

Artificial Intelligence Aids in Data Gathering for Cancer Research

To modernize and accelerate data collection in large population studies, American Cancer Society researchers explored the use of AI.

May 7, 2025 By American Cancer Society

New [research](#) led by [American Cancer Society](#) (ACS) scientists found that large language model (or LLM) -based systems, a form of artificial intelligence, can offer a time-efficient and accurate solution to abstract key data for use in cancer research studies. Clinical data elements that are crucial to oncologic research are often only captured in patients' unstructured medical records, which can contain hundreds of pages of clinical notes. To maximize the value of these data, medical records have traditionally been manually abstracted by oncology data specialists, which is a time and labor-intensive effort. To modernize and accelerate data collection in the large ACS population studies, ACS explored the use of LLM-based systems. The findings are to be presented today at this year's [annual meeting of the American Association for Cancer Research](#) (AACR) in Chicago, April 25-30, 2025.

Researchers in Population Science at ACS, led by associate scientist [Jillian Nelson](#), MPH, analyzed oncologic data abstracted from breast cancer-related medical records collected during routine follow-up of participants in the ACS's longitudinal Cancer Prevention Study 3, which initially enrolled 300,000 cancer-free adults nationwide from 2006-2013. Medical records and associated data, abstracted by Oncology Data Specialists, were used to develop and test the performance of the LLM-based platform Distill to abstract seven data elements: cancer behavior, laterality, neoadjuvant therapy status, and presence of key biopsy or surgery with associated procedure dates. Development began with aligning the Distill platform with 200 medical records and associated data. Abstraction guidelines and decision trees following the guidance of the North American Association of Central Cancer Registries were used as part of the development process. [Den E Bloodworth](#), an ACS oncology data specialist, adjudicated disagreements between human and LLM-based abstraction during development.

The Distill platform was then tested using the remaining 100 medical records, and the platform completed abstraction for this test set within 1 day. Accuracy across all of the abstracted data elements matched or exceeded human abstraction, ranging from 94% -100% accuracy.

This news release was published by the [American Cancer Society](#) on April 30, 2025. It is

republished with permission.

Click here for more news from [AACR 2025](#).

© 2026 Smart + Strong All Rights Reserved.

<http://beta.docker.cancerhealth.com/article/artificial-intelligence-aids-data-gathering-cancer-research>