ARTIFICIAL INTELLIGENCE

Inferring Negative Molecular Biomarker Data at Scale

Ashleigh E. McBratney, MS[1]; Benjamin A. Holmes, MS[1]; Giles S. Brown, MEng[1]; Raghu Warrier, MS, MT[1]; and Anna B. Berry, MD[1]

original reports abstract

PURPOSE Patients who represent the negative biomarker population, those tested for a biomarker but found to be negative, are a critical component of the growing molecular data repository. Many next-generation sequencing (NGS)–based tumor sequencing panels test hundreds of genes, but most laboratories do not provide explicit negative results on test reports nor in their structured data. However, the need for a complete picture of the testing landscape is significant. Syapse has created an internal ingestion and data transformation pipeline that uses the power of natural language processing (NLP), terminology management, and internal rulesets to semantically align data and infer negative results not explicitly stated.

PATIENTS AND METHODS Patients within the learning health network with a cancer diagnosis and at least one NGS-based molecular report were included. To obtain this critical negative result data, laboratory gene panel information was extracted and transformed using NLP techniques into a semistructured format for analysis. A normalization ontology was created in tandem. With this approach, we were able to successfully leverage positive biomarker data to derive negative data and create a comprehensive data set for molecular testing paradigms.

RESULTS The application of this process resulted in a drastic improvement in data completeness and clarity, especially when compared with other similar data sets.

CONCLUSION The ability to accurately determine positivity and testing rates among patient populations is imperative. With only positive results, it is impossible to draw conclusions about the entire tested population or the characteristics of the subgroup who are negative for the biomarker in question. We leverage these values to perform quality checks on ingested data, and end users can easily monitor their adherence to testing recommendations.

JCO Clin Cancer Inform 7:e2200158. © 2023 by American Society of Clinical Oncology

INTRODUCTION

As genetic sequencing becomes more available to patients, especially in the oncology space, there is an ever-growing database of biomarker information. However, very few laboratories provide explicit negative results on their reports or in structured data feeds. Currently, the standard among molecular laboratories is to list pertinent negatives—explicit negatives in a subset of genes that are relevant to the cancer type and/or targeted therapies, typically <10 genes. Thus, many databases relying on result ingestion from partners are limited in that they only include patients for whom a genetic alteration was detected or explicit negatives for 10 or fewer genes. With these limited data, it is not possible to make any claims about the entirety of the population who is being tested nor the population who is negative.

Additional challenges arise from the fact that laboratories change the composition of their sequencing panels. Genome-based oncology research is continuously advancing, and as new genes and alterations are linked to cancer or novel therapeutics are developed to target pathogenic mutations, these genes and/or variants must be included in tumor sequencing panels. As an example, one prominent laboratory's panel contained 595 genes on its initial launch and now contains 649 genes.[1,2] The challenge for informaticians is identifying these changes in real time to accurately reflect the panel size and composition.

Many guidelines now recommend genetic testing for oncology[3-6] and as it becomes standard practice, it is critical to capture both positivity rates and testing rates. Positivity rates are vital for life science companies seeking to develop targeted therapies, identify acquired resistance mutations, and to analyze real-world data on the efficacy of targeted agents.[7] As real-world data become more available, regulatory agencies have issued guidance in using these data in approval decisions for new drugs.[8] Testing rates are valuable to

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on January 24, 2023 and published at ascopubs.org/journal/cci on March 8, 2023: DOI https://doi.org/10.1200/CCI.22.00158

CONTEXT

Key Objective

To develop a methodology for computationally inferring negative genomic results not explicitly stated in molecular laboratory reports.

Knowledge Generated

A negative inference process and pipeline was developed to allow for complete testing data to be reflected in our molecular database. After the implementation of our design, we show a dramatic improvement in completeness of biomarker data, as well as alignment with expected population rates.

Relevance

Currently, most molecular laboratories report positive test results and pertinent negative results but do not provide negative results for all genes tested. This practice impairs the ability of data analysts and informaticists to determine exactly which patients were sequenced for which genes and who were negative. Our negative inference pipeline has allowed for greater completeness of biomarker data, improving the ability of health care systems, scientific investigators, and regulatory agencies to have a more complete picture of the molecular testing landscape.

health care systems as they seek to improve their practices and ensure guidelines are met.[9] Additionally, this information is beneficial to determine if rates of utilization of targeted therapeutics, on the basis of biomarker results, are in line with the field. Both positivity rates and testing rates rely on negative results to be calculated appropriately.

Negative biomarker results are also critical to ensuring quality control in real-world oncology data. Positivity rates are readily available in public biomarker data sets[10-13] and are often published. Completeness of real-world biomarker data, including negative results, is imperative to perform checks against published values.

Historically, use of tumor sequencing results in a real-world setting relied on manual review and abstraction of clinical reports. The mechanism by which reports are delivered to practitioners varies from faxes, online portals, PDF document delivery, and automated data feeds.[14] To complicate matters further, there is suboptimal standardization in molecular reporting, leading to variability in abstraction and increased time to locate relevant biomarkers between report types.[15] Considering that most laboratories do not explicitly call out negative results, a manual review of hundreds of genes and comparison with reported positives is not a feasible solution. Current collection methods for negative results are limited to this manual review and comparison and are heavily affected by changing panels and the lack of conformity between laboratory reports.

