i2b2 Integration and Use of Molecular Diagnostic Genomic Assay Results for Clinical Research

Jack London PhD

Research Professor of Cancer Biology
Thomas Jefferson University
and Informatics Director
Sidney Kimmel Cancer Center
Philadelphia Pennsylvania USA

October 5, 2017
i2b2 2017 European meeting

Jefferson
HEALTH IS ALL WE DO

Jefferson University’s Evolution with Research Data Analytics

Jefferson research informatics timeline

2010

Quick start i2b2 pilot project completed by Recombinant Data Corporation (RDC)

2011

- Deployment of first phase Thomas Jefferson University Hospital (TJUH) Data Trust clinical data warehouse (CDW) and i2b2 research data mart (RDM).
- Expansion of RDM to include biospecimens and tumor registry ontologies and data

2012

Development of methodology for the prediction of clinical trial accrual using the Jefferson RDM

2013

- RDC starts expansion of TJUH CDW.
- Addition of i2b2 RDM omic data ontology and data started.

2014

- TriNetX project starts.
- Neuroscience Biobank project starts.
- Additional development of RDM omic data ontology and data .

2015

- TriNetX deployment includes tumor registry data.
- Neuroscience Biobank project.

2016

- TriNetX deployment includes genomics data.
- TriNetX collaborative network with UTSW.
- Project to include whole exome sequencing data started.

2017

- Exome data project continues.

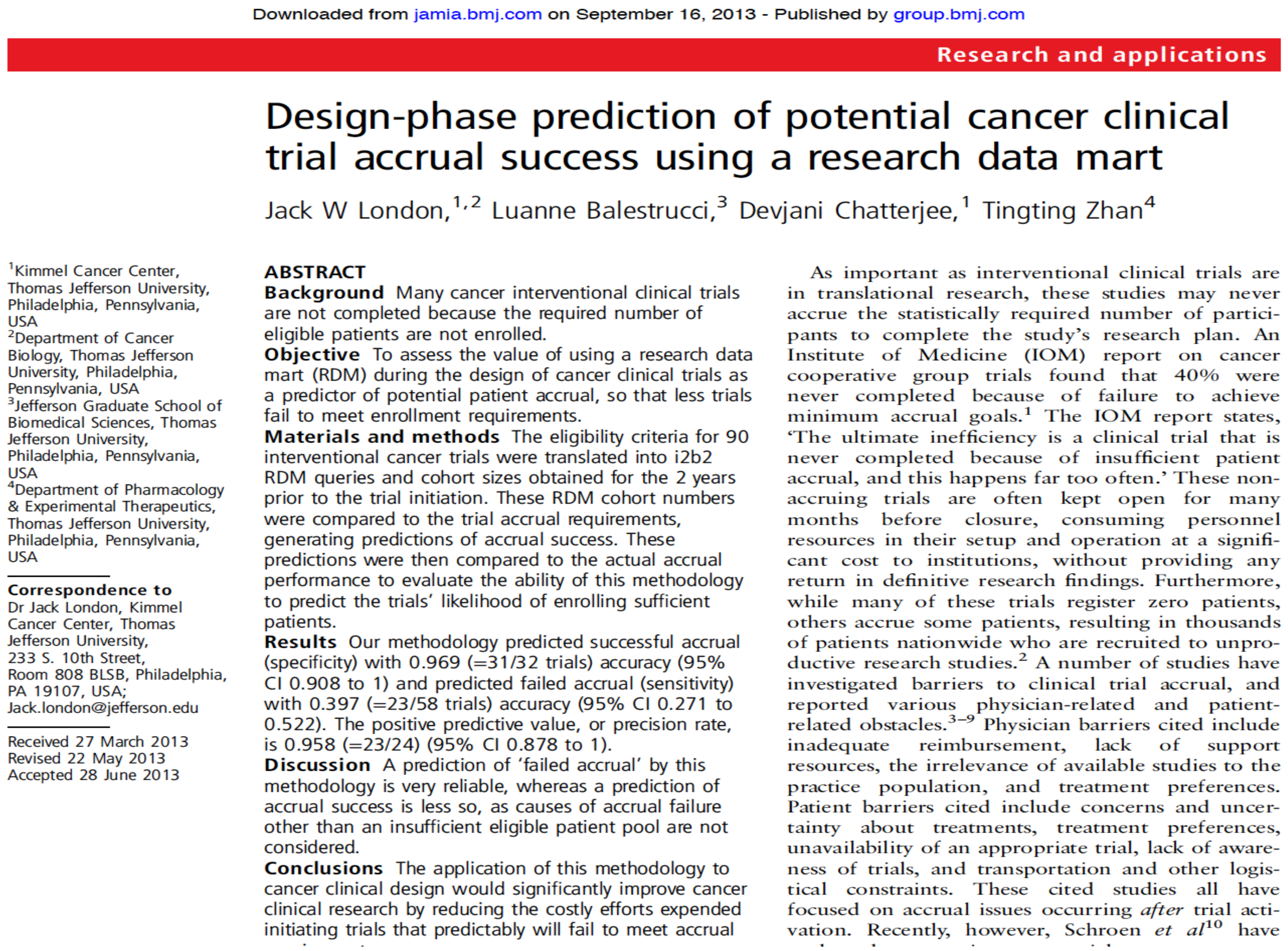
Jefferson's i2b2 research data mart deployment

