Re-Training of the Artificial Intelligence Tool LiveRef™: Improved Accuracy and Performance in Data Extraction

Mengmeng Zhang,[1] Reza Jafar,[2] Rozee (Junhan) Liu,[1*] Maria Rizzo,[3*] Sara Lucas,[3] Victoria Young[3]
[1]Cytel, Inc., Toronto, ON, Canada; [2]Cytel, Inc., Vancouver, BC, Canada; [3]Cytel, Inc., London, United Kingdom. *Affiliation at the time of the study.

Background

- Clinicians, pharmaceutical companies and decision-makers frequently rely on literature to gain insights on unmet needs, collect data inputs and understand current treatment options.[1-3]
- LiveRef™ is an innovative tool that leverages artificial intelligence (AI) to extract summary data from publications and offers an opportunity for stakeholders to continuously keep up-to-date on new evidence.
- Initially, LiveRef™ was trained on data that were inconsistently annotated and lacked a standardized approach to extraction. With these training data, LiveRef™ AI tool achieved an average accuracy of 56%.[4]

Objective

To improve the performance of the LiveRef™ AI tool and increase confidence in its predictions; and to enhance the ability to filter for studies with common data elements in an interactive literature review platform.

Methods

Data preparation

- A dataset of 1,000 congress abstracts and 1,000 references from Ovid searches across various indications were collected.
- Two independent and experienced reviewers manually extracted and annotated summary data for the variables included in the re-designed extraction scheme, which would typically be collected by stakeholders, plus subjective interpretation of the main message and summary of results.

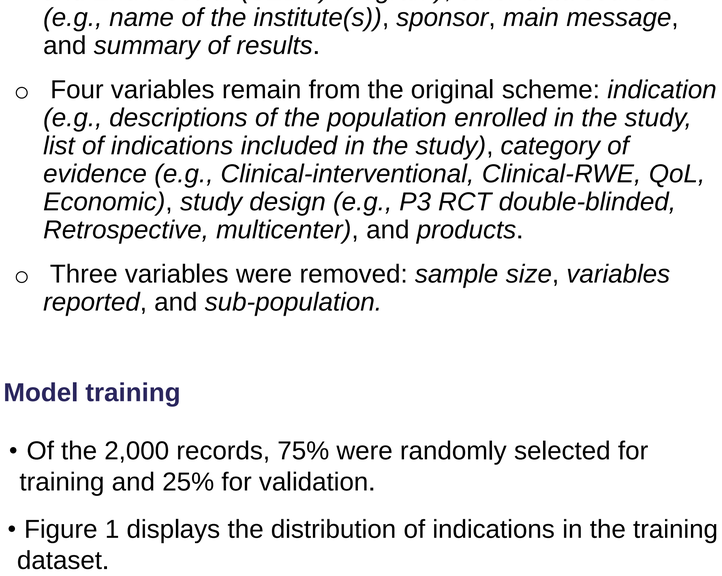
New annotation scheme

- An annotation guideline was developed to curate data according to a new annotation scheme, modified from the scheme used in the initial AI model.
- Compared to the initial AI model's annotation scheme:
  - Eight new variables were added: *regimens (e.g., monotherapy, combinations)*, *study registry name*, *study registry ID*, *database (e.g., the Surveillance, Epidemiology, and End Results (SEER) Program)*, *other data sources (e.g., name of the institute(s))*, *sponsor*, *main message*, and *summary of results*.
  - Four variables remain from the original scheme: *indication (e.g., descriptions of the population enrolled in the study, list of indications included in the study)*, *category of evidence (e.g., Clinical-interventional, Clinical-RWE, QoL, Economic)*, *study design (e.g., P3 RCT double-blinded, Retrospective, multicenter)*, and *products*.
  - Three variables were removed: *sample size*, *variables reported*, and *sub-population.*

Model training

- Of the 2,000 records, 75% were randomly selected for training and 25% for validation.
- Figure 1 displays the distribution of indications in the training dataset.
- Figure 2 depicts the distribution of evidence types in the training dataset.
- SciFive, a generative biomedical language model, was fine-tuned on the dataset to extract the summary data.