Molecular testing data is siloed and not well-coordinated across all data consumers, leading to inconsistent insights and representations of this data to end users.

Understanding all of the biomarkers tested for a patient that did not return any findings (a.k.a. negative results) is a significant research undertaking toward the outcome of accurately assessing testing and positivity rates at scale and using those negative findings for key insights into the molecular testing landscape. We have worked to address this data gap, and in doing so, we have enhanced the quality and depth of our data.

PATIENTS AND METHODS

After the development and implementation of the negative inference pipeline, a retrospective analysis was conducted to examine the impact on molecular biomarker data quality and completeness.

Data Source

The Syapse Learning Health Network (LHN) is an integrated data network encompassing data from large community-based health systems across the United States. Currently, there are more than 3M patients who are part of Syapse's LHN. The LHN currently consists of 450+ hospitals, 5,200+ outpatient clinics, 1,900+ oncologists, and 216,000+ newly diagnosed patients added annually.

Study Population

Patients included in the inferred negative population had a cancer diagnosis and at least one molecular report within the LHN performed by next-generation sequencing. Those reports which contained multiple testing technologies were excluded. Reports that were canceled or associated with samples were classified as quality not sufficient, did not have negatives inferred. Additional data requirements for molecular reports included a readable PDF or XML file containing an appropriate laboratory name, test name, and report date. The final data set consisted of 14,856 reports across 11,373 patients, 92% of which were solid tumor samples and the remaining being liquid biopsies.

Extraction and Processing of Gene Lists

For laboratories that do provide a gene list in their data feed, we were able to easily create entries in our test gene list ontology. However, for the majority of laboratories, to determine which genes were present on a panel at a given time, as well as the types of alterations that were sequenced, Syapse developed a way to leverage natural language

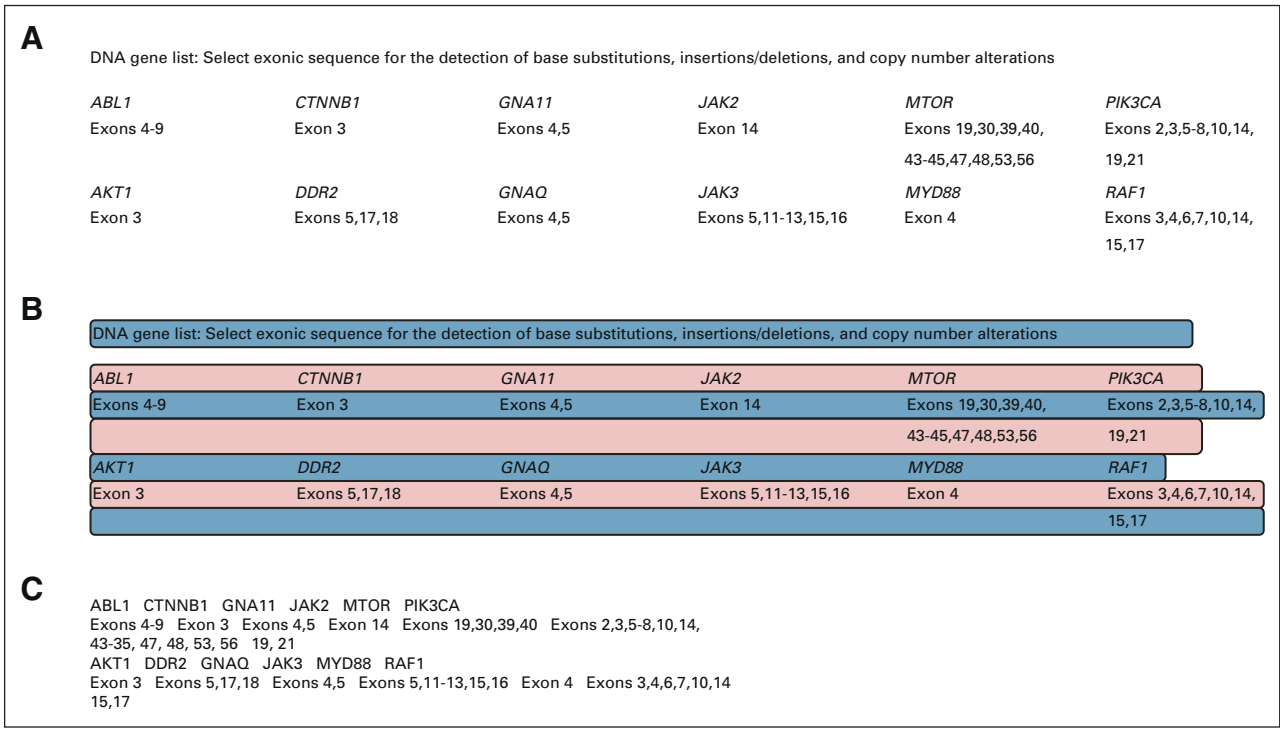

FIG 1. Example of processing of gene list using OCR and detectron model. (A) PDFs are normally formatted to be easily read, but text position is commonly lost in OCR PDF readers. These readers instead focus on reading the text line-by-line. (B) These highlights show the order in which PDF OCR reads the text—note that the association between exon and gene is lost. (C) The translated results of this PDF using a typical OCR program. It is now impossible to associate the correct exons with their genes if exon numbers take up more than one line. OCR, optical character recognition.

processing (NLP) to scan report PDFs to determine the appropriate gene alteration list.

