INTEGRATION OF THE GENOMIC PROFILE WITH OTHER PATIENT DATA FOR TREATMENT AND OUTCOMES RESEARCH

Jack London[1], Karen Fairchild[2], Matvey Palchuk[2]
[1]Jefferson-Sidney Kimmel Cancer Center, [2]TriNetX, Inc.

WHY INTEGRATE GENOMIC DATA WITH OTHER CLINICAL PATIENT DATA?

The integration of available genomic profile data with a patient's phenotypic data set (typically consisting of demographics, diagnoses, medications, procedures, laboratory results and other observations) is of increasing value to both clinical and basic researchers.

- Clinical trial eligibility criteria more frequently include gene variants, as targeted therapy relies more on mutational characteristics of disease mechanism.
- Basic biomedical investigations can use integrated phenotypic and genotypic data to suggest possible avenues for determining disease mechanism or potential treatment strategies.
- Tumor genotyping is important because many tumors have key "driver" oncogenes that can be controlled with highly targeted inhibitor therapies. A number of present-day treatment guidelines require molecular testing or certain diseases. Results of such testing are crucial in identifying patients who might be candidates for particular targeted therapy.

WHAT GENOMIC DATA ARE REQUIRED?

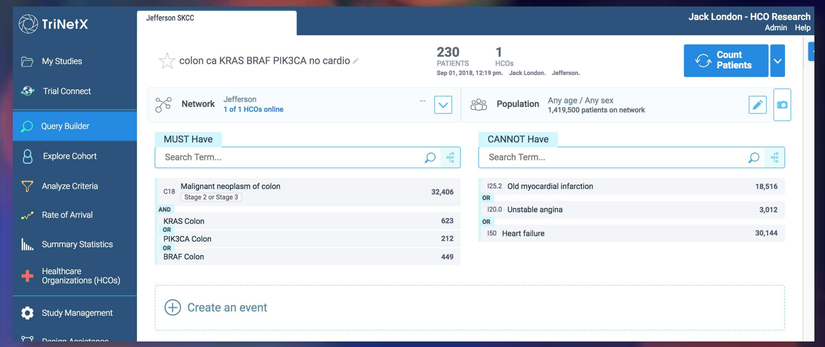
The results of molecular diagnostic tests which include the source of the specimen tested, the targeted gene(s), and whether any variants (mutation) were expressed is needed. For example, a colon tumor tissue sample assayed for the KRAS gene may report the expression of a variant (mutation) such as c.35G>A (G12D) for one patient, while the same molecular diagnostic test panel on another patient may report that no KRAS variants were found.

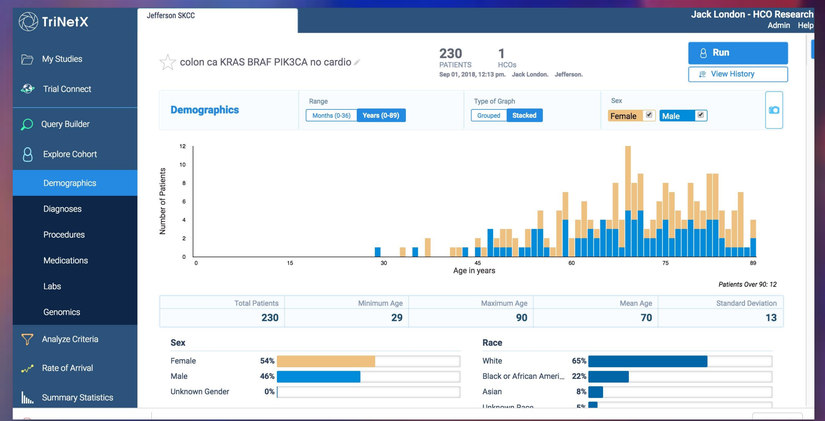
Minimally, the following variant types should be included: complex, copy number gain, copy number loss, deletion, duplication, fusion, indel, insertion, inversion, NT expansion, short repeat, single nucleotide variant, translocation, undetermined variant, and structural variants.

