Repeatable Auto-extraction Frameworks in Clinical Systematic Literature Review: Validating a Multi-Model Human-in-the-Loop Artificial Intelligence system for Extracting Study PICOs, Location, Size, and Type

P22

Joshua Twaites, MS, Kevin Kallmes, BS, MA, JD, Karl Holub, BS;

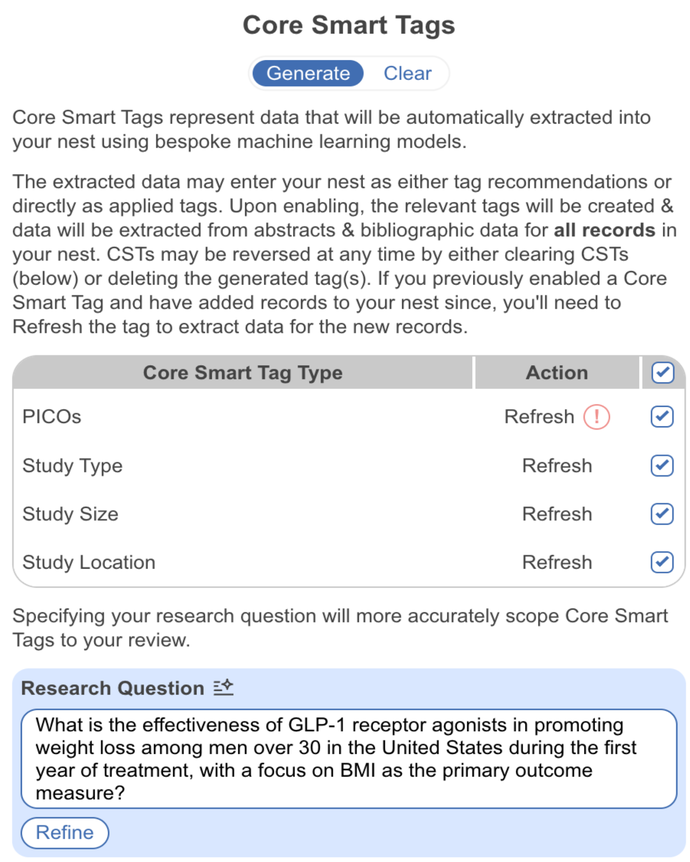
Nested Knowledge, Inc., St. Paul, MN, USA.

Background

Clinical systematic literature reviews (SLRs) are often framed around the Population, Interventions/Comparators, and Outcomes (PICOs) and basic designs of underlying studies. While Large language models (LLMs) have been tested for extraction of data based on user queries, research is lacking on the ability of artificial intelligence (AI) systems for building repeatable extraction structures for Clinical SLR. Specifically, we hypothesized that human-in-the-loop machine learning, natural language processing (NLP), and heuristics can provide reliable extraction without hallucination risk. We built and tested specialized, multi-model AI tools for both extracting and building hierarchies from key study elements, specifically PICOs, study type, location, and size.

Methods

We built 'Core Smart Tags' (CSTs), an integrated system employing machine-learning and heuristic-driven models to extract study type, location, and size from study abstracts and metadata, and integrated an existing NLP model to extract and structure PICOs hierarchically. We tested each element against existing gold standard datasets. For PICOs, the underlying model was tested against an open-source EBM-NLP dataset; for study location, we used ClinicalTrials.gov study locations from NCT-linked studies; for study type, we hand-labelled 1,000 studies; for study size, we tested against the PICO Corpus dataset.

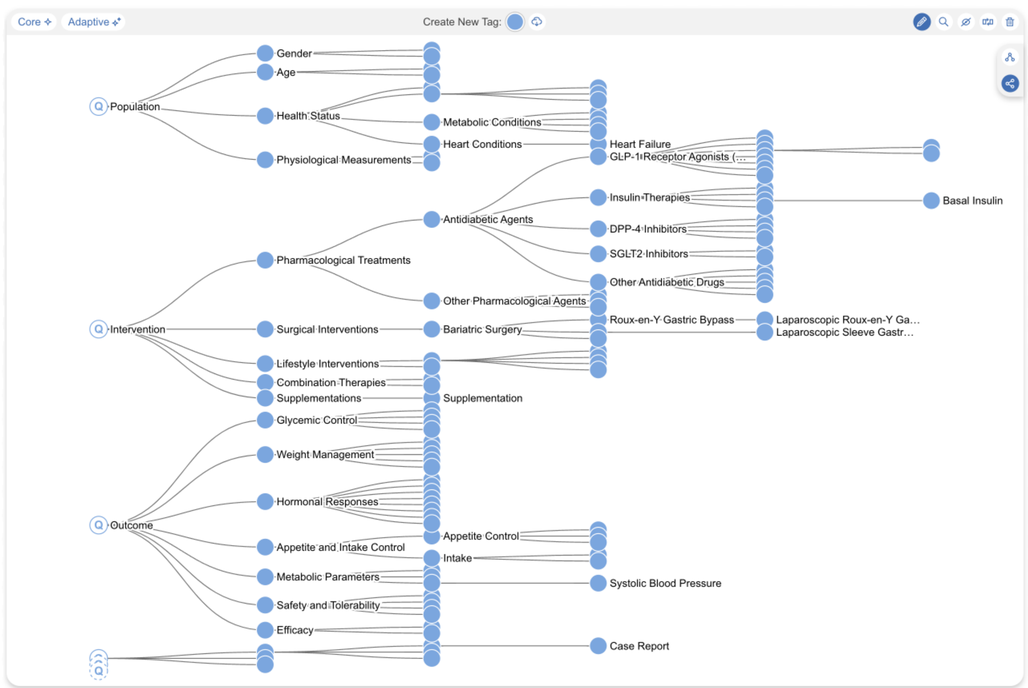


Results

In PICOs extraction, the model underlying CSTs achieved an F1 score of 0.74. In predicting study type, CSTs had overall F1 of 0.74, overall accuracy of 74%, and achieved 0.96 Recall for finding randomized controlled trials. In predicting location, CSTs had 78% accuracy, Recall of 0.79, and Precision of 0.90. In study size, CSTs had 91% accuracy.

Conclusions

When extracting evidence in a replicable extraction framework, specialized AI systems can find and structure elements with reasonable accuracy, even from abstracts alone. Furthermore, human-in-the-loop systems enable expert curation of the outputs from these AI tools, enabling faster, AI-informed structuring and execution of clinical SLRs.



{NESTED} KNOWLEDGE®