To cover the use case where gene lists were not provided in a structured form, the lists were extracted from unstructured documents. Unstructured documents arrive in a variety of formats, from PDF through HL7 which in turn presents a scalability challenge. To avoid manual extraction in these cases, we used NLP to pull the gene lists from these reports.

Laboratory reports were first normalized to a text-based format. For HL7 and text-based reports, this required only extraction of the relevant sections of the report from the larger file. In the case of PDFs that were provided without textual data, optical character recognition (OCR) was used for transformation. In some cases, PDFs were structured in such a way that a simple OCR resulted in text files with out-of-order text (Fig 1). In these cases, a detectron model was used to structure the PDF and extract text into coherent chunks.

Once extracted, reports were separated into segments, and a clinical metathesaurus[16] was used to scan the resulting text—gene lists were identified by finding clinically relevant terms, and information about the gene name and any associated amino acid or exon ranges, as well as the type of scan performed on that gene (mutations, fusions, etc), was extracted. As great a specificity as possible was used to represent the sections of gene tested: Where amino acid ranges were provided, these were used. If they were not, but exons were provided, these were used. Finally, if only the gene itself was listed, this was represented with no exon or AA information (Fig 2).

The output of this process is a raw file containing the genes and associated alterations that are sequenced across time. NLP-generated lists were then validated against the ingested report data. Gene lists extracted from PDF reports using NLP were manually checked against the PDF reports for each date range, ensuring that all tested genes, as reported on the PDR, were accounted for. Refinement of the NLP method for extracting gene lists was conducted until the generated gene lists matched the PDF report.

Validated gene alteration lists were then correlated, according to the date the report was issued, to ascertain the date on which a panel was changed. It is important to note that it was necessary to determine not only which genes were present on the panel but also the alterations tested for. Possible alterations included single-nucleotide variants, insertion deletions (indels), rearrangements/fusions, copy number variants, and splice site alterations. Typically these decisions are made based on the biological function of the gene and/or relevant therapeutics to target certain types of mutations; however, it is at the discretion of the testing laboratory. Additionally, although there are overlaps between the gene/alterations tested, there is no standard for oncology sequencing panels and thus

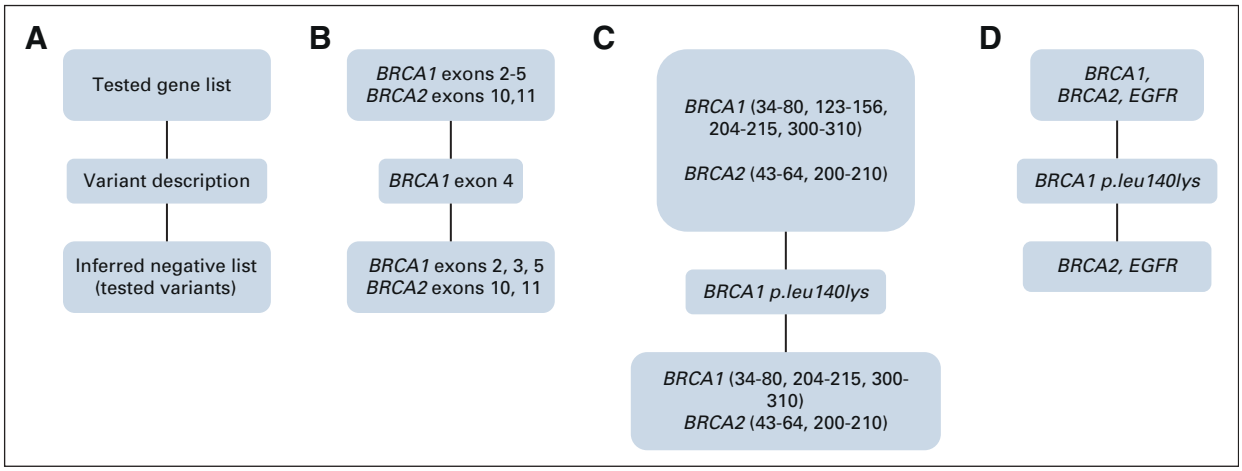

FIG 2. Diagrammatic representations of example report types and the resulting creation of an inferred negative list. (A) Sample diagram, showing how the inferred negative list is derived—the variants are subtracted from the total tested genes. (B) If the exons tested are given in the gene list and we are given which exons had the variant, only the exon associated with the variant is subtracted from the total list. Here, exon 4 contained the variant, so *BRCA1* exon 4 is removed from the inferred negative list. If all exons are removed, the gene itself is removed from the inferred negative list. (C) If the gene list provides amino acid ranges and the variant gives the amino acid at which the variant was found, only the amino acid ranges that contained the variants are removed from the list—here, range 123-156 contained the variant at position 140, so it is removed from the inferred negative list. If all ranges are removed, the gene itself is removed from the inferred negative list. (D) If only gene names are provided in the genes tested list, the entire gene is removed from the inferred negative list if any variant is found—this represents the most clarity we can achieve with the given testing specificity.

informaticians who rely on ingested data must derive another way to reliably obtain gene alteration lists.

