1503-Development and Validation of a Living Decision Support Tool (Living-DST) in Oncology Using Agentic AI-Augmented Systematic Literature Review (SLR)

Lee Schwartzberg 1, Anna Forsythe 2, Douglas Flora 3, Stefan Glück 4, Stacy Grieve 5, Rhiannon Campden 5, Rozee Liu 5, Jessicca Martin Rege 2, Peter A. Kaufman 6

1. Renown Health-William N. Pennington Cancer Institute, Reno, NV, USA 2. Oncoscope-AI, Miami, FL, USA. 3. Yung Family Cancer Center, St. Elizabeth Healthcare, Edgewood, KY, USA 4. Consultant Oncology Immunotherapy Biotech and AI, Miami Beach, FL, USA 5. Eviviz, Vancouver, BC, Canada. 6. The University of Vermont Medical Center, Burlington, VT, USA

Correspondence: anna.forsythe@oncoscope-ai.com. Disclosure statement: the authors of this publication have no conflict of interest to disclose.

PS3-04-10

INTRODUCTION

- Oncologists face growing difficulty in staying current with rapidly evolving data from congresses, journals, guidelines, and regulatory updates.
- Creating a well-organized, annotated clinical trial (CT) data library in an easily digestible format is time- and labor-intensive.
- To address this, we developed and validated a living-DST for breast cancer (BC), powered by an agentic AI system that supports daily human conducted / AI augmented SLR.

METHODS

Agentic AI training and accuracy validation

- An **agentic AI system** using GPT-4.1 & o3 (OpenAI), and Claude Sonnet-4 (Anthropic) was designed to emulate expert-led Cochrane-compliant[1] SLR processes,
- The system follows an **annotation manual**, constructed by human scientists manually annotating **29,236** clinical trial abstracts across **BC**, prostate cancer and non-small cell lung cancer.
- Accuracy was **assessed vs. 1,997 human annotations** as validation set.

Living DST development

- Each study was annotated with **4 review variables** (population, intervention/comparator, reported outcomes, study design) to determine the inclusion into the database that is the basis of the living DST.
- **32 data points** were extracted per included study on **patient and study characteristics** (TNM staging, histology, biomarker, risk factors, treatment line, interventions randomization, study phase and sample size, analysis type, follow-up period) and **reported outcomes** (median and landmark overall survival (OS), progression-free survival (PFS) other progression measures, response data, quality of life), subgroup analyses, safety/toxicity).
- AI-reviewed and -extracted data were **100% quality checked** and standardized by experienced human scientists and integrated with established guidelines (US and Europe) treatment paradigm.
- A **user interface - Oncoscope AI** (previously validated and published[3], *Figure 1*) - was built to filter this database, linking treatment pathways to real-time SLR, FDA approvals, and ongoing clinical trials.

Validation against chatbots

- DST evidence precision was evaluated against the AI chatbots: **ChatGPT**, **Perplexity**, **Consensus** and **OpenEvidence** using 6 criteria for 8 cases on treatment options in BC with different treatment path, subpopulation, and biomarker variables.
- Unlike Oncoscope-AI which employs human scientists to use AI to curate massive data faster, with guardrails like 100% quality-check, and publications grouped into trials, chatbots uses varying degrees of evidence curation that is much less robust, systematic and comprehensive. Some commonly used chatbots in the healthcare space are: ChatGPT, Perplexity AI, Consensus, and OpenEvidence. Highlights of these chatbots are provided in **Table 1**.
- Accuracy for each criterion was defined as the percentage of questions evaluated as positive over all 8 questions.

