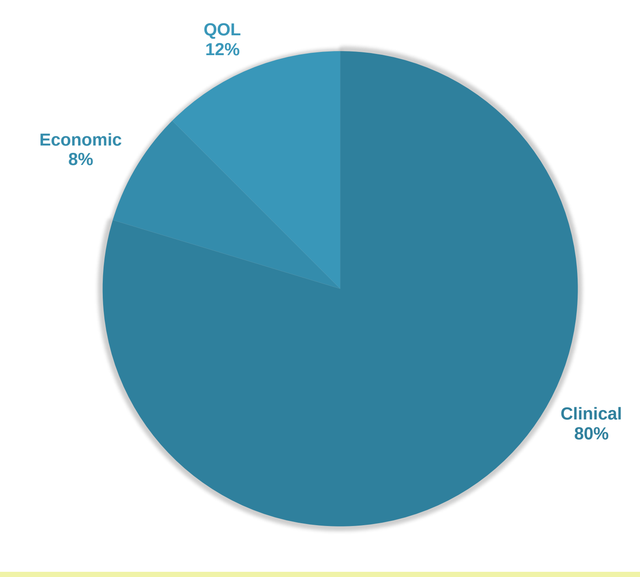


Figure 4. Types of Evidence Used to Train LiveSTART



Results

- LiveSTART validation showed an overall accuracy = 0.92, precision = 0.91, recall = 0.86, F1-score = 0.89, and area under the curve (AUC) = 0.91 when compared to the results generated by two independent reviewers and a third verifier.
- **Figure 5** shows the validation of LiveSTART by evidence type (Clinical, Economic or QOL), and indication type (Oncology vs. Non-oncology).
- LiveSTART reviews 1000 publications in ≈12.5 minutes with no additional preparation of the datasets as compared to manual review.
- An additional feature of LiveSTART is that it allows hierarchical rejection by PICOS criteria. Specifically, users can identify which PICOS criteria is higher priority. This allows traceability and flexibility of changes in SLR scope.
- LiveSTART output files are immediately ready for use in a Microsoft Excel format.
- An example of the output file is show in **Figure 6**.