In addition to **EMR patient data** (demographics, diagnoses, medications, labs, and procedures):

- Comprehensive data set for cancer patients from **Cancer Registry**
  - Tumor histology, stage, recurrence, treatment, disease-specific factors
- **"Omic" molecular diagnostic patient data** from Pathology A/P system and outsource vendor system
  - Currently > 350 genes with > 4,100 mutations (both in-house and outsourced Foundation Medicine results)
- **Biospecimen annotation** from biobanking system
  - Specimen anatomic origin, class, type, pathology, slide images

More than 125 million observations on 2.9 million patients, refreshed weekly

Jefferson Study Using a Data-Driven Approach to Clinical Trial Design



Downloaded from jamia.bmj.com on September 16, 2013 - Published by group.bmj.com

Research and applications

Design-phase prediction of potential cancer clinical trial accrual success using a research data mart

Jack W London,[1,2] Luanne Balestrucci,[3] Devjani Chatterjee,[1] Tingting Zhan[4]

[1]Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, Pennsylvania, USA
[2]Department of Cancer Biology, Thomas Jefferson University, Philadelphia, Pennsylvania, USA
[3]Jefferson Graduate School of Biomedical Sciences, Thomas Jefferson University, Philadelphia, Pennsylvania, USA
[4]Department of Pharmacology & Experimental Therapeutics, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

Correspondence to
Dr Jack London, Kimmel Cancer Center, Thomas Jefferson University,
233 S. 10th Street,
Room 808 BLSB, Philadelphia, PA 19107, USA;
Jack.london@jefferson.edu

Received 27 March 2013
Revised 22 May 2013
Accepted 28 June 2013

ABSTRACT

Background Many cancer interventional clinical trials are not completed because the required number of eligible patients are not enrolled.

Objective To assess the value of using a research data mart (RDM) during the design of cancer clinical trials as a predictor of potential patient accrual, so that less trials fail to meet enrollment requirements.

Materials and methods The eligibility criteria for 90 interventional cancer trials were translated into i2b2 RDM queries and cohort sizes obtained for the 2 years prior to the trial initiation. These RDM cohort numbers were compared to the trial accrual requirements, generating predictions of accrual success. These predictions were then compared to the actual accrual performance to evaluate the ability of this methodology to predict the trials' likelihood of enrolling sufficient patients.

Results Our methodology predicted successful accrual (specificity) with 0.969 (=31/32 trials) accuracy (95% CI 0.908 to 1) and predicted failed accrual (sensitivity) with 0.397 (=23/58 trials) accuracy (95% CI 0.271 to 0.522). The positive predictive value, or precision rate, is 0.958 (=23/24) (95% CI 0.878 to 1).

Discussion A prediction of 'failed accrual' by this methodology is very reliable, whereas a prediction of accrual success is less so, as causes of accrual failure other than an insufficient eligible patient pool are not considered.