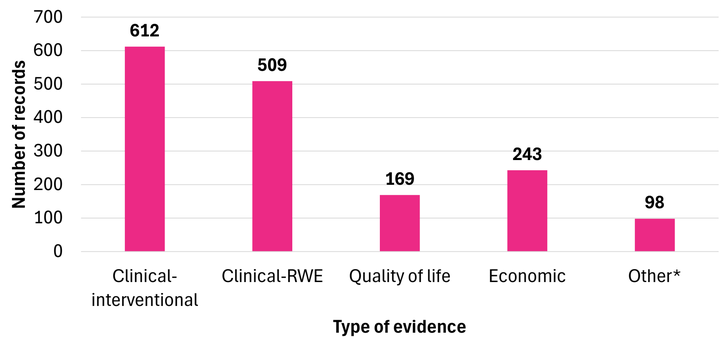
Figure 1. Indications used to train LiveRef™



*Other includes indications with a percentage of <1%, for example, colorectal cancer, head and neck cancers, gastroesophageal cancer etc. Abbreviations: AML, acute myeloid leukemia; DLBCL, diffuse large B-cell lymphoma; MCC, Merkel cell carcinoma; MDS, myelodysplastic syndrome; NMIBC, non-muscle-invasive bladder cancer; NSCLC, non-small cell lung cancer; RCC, renal cell carcinoma; SCLC, small cell lung cancer; SLE, systemic lupus erythematosus; UC, urothelial carcinoma.

Methods (cont.)

Figure 2. Type of evidence used to train LiveRef™



Some records are included in multiple categories. *Other categories of evidence include systematic literature reviews, meta-analyses, indirect treatment comparisons, HTA reports etc. Abbreviations: HTA, health technology assessment; RWE, real-world evidence.

Model validation and assessment

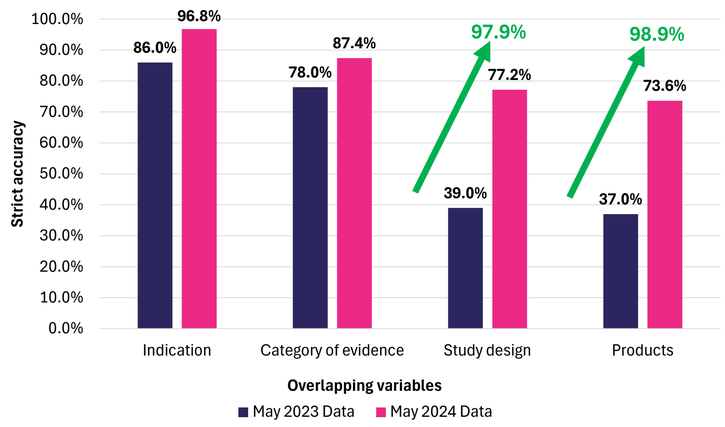
- The model accuracy was evaluated by a human reviewer who scored the model predictions on the validation set against manually extracted data using three categories: completely correct, partially correct, or incorrect.
- We defined two sets of evaluation metrics:

  1) Strict accuracy measures the proportion of predictions that are completely correct. The strict accuracy score considered partially correct predictions as incorrect.

  2) Lenient accuracy measures the proportion of predictions that are completely or partially correct. The lenient accuracy score considered partially correct predictions as correct.

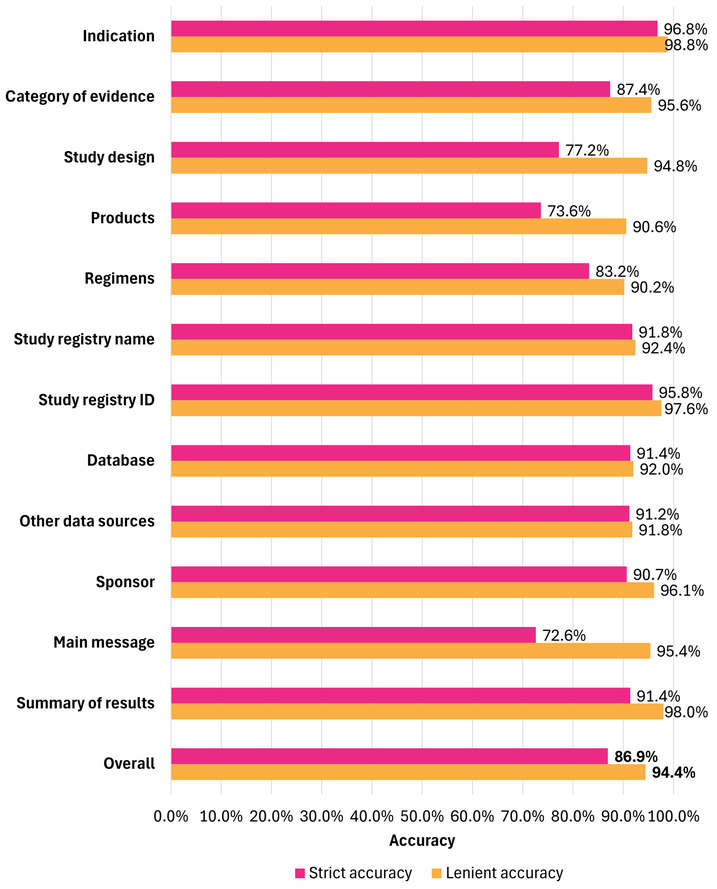
Results

Figure 3. LiveRef™ strict accuracy improvement for overlapping variables



Green arrows indicate the percentage improvement in accuracy between the original training and re-training.

