Pragmatic precision oncology: the secondary uses of clinical tumor molecular profiling

RECEIVED 28 August 2015

REVISED 7 December 2015

ACCEPTED 7 January 2016

PUBLISHED ONLINE FIRST 28 March 2016

Matthew J Rieth,^{1,2} Ramya Thota,¹ David B Staggs,² Douglas B Johnson,¹ and Jeremy L Warner^{1,2}

ABSTRACT

Background Precision oncology increasingly utilizes molecular profiling of tumors to determine treatment decisions with targeted therapeutics. The molecular profiling data is valuable in the treatment of individual patients as well as for multiple secondary uses.

Objective To automatically parse, categorize, and aggregate clinical molecular profile data generated during cancer care as well as use this data to address multiple secondary use cases.

Methods A system to parse, categorize and aggregate molecular profile data was created. A naïve Bayesian classifier categorized results according to clinical groups. The accuracy of these systems were validated against a published expertly-curated subset of molecular profiling data.

Results Following one year of operation, 819 samples have been accurately parsed and categorized to generate a data repository of 10,620 genetic variants. The database has been used for operational, clinical trial, and discovery science research.

Conclusions A real-time database of molecular profiling data is a pragmatic solution to several knowledge management problems in the practice and science of precision oncology.

INTRODUCTION

Cancer is fundamentally a disease that arises from disorder in our genome and represents the complex convergent end state of a cacophony of genetic variants.¹ While large scale sequencing projects have recruited thousands of patient samples,² the heterogeneity of cancer ensures that ongoing study of tumor genomics will be required to understand not only the genetic origins of the disease, but also its multiple means of escape from treatment.^{3–5} Tumor genetic testing, also known as molecular profiling, is becoming the standard of care for an increasing number of cancer types.^{4,6–8} Molecular profiling of patient tumors is expanding beyond individual base pair sequencing into next generation sequencing panels of the exons of numerous potential genetic drivers of the cancer.⁹ The increasingly routine use of next generation sequencing molecular profiling in oncology clinics generates data to inform the care of the individual patient, but can also in secondary use inform the science of precision oncology. Using informatics tools to automatically parse and aggregate molecular profiling reports into a real-time database represents a pragmatic use of clinical data to create resources that can be used for robust secondary uses.

In the course of clinical care, the results of molecular profiling are reported back to the ordering provider. The mechanism of results reporting varies from faxes, online portals internal or external to an electronic health record (EHR), or direct integration into the EHR. While there exists an HL7 V2.5.1 standard for the transmission of some clinical genomic results, the use cases for molecular profile reporting are not implemented widely.¹⁰ Previously, we have described a clinical transmission system of tumor molecular profiling reports from a third-party sequencing laboratory into an EHR^{11,12} for real-time clinical practice. In addition, this structured data feed allows molecular results to be parsed, restructured, and stored into a database that can be used to further inform the science and practice of precision oncology. This is similar to other pragmatic databases that are automatically populated from clinical data from EHRs.¹³ Herein, we describe the creation and use cases for an automated, real-time database of annotated tumor variants.

METHODS

Oncology providers at Vanderbilt University Medical Center (VUMC) order tumor molecular profiling of metastatic, advanced, or recurrent cancers according to the individual clinical situation (Figure 1). The tumors undergo molecular profiling at a third party laboratory of a panel of exons according to the laboratory's protocol. The results of the molecular profiling are packaged into a custom extensible markup language (XML) document, which is transmitted securely from the laboratory server to VUMC servers on a nightly basis. Variants are represented with Human Genome Variant Society standardized syntax for naming and genomic alterations.^{14,15} Variants with known significance for the diagnosis, prognosis, or therapy of the specific cancer type are structured in the XML tags, while variants lacking that established clinical significance, “variants of unknown significance” (VUS), are only found within the attached PDF. The reports are parsed and displayed in the EHR for clinical use, but also reflected to a secure server for database storage. This database creation mechanism was approved by the VUMC institutional review board as an information repository preparatory to research.

The daily incoming XML documents are parsed by a Perl (v 5.14.2) script using the XMLSimple package (v2.20) for variants of known clinical significance as well as metadata regarding histology, anatomic site, and logistical information. The PDF undergoes command line conversion to a text document. This text document is then processed using regular expressions in Perl to extract the VUS. The data is then stored in multiple tables in a relational database in a secure server maintained by the Vanderbilt-Ingram Cancer Center's Research Informatics Core.