Ontology Construction

After review of NLP output, an ontology was then constructed using this data. The ontology was designed with the concept of a panel list and sequenced variant types changing over time in mind (Fig 3).

Once the ontology was built and the relevant data were populated, the ontology was published to BioPortal[17]—a service for ontology management. In addition to this test gene alteration ontology, Syapse currently maintains a number of ontologies that are used in the normalization of incoming data to aid in merging and cleaning disparate data sets.

Pipeline Development and Algorithm

The negative inference pipeline consists of three key steps: extraction, lookup, and inference (Fig 4). Extraction involves querying multiple tables that contain report-level and biomarker-level information and creating an in-memory representation of this information, called a report. The lookup step establishes the link between the test gene alteration ontology and the ingested data. For each report, on the basis of the laboratory name, test name, test version code, alteration type, and report date, a full gene list is determined. When a test version code is not available, a date bisection algorithm is applied to account for the fact that laboratories will update a gene list but continue to use the same test name. Once lookup has been performed, the full gene list and the explicit biomarker results from the molecular report are available. The set difference between the two gene alteration lists is taken, and the results are the inferred negatives for the report.

After completion of the negative inference pipeline, results are written into the corresponding tables, noting that these results have been inferred by this process. Secondary verification of inferred negative results produced by the pipeline was also performed, comparing the reported positive and negative results with the inferred negatives. A randomized set of 10 reports were selected for each test date range. Any errors discovered during the verification process were remedied and reverified.

Impact Analysis

We conducted a positivity rate analysis comparing our internal data sets with other publicly available data sets and literature. Selected cancer types were lung, breast, and colon cancer as these tumor types have the most recommendations for genetic testing and standard-of-care biomarker-based targeted therapeutics.[18] Specific variant types were also selected in each gene/tumor pair to highlight the impact of using a gene alteration ontology. Positivity rates were compared with publicly available data sets and publications specific to each tumor gene variant group. Testing rates were examined in a similar fashion, comparing the testing rates before and after implementation of the negative inference pipeline for select cancer and variant types.

RESULTS

We were able to infer an accurate and comprehensive set of negative biomarkers that enabled us to better measure

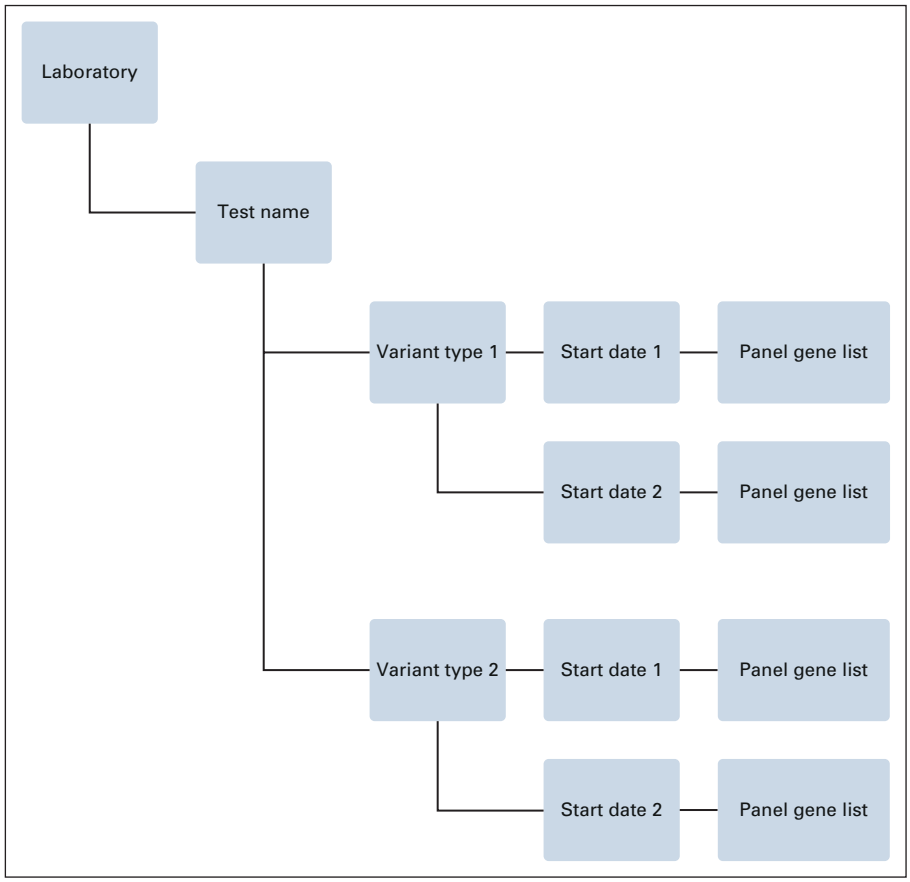

FIG 3. Example of the ontological hierarchy that allows for inference of negative results.

testing rates and positivity rates and to improve our biomarker completeness.

