Speeding up literature search for diversity action plans with AI

Marco Virgolin [1], Valerio Valente [1], Giuseppe Pasculli [1], Daniel Roeshammar [1], Sofia Stathopoulos [1]

[1] InSilicoTrials Technologies S.p.A, Riva Grumula 2, 34123, Trieste, Italy

INTRODUCTION

Underrepresentation in clinical trials poses significant risks, including insufficient safety and effectiveness of drugs for minority groups. FDA's Diversity Action Plans (DAPs) aim to address this by setting well-motivated enrollment goals stratified by age, sex, race, and ethnicity (FDA, 2024).
An important part of setting enrollment goals consists of reviewing scientific literature on incidence/prevalence of a disease in different subpopulations. This step requires expert personnel and dozens of hours. We introduce a novel, AI-based tool leveraging large language models (LLMs) to automatically scan and extract relevant epidemiological data.

METHODOLOGY

The steps performed by our tool are as follows.

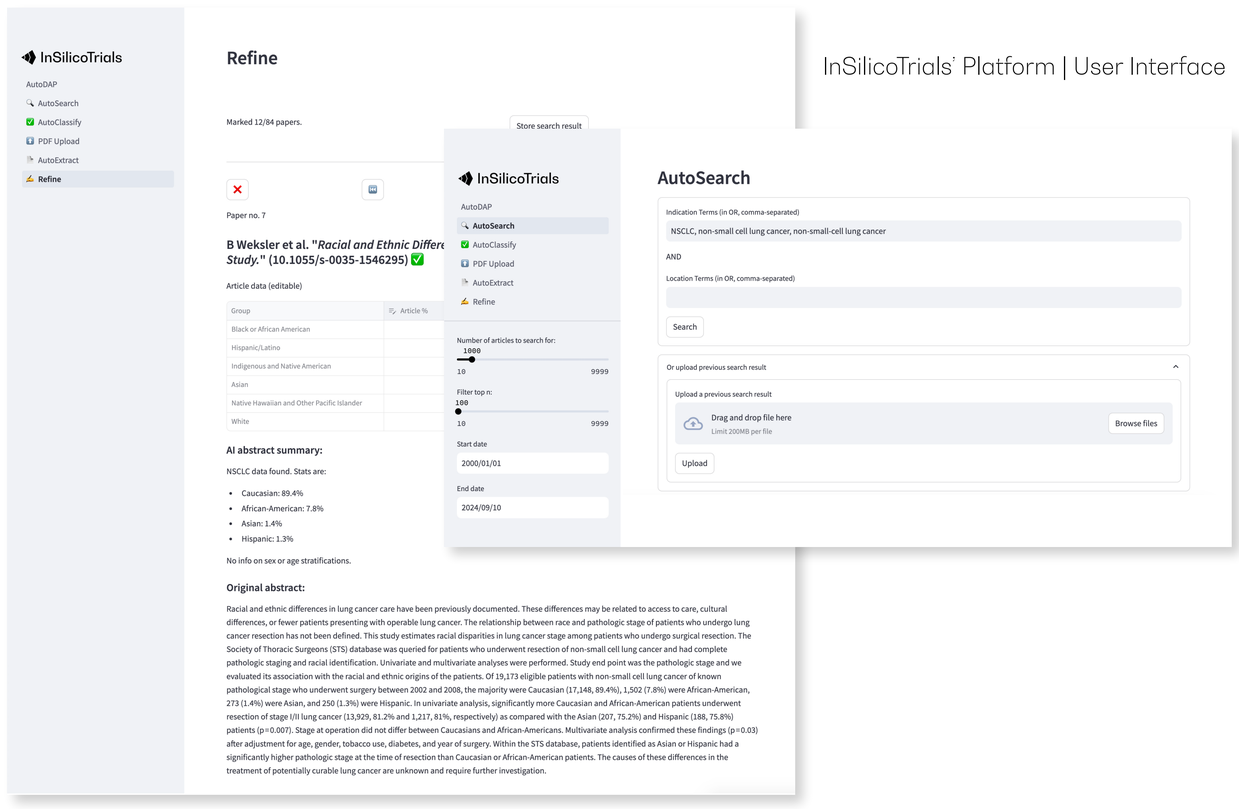
01 **Obtain a large (hundreds/thousands) collection of paper titles and abstracts for a given indication (e.g., "breast cancer"). This can be executed quickly using APIs like Pubmed's.**
No. papers: 100-10,000 Time: seconds

02 **Rank the abstracts to prioritize those that include numerical results on incidence/prevalence, then filter to keep top N. Here, an embedding model converts text to vectors, and cosine similarity ranks these vectors against a pre-defined vector. The latter is pre-optimized to represent abstracts with data on incidence/prevalence in different groups.**
No. papers: 100-1000 Time: seconds

03 **Classify the top abstracts as relevant (Y/N) using a relatively fast LLM.**
No. papers: 100-500 Time: minutes

04 **For the relevant abstracts, attempt to extract numerical information about incidence/prevalence with a more powerful LLM.**
No. extractions: 10-100 Time: minutes-hours

The tool can further work at the level of a paper's PDF, making a summary of the results contained therein and extracting tables.