Figure 5. LiveSTART Validation by Evidence or Indication Type

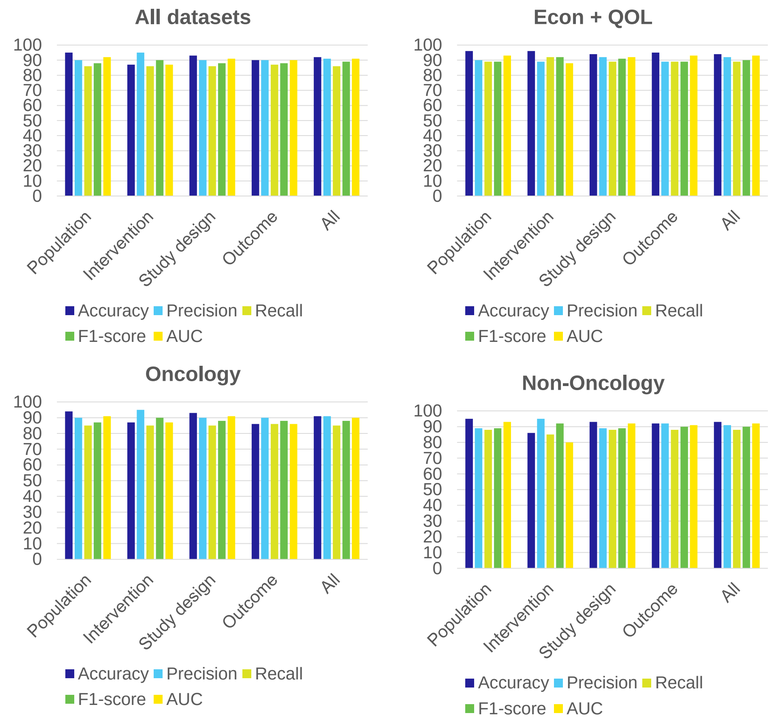


Figure 6. LiveSTART™ Output File Example

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	ORN	DB	PT	AU	SO	TI	AB	Shortened AB	Duplicate group #	Population probability	Intervention probability	Study design probability	Outcome probability	Predicted Selection	Predicted Reason for Rejection	Comment	Selection	Reason for Rejection
2	143	CCTR	Journal: Conferen	Burger JA, Sivina	Blood. Vol.130,	Randomi zed trial	Backgrou nd: Single		26	1	1	0.98	0.99	YES				
3	199	CCTR	Conferen ce	Conferen ger	Blood.	Randomi zed trial	Backgro nd:		26				.99	N				
4	140	CCTR	Jo Conferen	Smolej L, Salles G,	Vol.130, Blood.	Single-agent Si	uct ntroduct	PFS and OS outcomes	25				.04	N				
5	196	CCTR	ce Jou				u		25				.02	N				
6	139	CCTR	Co						4				.97	Y				
7	195	CCTR	ce Jou				u						.97	N				
8	138	CCTR	Co Conferen	Davids	Blood.	Initial	ct Introduct	Pts are randomized		0.98	1	0.01	0.94	NO	STUDY DESIGN			
9	194	CCTR	ce Journal:	MS, Soumerai	Conferen Blood.	results of A	ion Intro			0.98	1	0.01	0.94	NO	DUPLICATE			
10	132	CCTR	Conferen Conferen	JD, Ni A, Soumerai	Vol.130, Blood.	validated A	ion: Intro			1	1	0.04	0.96	NO	STUDY DESIG			
11	188	CCTR	ce Journal:	JD, Ni A, Lockmer	Conferen Blood.	validated Long-	ion: Backg			1	1	0.04	0.96	NO	DUPLICATE	ORN# 132		
12	131	CCTR	Conferen Conferen	S, Lockmer	Vol.130, Blood.	term Long-	nd: F Backg			0	1	0.99	0.99	NO	POPULATION			
13	187	CCTR	ce Journal:	S, Eyre TA,	Conferen Blood.	term Idelalisib-	nd: For Backgrou	Background:	21	0	1	0.99	0.99	NO	DUPLICATE	ORN# 131		

User Input

AI-Shortened Abstract
Sentences that can be used to validate the AI decision

AI prediction of probability of inclusion
1: very sure to be included
0: very sure to be excluded
Any number close to 0.5: not very sure

AI's decision of inclusion/exclusion
Reason of rejection is based on the pre-selected hierarchy
Comment includes the duplicate ID

Groups of duplicates
Citations with the same ID belong to the same duplicate groups

User's decision of inclusion/exclusion

Conclusions

- **With the combination of the unique algorithm, rigorous training on broad datasets, and highly reliable and transparent output, LiveSTART AI combined with a single reviewer could potentially yield comparable accuracy with significant time savings.**
- **However, adoption by regulatory and HTA authorities will be required.**

Limitations

- LiveSTART was primarily trained using clinical evidence in oncology indications. Although economic and QOL evidence, as well as non-oncology indications were also used to train the models, the accuracy is slightly lower for these evidence types. However, LiveSTART actively refines itself by re-training with up-to-date data, and therefore will be improved continuously.
- Currently, although the use of AI in SLRs is not specifically prohibited, it is not validated and integrated into most HTA guidelines. However, there is continued effort in validating LiveSTART and publishing the results for HTA adaptation.

rozee.liu@cytel.com

REFERENCES

1. Sarri G, Forsythe A, Elvidge J, and Dawoud D (2022). Living HTAs; How Close to Living Reality?. BMJ (In print).
2. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71

ABBREVIATIONS/GLOSSARY

AI, artificial intelligence; APL, acute promyelocytic leukemia; AUC, area under the curve; CML, chronic myeloid leukemia; DEB, dystrophic epidermolysis bullosa; GEC, gastroesophageal cancer; GIST, gastrointestinal stromal tumor; HP, Hypoparathyroidism; HTA, health technology assessment; IC AML, Acute myeloid leukemia not eligible for intensive chemotherapy; mCRPC, metastatic castration-resistant prostate cancer; MDS, myelodysplastic syndrome; NDMM newly diagnosed multiple myeloma; NIC AML, acute myeloid leukemia not eligible for intensive chemotherapy; NSCLC, non-small cell lung cancer; PICOS, population, intervention/comparator, outcome, and study design; PNH, Paroxysmal nocturnal hemoglobinuria; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PROC, platinum-resistant ovarian cancer; QOL, quality of life; RR ALL, relapsed/refractory acute lymphoblastic leukemia; RRMM, relapsed/refractory multiple myeloma; SCLC, small cell lung cancer; SLR, systematic literature review; TiAb, title and abstract