To evaluate the utility of the inferred negative process, we calculated individual biomarker testing rates within the analysis cohort before and after the implementation of the pipeline (Table 1). All selected biomarkers are associated with an US Food and Drug Administration–approved targeted therapy in the selected cancer type.[19] Patients included had at least one next-generation sequencing (NGS) molecular report containing data required for negative inference,

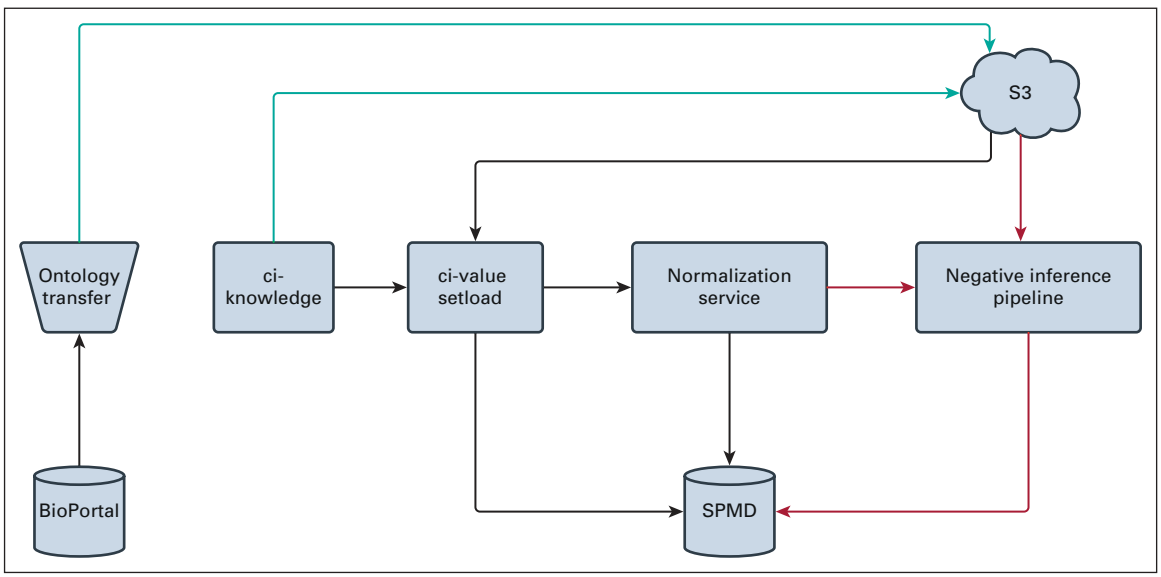

FIG 4. Schematic representation of the data transformation process. SPMD, Syapse Precision Medicine Database.

TABLE 1. Testing Rates of Select Alterations in the Analysis Cohort Before and After Addition of Inferred Negative Data

Gene	Alteration Type	Tumor Type	Targeted Therapy	Test Rate, % (before inferred negatives)	Test Rate, % (after inferred negatives)
KRAS	Sequence variant	Colon	Sotorasib	69.42	98.71
EGFR	Sequence variant	Lung	Afatinib, dacomitinib, erlotinib, gefitinib, mobocertinib, osimertinib	58.36	96.17
PIK3CA	Sequence variant	Breast	Alpelisib	47.88	96.32
NTRK1	Fusion	All	Entrectinib, larotrectinib	0.05	77.19
NTRK2	Fusion	All	Entrectinib, larotrectinib	0.04	58.38
NTRK3	Fusion	All	Entrectinib, larotrectinib	0.05	29.61

including laboratory name, test name, and report issue date. Rates improved for each biomarker evaluated, as demonstrated by paired T test ($P = .0008$).

To evaluate the accuracy of positivity rates after implementation of the negative inference pipeline, we searched several publicly available data sets and peer-reviewed articles to obtain comparator values. Reporting laboratories do not readily provide their internal positivity rates, so alternative data sets were chosen for comparison. Large, publicly available data sets were selected because of their similarity to the LHN—their populations contain a diverse array of tumor types and sequencing results primarily derived from NGS-based tests. The Sanger Institute's Catalog of Somatic Mutations in Cancer (COSMIC) contains data from both peer-reviewed papers and sequencing results and is a trusted source for assessment of somatic variants.[20] Tissue types were selected to match our comparison groups: breast, lung, and large intestine. cBioPortal, developed at Memorial Sloan Kettering Cancer Center, was selected because of its breadth of data and open-source platform allowing for easy data set creation and analysis. cBioPortal's curated set of nonredundant studies for each tumor type group (breast, bowel, and lung) was used. Finally, the National Cancer Institute's Genomic Data Commons (GDC) database, which includes data from National Cancer Institute programs such as The Cancer Genome Atlas (TCGA) and TARGET, was used as a large database comparator.[21] Primary sites of breast, colorectal, and lung were applied for this assessment. Analysis using these databases was completed between August 1, 2022, and August 31, 2022. Additional publications focused on each tumor alteration pair were also included to ensure that a comprehensive set of positivity rates were available for comparison. Our investigation demonstrated that Syapse's molecular data falls within the range of positivity rates published by comparators[10-13,21-28] (Fig 5).

