Revealing What Was Missed: AI Advances Discovery in Large Scale Cancer Data

Kelly Paulson MD PhD[1], Sasha Stanton MD PhD[2], Zachary Reitz PhD[1]
[1]Paul G Allen Research Center at Swedish Cancer Institute
[2]Providence Cancer Institute, Earle A Chiles Research Institute

Main quantitative results in this report are derived by AutoDiscovery, a discovery tool built by Allen Institute for AI

Despite major advances in cancer prevention, screening, and early detection, the incidence of many cancers continues to rise. Breast cancer now is now diagnosed in around 300,000 people each year in the United States (*1*), and melanoma for another 100K (United States Cancer Statistics) (*2*). These malignancies are among the most common cancers diagnosed in young adults making the need for deeper understanding especially urgent (*3,4*). While recent therapeutic innovations have brought meaningful progress, too many patients still face later diagnosis, limited therapeutic options, or poor outcomes. To drive the next generation of breakthroughs, it is essential to better understand the clinical, biological, and tumor-specific factors that shape cancer risk, presentation, and response to treatment.

The United States has invested heavily in large-scale cancer data resources, most notably The Cancer Genome Atlas (TCGA), which generated an unprecedented public repository of comprehensive, anonymized, clinical and molecular data across thousands of tumors. This effort included more than 1,000 breast cancers and approximately 500 melanomas and required an investment of roughly $1 billion to complete. Yet the sheer scale and complexity of these datasets have made them challenging to fully explore. As a result, many analyses have relied on targeted, "fishing" approaches where a predefined hypothesis is cast like a lure and thus largely recovers what was expected (*5*). We hypothesize that systematic exploration can reveal important patterns missed by earlier methods and unlock new insights from these substantial existing investments. AutoDiscovery (*6*) introduces a key breakthrough by combining hypothesis generation with surprisal detection, allowing systematic discovery that is not possible with other LLMs that rely solely on user supplied hypotheses.

Physician-scientists, data scientists, and AI experts at the Paul G. Allen Research Center (Providence Swedish Cancer Institute in Seattle), the Earle A. Chiles Research Institute (Providence Cancer Institute in Portland), and the Allen Institute for Artificial Intelligence (Ai2) are working together to apply AutoDiscovery to melanoma and breast cancer datasets from The Cancer Genome Atlas (TCGA). Reassuringly, for well-established findings that inform current therapies, such as the higher frequency of BRAF mutations in melanoma among younger patients or the way PIK3CA mutations shape PI3K pathway gene expression and prognostic differences in breast cancer, the tool has identified these hypotheses as clinically and biologically important, reproduced the expected outcomes, and categorized them as non-surprising. This performance supports the accuracy and reliability of AutoDiscovery.

More importantly, AutoDiscovery revealed several key areas for further investigation and generated new plausible hypotheses. In melanoma, it identified novel associations involving immune cell

infiltration that are now undergoing secondary validation. In breast cancer, it highlighted multiple factors associated with lymph node spread, a finding that could influence surgical strategies in early-stage disease. Current efforts are focused on streamlining hypotheses, reducing redundancy, and incorporating clinical expertise input to prioritize areas with the most immediate potential for research advancement and clinical impact.

References

1. A. N. Giaquinto, *et al.*, Breast cancer statistics 2024. *CA: a cancer journal for clinicians* **74** (6), 477–495 (2024).

2. K. G. Paulson, *et al.*, Age-specific incidence of melanoma in the United States. *JAMA dermatology* **156** (1), 57–64 (2020).

3. H. J. Kim, S. Kim, R. A. Freedman, A. H. Partridge, The impact of young age at diagnosis (age¡ 40 years) on prognosis varies by breast cancer subtype: A US SEER database analysis. *The Breast* **61**, 77–83 (2022).

4. C. H. Kugel III, *et al.*, Age correlates with response to anti-PD1, reflecting age-related differences in intratumoral effector and regulatory T-cell populations. *Clinical Cancer Research* **24** (21), 5347–5356 (2018).

5. R. Akbani, *et al.*, Genomic classification of cutaneous melanoma. *Cell* **161** (7), 1681–1696 (2015).

6. D. Agarwal, *et al.*, AutoDiscovery: Open-ended Scientific Discovery via Bayesian Surprise, in *The Thirty-ninth Annual Conference on Neural Information Processing Systems* (2025).

Acknowledgments

We thank the Allen Family, patients and their families for their trust and support. We are grateful to the team at the Paul G. Allen Research Center at Swedish, including Doug Kieper, Ash Rajput, Sid Devarakonda, Hank Kaplan, Chuck Drescher, and the broader lab team. We also thank our collaborators at the Allen Institute for AI—Bodhisattwa Prasad Majumder, Dhruv Agarwal, Reece Adamson, Ruben Lozano Aguilera, and Peter Clark—for their contributions. In addition, we acknowledge our partners at Earl A. Chiles Research Institute at Providence Portland, including the Stanton Laboratory, Bryan Bell, and Michael Coleman, and we thank Bill Wright and the Providence team for their support.