Figure 4. LiveRef™ retraining strict and lenient accuracy for all variables

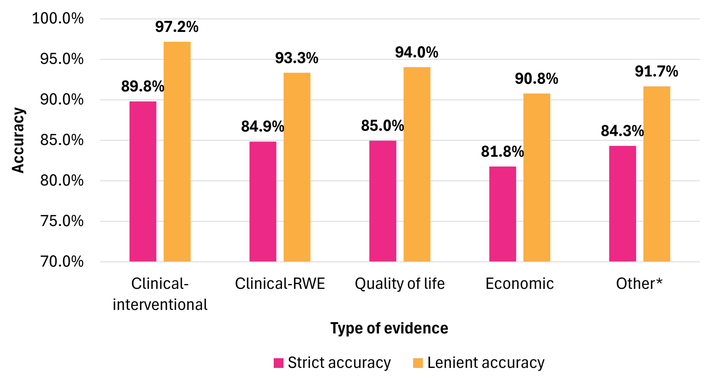


Abbreviations: ID, identifier.

Results (cont.)

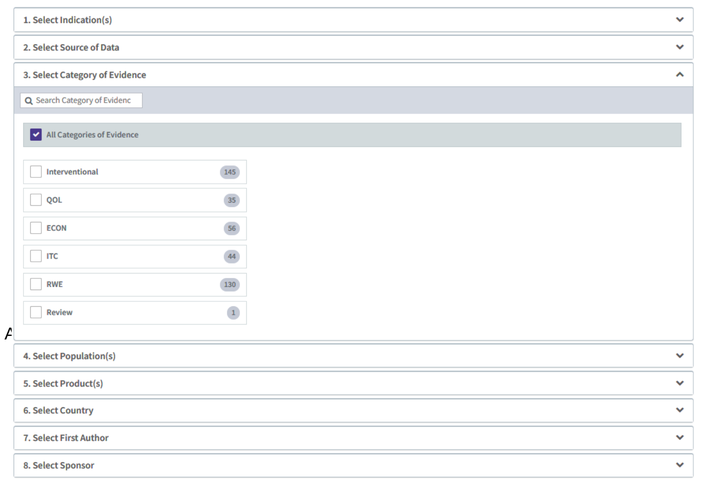
- The updated LiveRef™ model showed an increase of 12.6%, 12.1%, 97.9%, and 98.9% in strict accuracy of predictions for indication, category of evidence, study design, and products, respectively, versus the original results published in May 2023.
- Figure 3 demonstrates the improved strict accuracy.
- The three variables with the highest strict accuracy were indication, study registry ID, and study registry name
- The three variables with the highest lenient accuracy are indication, study registry ID, and summary of results.
- Figure 4 presents the LiveRef™ AI-assisted extraction accuracy for all variables in the new annotation scheme.
- LiveRef™ demonstrated excellent grammar, syntax, and logical editing for subjective interpretations.
- Figure 5 presents the overall accuracy of LiveRef™ extraction for different types of evidence.
- The strict accuracy and the lenient accuracy of predictions for interventional studies were the highest (89.8% and 97.2%) and the strict accuracy for all categories of evidence was 81.8% or higher.

Figure 5. LiveRef™ retraining accuracy for different types of evidence



*Other categories of evidence include systematic literature reviews, meta-analyses, indirect treatment comparisons, HTA reports etc. Abbreviations: HTA, health technology assessment; RWE, real-world evidence.

Figure 6. Filters of an example project on LiveRef™



ITC: indirect treatment comparison; ECON: economic; QOL: quality of life; RWE: real-word evidence.

- Figure 6 shows the filters for category of evidence, population, products, country, first author and sponsor of an example project on the LiveRef™ platform.

Conclusions

- The accuracy and performance of the LiveRef™ AI tool were substantially improved through controlled data collection and annotation, and supervised training of a biomedical language model.
- This improvement could allow us to create more structured and standardized data sets, improve filtering for studies interactively and yield time and resource savings.

References

1. National Institute for Health and Care Excellence. https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-technology-appraisal-guidance. Accessed Oct 29, 2024.
2. Pharmaceutical Benefits Advisory Committee. https://pbac.pbs.gov.au/ Accessed Oct 29, 2024.
3. Agoritsas T, et al. Finding Current Best Evidence. In: Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 3rd ed. McGraw-Hill Education; 2015.
4. Liu J, et al. Value in Health, 2023;26(6):S2,S279.

Disclosures

Conflict of interest: Mengmeng Zhang, Reza Jafar, Rozee (Junhan) Liu, Maria Rizzo, Sara Lucas, and Victoria Young were employees of Cytel Inc at the time of the study.

Funding Information: This study was funded by Cytel Inc.

Cytel