As a measure of completeness, we assessed the total number of biomarker results in our data set. After the implementation of the negative inference pipeline, Syapse was able to increase our biomarker result repository 8-fold (Table 2).

DISCUSSION

The work presented here highlights a critical gap in the current biomarker data ingestion paradigm. Clinical laboratories that run large NGS-based tumor sequencing panels often do not explicitly report negative results. This requires additional steps to be taken to identify tested, but negative, genes. At present, there are several methods by which negative results are presented and collected. In the clinical

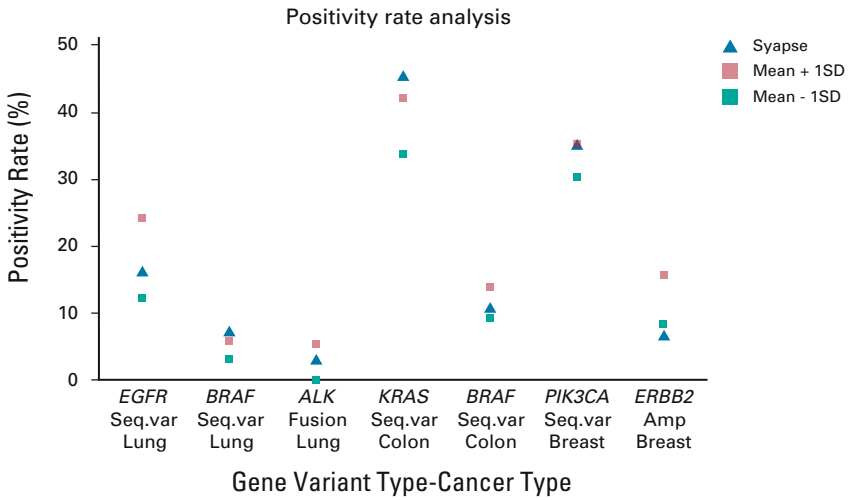

FIG 5. Positivity rate analysis comparing Syapse data set with other publicly available external sources. Mean positivity rate among comparator values was calculated; points represent a range of one standard deviation. Values can be found in Appendix Table A1. SD, standard deviation.

TABLE 2. Results of the Implementation of the Negative Inference Pipeline on Syapse's Molecular Biomarker Repository

Explicit NGS Biomarker Results	Inferred NGS Negatives
1,678,966	8,884,559

Abbreviation: NGS, next-generation sequencing.

setting, a physician can reference the list of tested genes at the end of the report and compare them with the reported positive results to ascertain the negative results in real time. These results may be recorded separately in the physician's note or could also be partially or fully replicated in the original surgical pathology report, at the discretion of the physician or pathologist involved. Physician notes typically do not detail all genes tested and only highlight positive findings. On the data ingestion side, negative results are limited to those that are included in the structured report data, which typically consists of a few pertinent negatives, often related to tumor type and clinical guidelines, only. To use negative results to create a comprehensive database of all tested genes, large scale data companies are therefore reliant on manual abstraction of negative results. Many large panels number more than 500 genes and require a sizable time investment for manual capture, which is also susceptible to human error. Additionally, as gene lists are not stable over time, a set gene list for each test cannot be relied on. In some cases, a full PDF of the molecular report is not available for manual review, and abstractors are reliant on physician notes included in the electronic medical records.

For any population-based study, internal health system quality control, observational research, external collaboration, or regulatory purpose, depth and uniformity of tested, but negative, data is critical for calculation of metrics such as the positivity rate and testing rate. To address this, we have created an AI and informatics-based approach that allows for inference of negative results using the gene lists provided by the laboratory in PDF reports and/or data feeds. With the implementation of this pipeline, we demonstrate a substantial improvement in data quality and completeness.

Testing rates in the LHN show a significant increase before and after the application of the inferred negative pipeline, illustrating an important role in accurately reflecting the tested population. These rates are a critical measure that rely on accurate counts of tested patients. National Comprehensive Cancer Network guidelines recommend genetic testing for all selected alterations in our analysis, for their respective cancer types, either at initial disease presentation or after the failure of first-line therapy.[3-5] Additionally, pan-tumor testing for *NTRK1/2/3* fusions is becoming increasingly common among physicians because of the approval of several TRK inhibitors.[29]

Our analysis on the inferred negative data set, which aims to better reflect testing data, shows that these recommendations and trends are being adopted in clinical testing panels. As precision oncology continues to advance, molecular testing is becoming a cornerstone in therapeutic decision making. In-depth data surrounding the utilization of molecular sequencing, beyond the presence of a mutation, is therefore a meaningful metric for providers.