METHODS (CONTINUED)

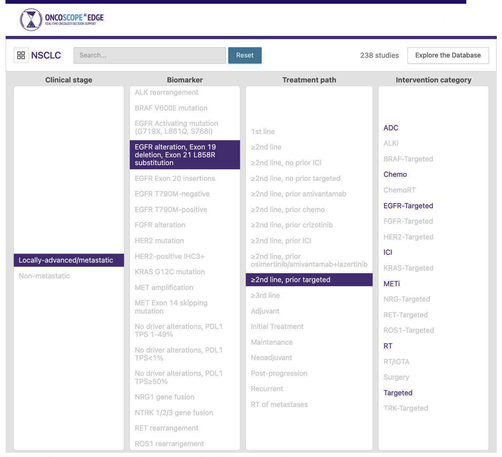


Figure 1. Screenshot of the Oncoscope-AI validated tool filtering process.

Table 1. Summary of the chatbots: ChatGPT, Perplexity, Consensus, OpenEvidence.

Tool	Knowledge Base	Source Currency	Content Generation Style	Intended Users / Popularity
ChatGPT	Large language model (GPT) trained on broad, static datasets up to its latest cutoff; custom GPTs allow user-added data, but base model itself is not web-connected by default [4]	Static (training data up to 2022–2023; some modes enable limited browsing plugins)	Highly creative, conversational, supports code, stories, essays, brainstorming as well as factual Q&A [4 5]	General public, students, creators, businesses; extremely popular with 100M+ users, especially for creative/content generation, chat, and productivity
Perplexity	Aggregates real-time web data for each query via multiple LLMs (GPT-4o, Claude 3, etc.); integrates web search and document upload; built for research and up-to-date info [6 7 8]	Real-time search (retrieves current sources; always cites them)	Primarily factual, concise, evidence-based with source citations; supports creativity if prompted, but less writing finesse than ChatGPT [6 7 8]	Researchers, professionals, students, marketers, analysts, B2B and B2C; especially popular with educated, affluent users in research, business, and education [6 7 8]
Consensus	Focused on scientific, academic, and credible sources; aggregates and analyzes studies, journals, news; expert-area specialization, especially for research and evidence synthesis [9 10]	Real-time analysis of latest studies, research, news—updates as new sources appear	Factual, research-driven, synthesis-style reports; emphasizes data integration over open-ended creativity [10]	Researchers, analysts, academics, professionals needing insights from scientific literature and evidence-based synthesis [10]
Open Evidence	Specialized, proprietary ensemble of models trained exclusively on peer-reviewed medical literature; not general-purpose; optimized for clinical accuracy and transparency [11 12 13]	Real-time access to latest medical literature; continuous ingestion of new peer-reviewed studies	Highly factual, evidence-based, structured syntheses tailored for clinicians; references every claim	Physicians and clinicians, medical professionals in the US; extremely popular and rapidly adopted in hospitals and clinics for clinical decision-making [12]

RESULTS

- Accuracy across review variables ranged 93.12–97.1%. For extraction variables, accuracy exceeded 90% for all variables (92%–100%) with 50% of variables above 95%.
- Agentic AI annotated 1,997 publications in 7.25 hours vs. an estimated 727.4 hours required by human experts—yielding 99% of time savings.
- In all 8 test cases, our DST met the 6 criteria 100%, outperforming ChatGPT (12.67%), Perplexity (12.50%), Consensus (21.00%), OpenEvidence (14.83%) by providing more complete and accurate evidence, retrieving recently published, potentially practice-changing pivotal trials, and linking to original citations and regulatory labels where available.