For operational and other needs, it was necessary to classify cases into “clinical groups”, which roughly correspond to the subject matter expertise of the treating physicians (eg, lung, breast, central nervous system). To classify incoming samples to the clinical groups, multiple potential classifiers were tested. A naïve Bayesian classifier was trained with 10-fold cross validation using R (v 3.2.2), the e1071 package (v. 1.6), and the ranger (0.3.0) package. Ordering provider, patient gender,

Correspondence to Matthew J Rieth, Vanderbilt Department of Medicine, Division of Hematology & Oncology, Vanderbilt University, 2220 Pierce Ave Preston Research Building 777, Nashville, TN 37232, USA; matthew.j.rieth@vanderbilt.edu For numbered affiliations see end of article.

©The Author 2016. Published by Oxford University Press on behalf of the American Medical Informatics Association. All rights reserved.

For Permissions, please email: journals.permissions@oup.com

Figure 1: Flow diagram of data processing. Tumors undergo molecular profiling according to the clinical situation. Molecular profiling reports are packaged into custom XML documents and transmitted daily to clinical servers to incorporate into the chart. The XML is parsed for variants of known clinical significance and sample metadata. A PDF of additional information is converted to text and parsed for “variants of unknown significance.” These data sources are merged with extant clinical data warehouses. A classifier takes heterogeneous metadata strings and classifies reports into one of 18 clinical categories. The resultant data is stored in an integrated molecular and clinical database for multiple use cases.

Figure 2: Automated parsing of variant information is more accurate than manual abstraction. A histogram of discordant reports between manual and automated data extraction. Both methods were concordant in 81 of 94 charts reviewed ($\kappa = 0.95$). All observed discordant reports ($n = 13$) were false negative variants of the manual method verified by review of original report. The automated method did not produce any detected false negative nor false positives.

and submitted histology text variables for 800 expertly-annotated samples were parsed and trained for classification into 18 clinical categories.

To evaluate the accuracy of the parsing and storage protocols, results were compared to a published dataset manually abstracted from the reports of some of the same patients.¹⁶ Samples of 80 reports were manually re-evaluated by two investigators and compared to

those generated by the parsing algorithm. The R statistical package was used to calculate descriptive statistics as well as Cohen's kappa, accuracy, precision, and recall. All appropriate statistical tests were two-sided.

The existence of the database was communicated to basic science, translational, operations, and clinical researchers within VUMC. Data

use cases information was collected and managed using REDCap electronic data capture tools hosted at VUMC.¹⁷ Institutional review board approval was required for data requests from investigators for subsequent research involving personal health information.

RESULTS

The database created by parsing molecular profile reports is accurate and useful across a broad range of precision medicine secondary uses. Over one year of operation it contained 819 unique molecular profiling reports. The reports were parsed to 3435 variants of known significance and 7185 VUS, with a median of 13 variants per report (4 known and 9 VUS per report). Duplicate reports were purged by requiring unique pathology sample identifiers. Reports received were dependent on the test ordered by providers.

In comparison to the manually extracted cohort of 103 patients,¹⁶ the automated XML parsing method was concordant to those data ($\kappa = 0.95$) with 13 discordant variants (median 0, range 0–7) out of 365 variants of known significance (Figure 2). All discordant variants were false negatives by the manual method and verified to be correctly parsed by the algorithm by repeat manual review of the original molecular profiling report. Of the VUS parsed from text using regular expressions, 100% (308 of 308) from 20 charts were verified from the original report to be accurate. No false positive variants were observed.

A Bayesian classifier was trained to classify incoming samples into one of 18 clinically relevant groups, performing with accuracy of 0.994 (795 of 800 samples correctly classified) [Supplemental Table]. Of the five misclassified examples, two were sarcoma likely reflecting the disparate anatomic sites and nomenclatures for sarcomas. Three other misclassifications reflected uncommon tumor types seen by a provider who predominantly sees a common tumor type. In prospective evaluation of 106 cases, the classifier performed with 0.953 accuracy.

Multiple users presented requests for data to answer varied precision oncology use cases. Fifteen use cases were classified into operational, clinical trial, and basic science research projects (Table 1). Data extraction cohorts ranged from 17 reports to all reports (median = 83 reports). Variables of interest included ordering statistics by month and histology, variants information for evaluation of a novel assay, molecular biomarker estimates for clinical trial feasibility estimates, visualization of VUMC population variant data relative to other published data sets, and variant information from a previously-defined cohort of interest. Many informal comments were positive regarding the existence, utility, and accuracy of this database.