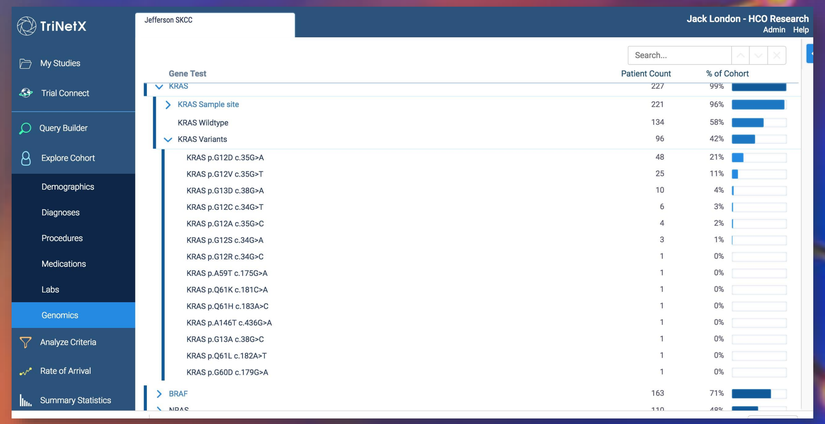
For cancer, the focus is on tumor genotyping. Genotyping results are represented as the presence or absence of specific variants in particular DNA/RNA sequences and/or protein structures. The results do not include any large swathes of individuals' genome and are similar to other clinical observations about patients, such as their traditional lab results.

WHAT STANDARDS-BASED TERMINOLOGY SHOULD BE USED?

- Gene names and symbols should be rigorously defined based on the HUGO Gene Nomenclature Committee (HGNC). Gene symbols are augmented with "previously used" synonyms.
- Variant descriptions should be augmented with synonyms from dbSNP (rsids, chromosome location), OMIM Allelic Variants, UniProtKB, and Breast Cancer Information Core (BIC). Variants (mutations) are mapped to those with clinical significance drawn from ClinVar, dbSNP, dbVar, and Ensembl.
- Variant descriptions should follow the Human Genome Variation Society (HGVS) sequence variant nomenclature for single nucleotide variants (SNV) and other simple variants.
- The International System for Human Cytogenomic Nomenclature (ISCN) can be used for complex structural variants and support both one and three letter amino acid abbreviations or protein sequence variants.







WHERE DOES THE DATA RESIDE?

- Patient molecular diagnostic testing may be performed in an in-house pathology department laboratory, or on samples sent to an outside vendor (such as Foundation Medicine or Caris). It is common for both in-house and external testing sites to be used which may result in multiple sources of genomic data.
- To obtain in-house data requires either direct access to molecular testing results maintained by the pathology department, such as annotated VCF files, or results stored in the institution's clinical data warehouse or EMR system. Molecular diagnostic results may be obtained from external vendors in structured electronic format (e.g., XML or CSV file).
- Natural Language Processing (NLP) may be used on unstructured molecular diagnostic reports, if need be.

To date, integration of tumor genomic profiles with other patient clinical data has been successfully deployed for a number of TriNetX Research Network members, and tens of thousands of their patients.

REFERENCES

- [Gymrek 2013] GYMREK M, MCGUIRE AL, GOLAN D, HALPERIN E, ERLICH Y. Identifying personal genomes by surname inference. Science. 2013; 339: 321-324.
- [FSI 2007] FSI GENETICS. Forensic validation or the European SNPforID 52-plex assay. Available Online at: https:/www.fsigenetics.om/article/S1872-4973%2807%2900041-5abstract
- [NIH 2015] NATIONAL INSTITUTES OF HEALTH. Notice for use of cloud computing services for storage and analysis of controlled-access data subject to the NIH genomic data sharing (GDS) policy. NOT-OD-15-086. March 27, 2015.
- [OIRA 2015] OFFICE OF INFORMATION AND REGULATORY AFFAIRS, Office of Management and Budget, Executive Office of the President. Handout at meeting at the Council on Governmental Relations. March 27, 2015. Available online at: https:/www.reginfo.gov/public/do/viewEO12866Meeting?viwRule=true&rin=0937-AA02&meetinId=888&acro-nym=0937-HHS/OASH Last accessed: November 13, 2016.
- [PCB 2012] Presidential omission for the Study of Bioethical Issues. Privacy and progress in whole genome sequencing. October 2012. Available Online at: https:/bioethicsarchive.georgetown.edu/pcsbi/sites/default/files/rivacyProgress508_1.pdf
- [2014 GenSeq] MIT Technology Review. September 24, 2014. EmTech: Illumina Says 228,000 Human Genomes Will Be Sequenced This Year. https:/www.technologyreview.com/s/531091/emtech-illumina-says-228000-human-genomes-will-be-sequenced-this-year
- [2018 UK NHS WGS] GOV.UK 2 October 2018. Matt Hancock announces ambition to map 5 million genomes. https:/www.gov.uk/government/news/matt-hanock-announces-ambition-o-map-5-million-genomes