Table 2. Living DST and chatbots performance analysis and comparison

Assessment criteria	Living-DST (Oncoscope AI)	ChatGPT	Perplexity	Consensus	OpenEvidence
Comprehensive Treatment Listings	100%* All guideline-backed and practice-changing treatment options, including recent approvals and off-guideline evidence. *Although all treatment options were provided, some tumor agnostic trials were missed and later added.	38% Consistently missed important practice-changing treatment options like ADCs.	25% Provided mostly only classes of drugs without individual options.	50% Consistently missed important practice-changing treatment options like ADCs but performed better than ChatGPT.	13% Consistently missed both practice-changing treatment options, as well as on-guideline treatment options.
Guideline Integration and Evidence Grading	100% Each treatment is linked to current national and international guidelines. The strength of evidence critical for informed choices are provided.	0% Did not mention guideline recommendation status.	0% Did not mention guideline recommendation status.	0% Did not mention guideline recommendation status.	13% Consistently mentioned some ASCO guidelines, but some are either inaccurate or irrelevant to the questions asked.
Included treatment with positive results in trials but not yet in guidelines?	100% All prospective clinical trials with published data are included with data presented and linked back to clinical trial registry.	38% Mentioned trials with results but no details on trial name or registration.	25% Mentioned a few treatments under investigation, but with no data or trial details.	38% Mentioned a few treatments under investigation, but with no data or trial details.	13% Only mentioned one trial in one test scenario, but no data or trial details were provided.
Accurate Efficacy Data	100% Recommendations are accompanied by pivotal trial outcomes (e.g., PFS, OS, ORR), not just observational or real-world data.	0% Did not mention any efficacy data.	25% Mentioned some efficacy data, but the data provided were incomplete (e.g. only PFS, no OS), or inaccurate (e.g. PFS and OS from real-world data instead of clinical trials)	38% Mentioned some efficacy data, but the data provided were incomplete (e.g. only show comparison between treatment arms, not individual arm data).	0% Did not mention any efficacy data.
Transparent Source Referencing	100% Provides ability to trace every recommendation to original clinical trials.	0% Did not provide source to evidence.	0% Did not provide source to the primary publication of the clinical trial.	0% Did not provide source to the primary publication of the clinical trial.	50% Provided primary publication references to some trials only.
Provided links/reference to FDA approval	100% Every regulatory approved treatment option has a direct link to its most updated FDA label.	0% Did not provide link to FDA approval or label.	0% Did not provide link to FDA approval or label.	0% Did not provide link to FDA approval or label.	0% Did not provide link to FDA approval or label.
Average	**100%**	**12.67%**	**12.50%**	**21.00%**	**14.83%**

DISCUSSION & CONCLUSION

By integrating human-conducted, AI-augmented daily SLR integrated with guidelines and regulatory approvals, our Living-DST delivers real-time, clinically actionable decision support. This system offers oncologists patient-specific insights and may improve cancer treatment outcomes.

REFERENCES & ACKNOWLEDGMENTS

1. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (updated August 2024). Cochrane, 2024. Available from www.cochrane.org/handbook.
2. Page M J, McKenzie J E, Bossuyt P M, Boutron I, Hoffmann T C, Mulrow C D et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews BMJ 2021; 372 :n71 doi:10.1136/bmj.n71
3. Saro Sarkisian MD et al. A living systematic literature review (L-SLR) for non–small-cell lung (NSCLC), prostate (PC), and breast cancer (BC), built with an agentic text annotation system powered by large language models (LLM) to assist treatment decision making.. JCO 43, e13657-e13657(2025). DOI:10.1200/JCO.2025.43.16_suppl.e1365
4. https://masterofcode.com/blog/chatgpt-statistics
5. https://www.chatbees.ai/blog/chatgpt-knowledge-base
6. https://team-gpt.com/blog/perplexity-review/
7. https://www.elegantthemes.com/blog/business/perplexity-ai
8. https://generativeai.pub/from-google-to-perplexity-navigating-the-next-generation-of-search-engines-f26f7963b377
9. https://myjotbot.com/blog/consensus-ai-tool
10. https://consensus.app/home/blog/welcome-to-consensus/
11. https://www.sequoiacap.com/podcast/training-data-daniel-nadler/
12. https://www.openevidence.com/announcements/openevidence-the-fastest-growing-application-for-physicians-in-history-announces-dollar210-million-round-at-dollar35-billion-valuation
13. https://pmc.ncbi.nlm.nih.gov/articles/PMC12142943/

Follow Oncoscope-AI on LinkedIn

Watch demo and read our other publications on the Oncoscope-AI website