DISCUSSION

In addition to informing clinical care, tumor genetic testing can be used to advance the practice and science of precision oncology. A system to parse and aggregate structure tumor genetic variant information is accurate in comparison to a manually curated reference standard and useful in addressing multiple investigations across an academic medical center. The discrepancies between the reference and the automated parser likely indicates filtering during the creation of reference standard of variants that were not clinically relevant to that analysis. This discrepancy illustrates that in practice, the definitions of “known significance” or “clinically relevant” are constantly evolving entities.

The evolution of variant classification will continue to be a challenge in the annotation of molecular profiling results. Of particular note is the role molecular biomarkers will play in future precision oncology trials. One illustrative example is that of an experimental drug, lucitanib, that targets fibroblast growth factor pathway alterations (Table 1, Trial example #3). This first-in-class drug was studied through a clinical trial (NCT02202746) with molecular inclusion criteria of fibroblast growth

Table 1: Precision Oncology Use Cases from Aggregated Tumor Molecular Profiling Data

Type	Data request use case
Operations	1. Figures for cancer center support grant
	2. Molecular profiling utilization statistics
	3. Validation of novel next generation sequencing assay
Trial	1. Prioritizing future clinical phase 1 trials of targeted therapeutics
	2. Solid tumors with MET and NTRK1, 2, 3 variants
	3. Breast cancer patients with FGFR mutations or 11q amplifications
	4. Solid tumors with MAP2K, MAP2K2 variants
	5. Tumors with ERBB family mutations
	6. Solid tumors with novel BRAF mutations
Research	1. Network analysis of VUS in breast cancer
	2. Triple negative breast cancer patients with PIK3CA variants
	3. Novel ERBB2 mutations for validation
	4. Potential germline implications of molecular profiling
	5. Genetics of gastrointestinal malignancies with signet ring histology or peritoneal metastases
	6. SRC variants in lung cancer

factor receptor mutations or chromosome 11q amplifications.¹⁸ Any patients meeting the molecular inclusion criteria prior to the trial opening (and thus prior to incorporation into the knowledge base used to annotate the reports) would not have been identified as eligible. Through the methods we have described, we were able to identify all patients who met these inclusion criteria so they could be screened for this trial.

Since the methodology we describe has been implemented, additional investigators have requested data to determine feasibility and to assist in screening for patients meeting molecular inclusion criteria. Precision medicine will increasingly rely on “basket trials” such as NCI-MATCH,^{19,20} where patients are included not by histology but rather by molecular inclusion criteria.^{21,22} In such trials, being able to screen across populations for patients meeting molecular inclusion criteria will be critical for efficient study conduct. Investigators can use this database to query across populations of cancer patients for molecular biomarkers to determine if a clinical trial is feasible, and once open, to identify eligible patients. Investigators conducting precision oncology clinical trials, particularly early phase clinical trials, have already used this database effectively.

Generally, the practice of precision oncology will require a greater awareness of molecular biomarkers that may fall outside the daily practice of an oncologist. Maintaining current knowledge of molecular inclusion criteria of basket trials is challenging as novel or low frequency biomarkers may be familiar to the physician. This becomes an increasingly difficult task in clinical practice as both molecular profiling and precision oncology trials increase. Additionally, future annotation of variants as clinically relevant relies on the molecular diagnostic lab to maintain a continually updating knowledge base of all molecular biomarkers for clinical trials. The failure at any of these steps prevents patients from being enrolled on precision oncology trials or offered potentially therapeutic treatments. A system such as this can provide another layer of knowledge to ensure all patients who may benefit from a molecularly defined clinical trial are identified.

In addition to clinical trial use cases, basic science and translational researchers have requested data to develop hypotheses for foundational research. Several requests for data came from cancer biologists studying genetic mechanisms of cancer genetics and biomarkers. Aggregating and structuring clinical data for secondary use represents a pragmatic and cost-effective means to facilitate translational science.^{13,23} As more patients accrue in the database, the power to discover novel cancer biomarkers and to drive discovery science increases. The strengths of this approach are modeled on the successful strategy of the Electronic Medical Records and Genomics network by linking molecular information to existing clinical data warehouses.²⁴ Furthermore, the patients who undergo molecular profiling in a pragmatic manner represent unique clinical substrates of exceptional responders or resistant tumors in many cases. Patients within this database constitute more advanced cancers than those represented in databases such as The Cancer Genome Atlas. These aggressive, refractory or unusual clinical situations may reveal new insights into cancer biology.

CONCLUSION

Our methodology is limited by the custom nature of the parser that was created for the custom XML document. As such, it is not extensible to other molecular profiling reports and constitutes a limitation of these methods. This limitation illustrates the need for a consensus means of structured molecular profiling reporting, such as the efforts underway with HL7 and Fast Healthcare Interoperability Resources²⁵ so that the information is interoperable and standardized tools can be built.