Positivity rates within selected tumor gene alteration groups were found to be comparable with other data sets and publications. However, it is difficult to make statistical comparisons as the rates in which mutations are observed in certain cancer types are variable and heavily dependent on the patient population. Aspects such as cancer stage, patient age, sex, smoking status, and geographical location, among others, have all been shown to affect positivity rate.[30,31] Although large data sets and studies exist as reference points, we expect that our patient population will differ based on the aforementioned variables. Accordingly, we opted to make a qualitative comparison of positivity rates in our investigation. With the creation of our negative inference pipeline, we hope to enable end users to get a more complete picture of their population of interest.

It is critical to note that the use cases for inferred negative results are intended to be population-based retrospective studies, regulatory studies, and quality improvement programs. The intent is not to replace or overwrite any results reported by a testing laboratory or to inform individual patient care decisions. We are leveraging the data already provided by the clinical laboratories in their PDF reports and making it more easily accessible and readable for insights and analytics. Clinical care impact for this work resides in the ability of health care systems, health care providers, life science investigators, and regulatory agencies to more readily determine if patients are being tested as recommended by guidelines, as required for large-scale clinical trial matching and as needed to provide clean uniform data pertaining to negative results for observational research studies.

An ever-present challenge in collecting and analyzing real-world data is biomarker completeness, even when PDF reports and structured and unstructured data are provided.[32] Although physicians are able to make decisions regarding patient care and clinical trial eligibility on the basis of those variants not reported to be positive but referenced as tested in a gene list, informaticians struggle with these implied negative results since they are not readily present in ingested data. Population-based studies thus become more difficult to conduct as the molecular data do not paint a complete picture beyond patients who tested positive for a biomarker. Machine learning–based inferred negatives are a way to alleviate this gap and enhance the completeness of real-world data.

AFFILIATION

[1]Syapse Inc, San Francisco, CA

CORRESPONDING AUTHOR

Ashleigh E. McBratney, MS, Syapse Inc, 2021 Fillmore St, Ste #1183, San Francisco, CA 94115-2708; e-mail: ashleigh.mcbratney@syapse.com.

SUPPORT

Support for the study was provided by Syapse Inc.

AUTHOR CONTRIBUTIONS

Conception and design: All authors
Administrative support: Raghu Warrier
Collection and assembly of data: Ashleigh E. McBratney, Giles S. Brown
Data analysis and interpretation: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Ashleigh E. McBratney
Employment: Syapse
Stock and Other Ownership Interests: Tempus, Syapse
Patents, Royalties, Other Intellectual Property: Patent application 17/546,049: Artificial Intelligence Driven Therapy Curation and Prioritization

Benjamin A. Holmes
Employment: Syapse
Stock and Other Ownership Interests: Syapse

Raghu Warrier
This author is a member of the *JCO Clinical Cancer Informatics* Editorial Board. Journal policy recused the author from having any role in the peer review of this manuscript.

Anna B. Berry
Employment: Syapse
Stock and Other Ownership Interests: Syapse
Research Funding: Tempus

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

The results shown in Figure 4 are in part based on data generated by the TCGA Research Network: https://www.cancer.gov/tcga. MetaMap software is courtesy of the US National Library of Medicine. The authors thank Syapse for providing the space, time, and resources to complete our work.

REFERENCES

1. Tempus Unveils Tempus xT, 595 Gene Panel Aimed at Empowering Data-Driven Cancer Care. 2017 https://www.globenewswire.com/en/news- release/2017/10/19/1150526/0/en/Tempus-Unveils-Tempus-xT-595-Gene-Panel-Aimed-at- Empowering-Data-Driven-Cancer-Care.html
2. xT Gene Panel: For Use With xT I 648 Gene Panel Reports, 2020 https://www.tempus.com/wp-content/uploads/2020/10/xTGene-List_100620-1.pdf
3. NCCN Clinical Practice Guidelines in Oncology: Breast Cancer Version 4. 2022 https://www.nccn.org/professionals/physician_gls/pdf/breast.pdf
4. NCCN Clinical Practice Guidelines in Oncology: Colon Cancer Version 1. 2022 https://www.nccn.org/professionals/physician_gls/pdf/colon.pdf
5. NCCN Clinical Practice Guidelines in Oncology: Non-Small Cell Lung Cancer Version 3. 2022 https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf
6. Somatic Genomic Testing in Patients with Metastatic or Advanced Cancer. 2022 https://www.asco.org/practice-patients/guidelines/molecular-testing- and-biomarkers#/168858
7. Sherman RE, Anderson SA, Dal Pan GJ, et al: Real-world evidence—What is it and what can it tell us? N Engl J Med 375:2293-2297, 2016
8. Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics Guidance for Industry. 2019. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-using-real-world-data-and-real-world-evidence-fda-drugs-and-biologics-guidance
9. Alberty-Oller JJ, Weltz S, Santos A, et al: Adherence to NCCN guidelines for genetic testing in breast cancer patients: Who are we missing? Ann Surg Oncol 28:281-286, 2021
10. Cerami E, Gao J, Dogrusoz U, et al: The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov 2:401-404, 2012
11. Gao J, Aksoy BA, Dogrusoz U, et al: Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 6:pl1, 2013
12. Tate JG, Bamford S, Jubb HC, et al: COSMIC: The catalogue of somatic mutations in cancer. Nucleic Acids Res 47:D941-D947, 2019
13. Weinstein JN, Collisson EA, Mills GB, et al: The Cancer Genome Atlas Pan-Cancer Analysis Project. Nat Genet 45:1113-1120, 2013
14. Rioth MJ, Thota R, Staggs DB, et al: Pragmatic precision oncology: The secondary uses of clinical tumor molecular profiling. J Am Med Inform Assoc 23:773-776, 2016
15. Jones GR, Legg M: Report formatting in laboratory medicine—A call for harmony. Clin Chem Lab Med (Cclm) 57:61-65, 2018
16. Aronson AR: Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. Proc AMIA Symp:17-21, 2001
17. Noy NF, Shah NH, Whetzel PL, et al: BioPortal: Ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 37:W170-W173, 2009
18. NCCN Biomarkers Compendium. 2022 https://www.nccn.org/professionals/biomarkers/content/
19. Table of Pharmacogenomic Biomarkers in Drug Labeling. https://www.fda.gov/drugs/science-and-research- drugs/table-pharmacogenomic-biomarkers-drug-labeling

