Re-Training of the Artificial Intelligence (AI) Tool LiveSTART™ with Additional Datasets: Updated Accuracy in the Title and Abstract Review Stage of Systematic Literature Reviews (SLR)

Poster #MSR33

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, US

Background

- Living health technology assessment (HTA) has been suggested as an innovative approach to address challenges in the current reimbursement processes.
- Systematic literature reviews (SLR) are labor intensive and time consuming; however, they are required for submissions to regulatory and HTA bodies.
- We previously reported a novel LiveSTART™ artificial intelligence (AI) tool utilizing transfer learning to perform the title and abstract review stage of SLR processes with reported accuracy = 0.92, precision = 0.91, recall = 0.86, F1-score = 0.89, and area under the curve (AUC) = 0.91.

Objective

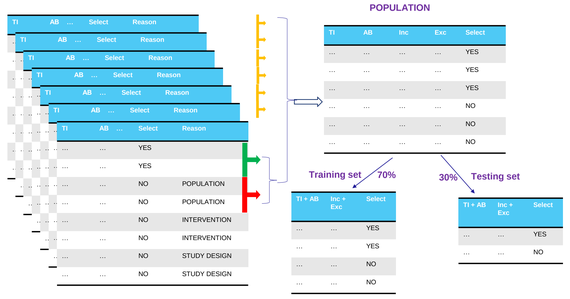
- The objective of this research was to increase confidence in the LiveSTART™ predictions by investigating whether re-training with more datasets could improve the performance of the AI tool.

Methods

- LiveSTART™ utilizes deep learning (a 12-layer neural network) to identify texts relevant to population, intervention, comparators, outcomes, and study design (PICOS) criteria, and then hierarchically predicts publication acceptance based on given inclusion/exclusion criteria.
- Fifty-nine SLR datasets with 65,328 publications were reported in November of 2022, and an additional 42 were prepared with 25,935 publications to be used toward the re-training of the existing machine learning model.
- All new datasets were manually annotated by two independent reviewers and any discrepancies were verified by a third, senior reviewer.
- A visual illustration of the training process is shown in Figure 1.

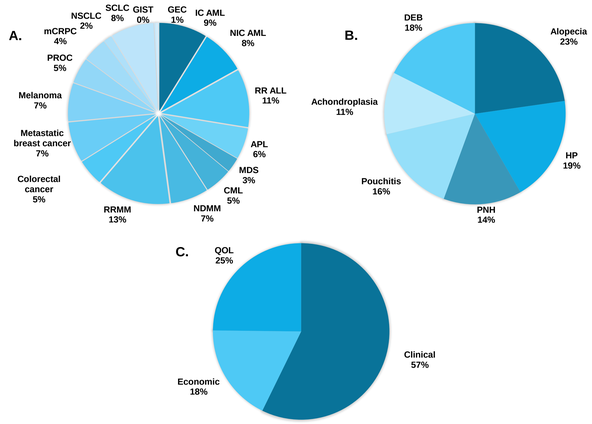
Methods (cont.)

Figure 1. Training LiveSTART™ with annotated SLR datasets



- In total, 111 annotated datasets were used to train LiveSTART™, covering 17 oncology and six non-oncology indications with 48 clinical, 32 economic, and 31 health-related quality-of-life (QoL) SLRs.
- Ninety-seven were for oncology in 17 unique indications (Figure 2A) and 14 were for non-oncology in six indications (Figure 2B).
- Regarding evidence types, 48 contained clinical datasets, 32 were economic datasets, and 31 were QoL (Figure 2C).

Figure 2. Oncology indications [A], non-oncology indications [B], and types of evidence [C] used to train LiveSTART™

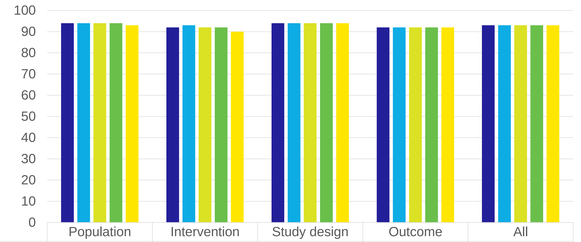


Abbreviations: (oncology indications not listed) DEB, dystrophic epidermolysis bullosa; HP, Helicobacter pylori; PNH, paroxysmal nocturnal hemoglobinuria; QoL, quality of life

Results

- After re-training with more datasets, LiveSTART™ showed improved overall accuracy = 0.93, precision = 0.93, recall = 0.93, F1-score = 0.93, and AUC = 0.93 when compared with results generated by two independent reviewers and a verifier.
- The validation of LiveSTART™ by evidence type (clinical, economic, or QoL), and indication type (oncology and non-oncology) is shown in Figure 3.
- LiveSTART™ can review 1,000 publications in approximately 12.5 minutes with no additional preparation of the datasets as compared with manual review.
- An additional feature of the new version of LiveSTART™ allows the user to customize the threshold for predictions, balancing the recall and precision of the model. The default probability threshold is set at 0.35 with recall = 0.953 and precision = 0.906. (Figure 4)

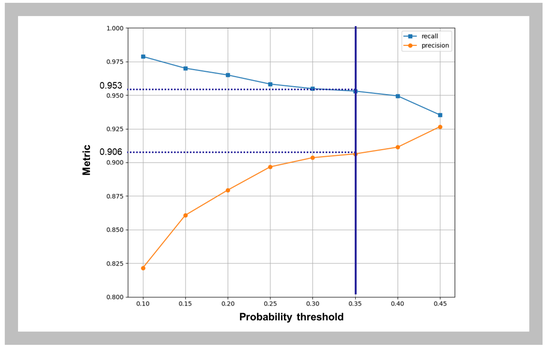
Figure 3. LiveSTART™ validation by PICOS



	Population	Intervention	Study design	Outcome	All
Accuracy	94	92	94	92	93
Precision	94	93	94	92	93
Recall	94	92	94	92	93
F1-score	94	92	94	92	93
AUC	93	90	94	92	93

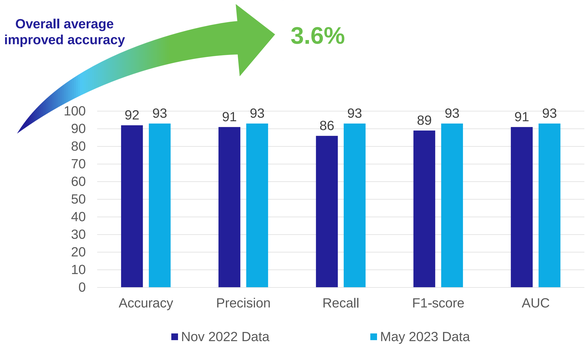
Abbreviation: AUC, area under the curve

Figure 4. Relationship between recall and precision



- Compared with the previously reported accuracy, the newly trained LiveSTART™ model has an overall average improved accuracy of 3.6%.
- Figure 5 presents the improved accuracy across all metrics.

Figure 5. LiveSTART™ re-training improvements



Limitations

- LiveSTART™ keywords for training datasets were prepared by the same analyst. There were some discrepancies in accuracy when the keywords for a project were prepared by a different person. This is currently managed by manual intervention at the keyword preparation step.
- 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.

Discussion/Conclusion

- Machine learning algorithms rely on supervised training to improve performance.
- There has been a positive correlation between the size of training datasets and performance.
- There were better predictions with a 58% increase in training data.
- LiveSTART™ AI could potentially yield significant time savings. However, buy-in from regulatory and HTA stakeholders will be required to realize this benefit.

Presented at ISPOR US | Boston, MA | May 7–10, 2023

Contact information: www.cytel.com