InSilicoTrials' Platform | User Interface

RESULTS

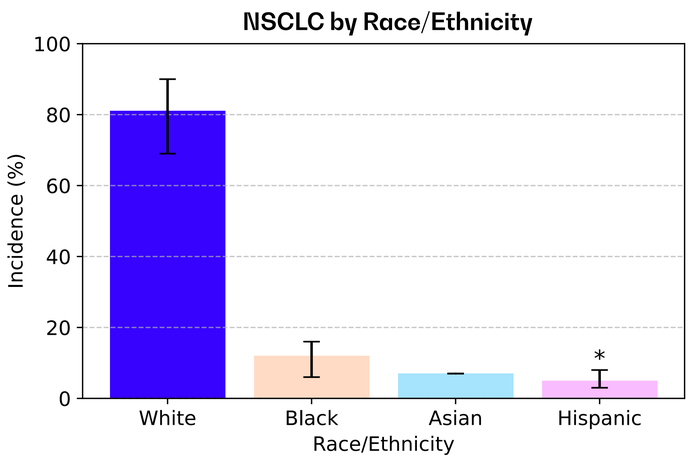
We compared the extractions of the tool against a manually-crafted DAP concerning non-small-cell lung cancer (NSCLC), focusing on race and ethnicity (Hispanic: yes/no). For confidentiality reasons, we cannot provide detailed information.

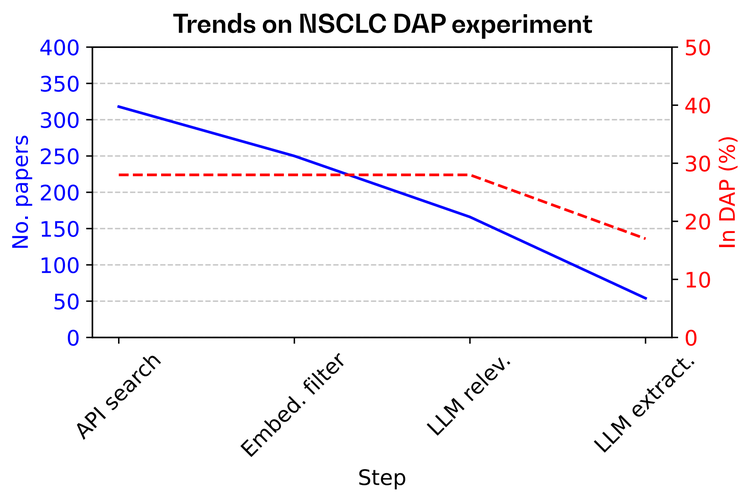
01 **Step 1, searching with PubMed API within 2000/1/1 to 2024/8/6 and "non-small cell lung cancer, non-small-cell lung cancer, NSCLC, lung cancer" as keywords, resulted in 318 papers. Of these, 28% matched those cited in the DAP. Typical missed papers had relatively broad abstracts and titles (e.g., "Key Statistics for Lung Cancer").**

02 **Filtering with step 2, with values of N=1000,500,250,100, respectively led to 28%,28%,28%,24% matches remaining.**

03 **The list of top 250 was submitted to step 3. The LLM took ~2 seconds per abstract and classified 166 papers as relevant, retaining the 28% matching papers.**

04 **Step 4, with a more precise LLM taking ~10s per abstract, extracted incidence/prevalence information out of 54 papers. Finally, the extractions were analyzed manually, in ~10 minutes.**





Incidence of NSCLC from extractions (min-max, mean %) were as follows:
White: 69-90,81; Black: 6-16,12; Asian: 7-7,7; Hispanic: 3-8,5. All found means of minorities were within the ranges expressed in the DAP, except for Hispanic, which was too low by 1%.

Finally, we also tested the tool extracting tables out of 4 random papers: 3/3, 3/3, 2/2, 2/2 tables were respectively extracted.

CONCLUSION

We presented a fast, scalable, data-driven solution for literature search in diversity action plans. In our experiment, extracted disease incidence found in minutes of human intervention were comparable with those found in days. Our tool can help professionals be more efficient and effective in defining diversity-aware enrollment goals, ultimately promoting health equity.

REFERENCE

FDA (2024) 'Diversity Action Plans to Improve Enrollment of Participants from Underrepresented Populations in Clinical Studies: Draft Guidance for Industry'. Available at: https://www.fda.gov.

DISCOVER

Scan to discover all of InSilicoTrials' solutions for pharma.