This system represents a pragmatic approach to the practice and science of precision medicine in oncology. Structuring and aggregating accurate molecular profiling data facilitates operational, clinical trial, and discovery science use cases that further the field and improve patient care. As more molecular profiling is utilized in cancer care, the data it produces can feed into a learning health system that continues to drive precision oncology practice and discovery.

FUNDING

This work was supported by National Institutes of Health grants 5T15LM007450-12, UL1TR000445, and K12 CA090652; and CCSG P30 CA068485.

COMPETING INTERESTS

None.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Joseph Burden, Lucy Wang, and Michael Nicholas with contributing to database creation and maintenance.

REFERENCES

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144:646–674.
2. (Chairperson) TJH, Anderson W, Aretz A, et al. International network of cancer genome projects. *Nature*. 2010;464:993–998.
3. Santarpia L, Qi Y, Stemke-Hale K, et al. Mutation profiling identifies numerous rare drug targets and distinct mutation patterns in different clinical subtypes of breast cancers. *Breast Cancer Res Treat*. 2012;134:333–343.

AUTHOR AFFILIATIONS

¹Department of Medicine, Division of Hematology & Oncology, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, 37232

4. Ellis MJ, Perou CM. The genomic landscape of breast cancer as a therapeutic roadmap. *Cancer Discov*. 2013;3:27–34.
5. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*. 2012;366:883–892.
6. Van Allen EM, Wagle N, Levy MA. Clinical analysis and interpretation of cancer genome data. *J Clin Oncol*. 2013;31:1825–1833.
7. Tsimberidou AM, Eggermont AMM, Schilsky RL. Precision cancer medicine: the future is now, only better. *Am Soc Clin Oncol Educ Book ASCO Am Soc Clin Oncol Meet*. 2014;61–69.
8. Dancey JE, Bedard PL, Onetto N, et al. The genetic basis for cancer treatment decisions. *Cell*. 2012;148:409–420.
9. Frampton GM, Fichtenholtz A, Otto GA, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol*. 2013;31:1023–1031.
10. Clinical Genomics. <http://www.hl7.org/Special/committees/clingenomics/>. Accessed November 15, 2015.
11. Rieth M, Staggs D, Warner J. Incorporation of externally generated next-generation tumor genotyping into clinical and research workflows: Successes and lessons learned. *J Clin Oncol*. 2014;32:abstr 156.
12. Rieth MJ, Staggs D, Hackett L, et al. Implementing and improving automated electronic tumor molecular profiling. *J Oncol Pr*. 2016;12:e332–e337.
13. Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inf*. 2014;52:28–35.
14. Human Genome Variation Society. <http://www.hgvs.org/>. Accessed November 15, 2015.
15. Dunnen JT den, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat*. 2000;15:7–12.
16. Johnson DB, Dahlman KH, Knol J, et al. Enabling a genetically informed approach to cancer medicine: a retrospective evaluation of the impact of comprehensive tumor profiling using a targeted next-generation sequencing panel. *Oncologist*. 2014;19:616–622.
17. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377–381.
18. Clovis Oncology. A Study to Assess the Safety and Efficacy of the VEGFR-FGFR Inhibitor, Lucitanib, Given to Patients With FGF Aberrant Metastatic Breast Cancer. In: *ClinicalTrials.gov*. Bethesda (MD): National Library of Medicine 2015.
19. Conley BA, Doroshow JH. Molecular analysis for therapy choice: NCI MATCH. *Semin Oncol*. 2014;41:297–299.
20. NCI-Molecular Analysis for Therapy Choice (NCI-MATCH) Trial. Natl. Cancer Inst. <http://www.cancer.gov/about-cancer/treatment/clinical-trials/nci-supported/nci-match>. Accessed August 22, 2015.
21. Ledford H. 'Master protocol' aims to revamp cancer trials. *Nature*. 2013;498:146–147.
22. Sleijfer S, Bogaerts J, Siu LL. Designing transformative clinical trials in the cancer genome era. *J Clin Oncol*. 2013;31:1834–1841.
23. Bowton E, Field JR, Wang S, et al. Biobanks and electronic medical records: enabling cost-effective research. *Sci Transl Med*. 2014;6:234cm3–234cm3.
24. Gottesman O, Kuivaniemi H, Tromp G, et al. The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet Med*. 2013;15:761–771.
25. Alterovitz G, Warner J, Zhang P, et al. SMART on FHIR Genomics: facilitating standardized clinico-genomic apps. *JAMIA*. Published Online First: July 21 2015;22:1173–1178.

²Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, 37203 USA