20. Li MM, Datto M, Duncavage EJ, et al: Standards and guidelines for the interpretation and reporting of sequence variants in Cancer. J Mol Diagn 19:4-23, 2017
21. Grossman RL, Heath AP, Ferretti V, et al: Toward a shared vision for cancer genomic data. N Engl J Med 375:1109-1112, 2016
22. Safaee Ardekani G, Jafarnejad SM, Tan L, et al: The prognostic value of BRAF mutation in colorectal cancer and melanoma: A systematic review and meta-analysis. PLoS One 7:e47054, 2012
23. Chia PL, Mitchell P, Dobrovic A, et al: Prevalence and natural history of ALK positive non- small-cell lung cancer and the clinical impact of targeted therapy with ALK inhibitors. Clin Epidemiol 6:423-432, 2014
24. Cizkova M, Susini A, Vacher S, et al: PIK3CA mutation impact on survival in breast cancer patients and in ERα, PR and ERBB2-based subgroups. Breast Cancer Res 14:R28, 2012
25. Dong Z, Kong L, Wan Z, et al: Somatic mutation profiling and HER2 status in KRAS-positive Chinese colorectal cancer patients. Sci Rep 9:16894, 2019
26. Nakamura K, Aimono E, Oba J, et al: Estimating copy number using next-generation sequencing to determine ERBB2 amplification status. Med Oncol 38:36, 2021
27. O'Leary CG, Andelkovic V, Ladwa R, et al: Targeting BRAF mutations in non-small cell lung cancer. Transl Lung Cancer Res 8:1119-1124, 2019
28. Melosky B, Kambartel K, Hantschel M, et al: Worldwide prevalence of epidermal growth factor receptor mutations in non-small cell lung cancer: A meta-analysis. Mol Diagn Ther 26:7-18, 2021
29. Marshall LZ, Klink AJ, Kavati A, et al: BPI21-006: Timing of NTRK gene fusion testing and treatment modifications following NTRK+ status among U.S. oncologists treating NTRK+ patients. J Natl Compr Cancer Netw 19:BPI21-006, 2021
30. Aye PS, Tin Tin S, McKeage MJ, et al: Development and validation of a predictive model for estimating EGFR mutation probabilities in patients with non-squamous non-small cell lung cancer in New Zealand. BMC Cancer 20:658, 2020
31. Mendiratta G, Ke E, Aziz M, et al: Cancer gene mutation frequencies for the U.S. population. Nat Commun 12:5961, 2021
32. Norden A, Wang C, Lane D: Real-world data reveals quality gaps in routine cancer care. Oncol Times 42:24-25, 2020

APPENDIX

TABLE A1. Data for Positivity Rates Represented in Figure 5

Gene- Variant Type- Cancer, Type	Syapse, %	COSMIC, %	cBioPortal, %	GDC, %	Publications, %	Mean, %	SD, %
EGFR seq.var lung	16.0	26.50	18.20	12.65	15.40	18.19	5.99
BRAF seq.var lung	7.0	2.42	4.90	5.56	5.00	4.47	1.40
ALK fusion lung	2.8	NA	0.70	NA	4.50	2.60	2.69
KRAS seq.var colon	45.2	32.32	37.60	41.89	40.00	37.95	4.15
BRAF seq.var colon	10.6	12.36	9.70	14.39	9.60	11.51	2.31
PIK3CA seq.var breast	34.9	29.07	34.50	34.13	33.40	32.78	2.51
ERBB2 amp breast	6.4	8.30	12.20	NA	15.56	12.02	3.63

NOTE. Data set did not have information labeled NA. Mean and standard deviation were calculated on the basis of comparator values, excluding Syapse.

Abbreviations: COSMIC, Catalog of Somatic Mutations in Cancer; GDC, Genomic Data Commons; SD, standard deviation.