Conclusions The application of this methodology to cancer clinical design would significantly improve cancer clinical research by reducing the costly efforts expended initiating trials that predictably will fail to meet accrual

As important as interventional clinical trials are in translational research, these studies may never accrue the statistically required number of participants to complete the study's research plan. An Institute of Medicine (IOM) report on cancer cooperative group trials found that 40% were never completed because of failure to achieve minimum accrual goals.[1] The IOM report states, 'The ultimate inefficiency is a clinical trial that is never completed because of insufficient patient accrual, and this happens far too often.' These non-accruing trials are often kept open for many months before closure, consuming personnel resources in their setup and operation at a significant cost to institutions, without providing any return in definitive research findings. Furthermore, while many of these trials register zero patients, others accrue some patients, resulting in thousands of patients nationwide who are recruited to unproductive research studies.[2] A number of studies have investigated barriers to clinical trial accrual, and reported various physician-related and patient-related obstacles.[3–9] Physician barriers cited include inadequate reimbursement, lack of support resources, the irrelevance of available studies to the practice population, and treatment preferences. Patient barriers cited include concerns and uncertainty about treatments, treatment preferences, unavailability of an appropriate trial, lack of awareness of trials, and transportation and other logistical constraints. These cited studies all have focused on accrual issues occurring *after* trial activation. Recently, however, Schroen *et al*[10] have

In 2012 we **used our research data mart** to determine recent cohort sizes of patients satisfying the eligibility criteria for a number of open trials at SKCC.

The study's hypothesis was that the RDM could have predicted whether a sufficient number of patients could be recruited for the trial during the trial's design phase.

Jefferson University's Evolution with Genomics Data Analytics

Precision Medicine's targeted therapies were the driver for integrating genomic data with other patient data.

- Clinical (trial) research is focused on determining genomic markers indicating diagnostic specificity, and predicative of treatment response and outcomes for personalized patient care.
- Inclusion of genomics data necessary for trial design and feasibility analyses.
- Integrated genomics/clinical data also a requirement for hypothesis generation (i.e., biomarker discovery).

Jefferson University's Evolution with Genomics Data Analytics

Consider a clinical trial for locally advanced colon patients "Neoadjuvant Treatment of Colon Cancer (NCT01108107)"

This study will investigate the effect of preoperative combination chemotherapy in patients with locally advanced colon cancer with mutation in the KRAS, BRAF or PIK3CA gene

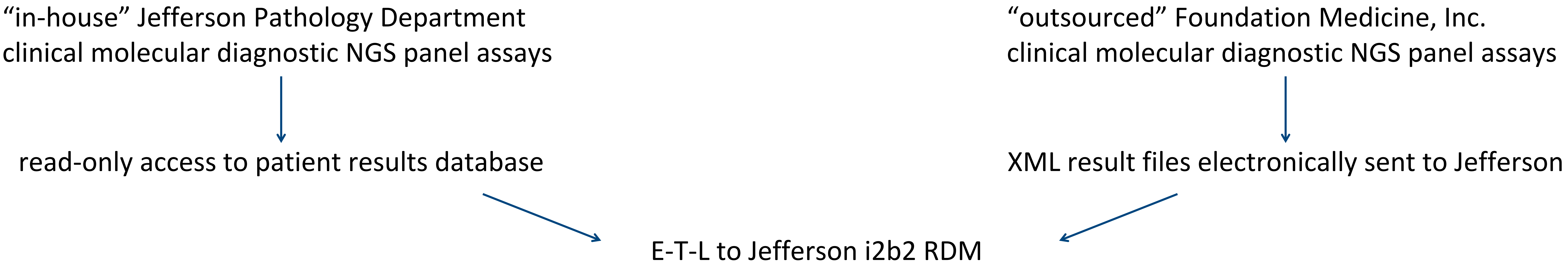
inclusion criteria

- stage 2 or 3 colon cancer
- KRAS, BRAF, and/or PIK3CA mutation testing determined in a CLIA-certified lab

exclusion criteria

- Clinically significant cardiovascular disease (including myocardial infarction, unstable angina, symptomatic congestive heart failure).

Genomic Clinical Data Sources



"in-house" Jefferson Pathology Department
clinical molecular diagnostic NGS panel assays

read-only access to patient results database

"outsourced" Foundation Medicine, Inc.
clinical molecular diagnostic NGS panel assays

XML result files electronically sent to Jefferson

E-T-L to Jefferson i2b2 RDM

Jefferson University’s Evolution with Genomics Data Analytics

Initial effort to integrate genomic data with other clinical data started with developing an i2b2 ontology for patients’ molecular diagnostic gene panel results.

All available sequencing data elements were included

- gene
- chromosome
- Start/end position
- reference/alternate allele
- DNA/amino acid change
- mutation type
- COSMIC id
- read depth
- aletrnate allele frequency
- …
- …

Molecular diagnostic assays – gene panels

Discussions with pathologists and oncologists brought the realization that – as in many other situations – the needed data classification is a function of the specific use case being addressed.

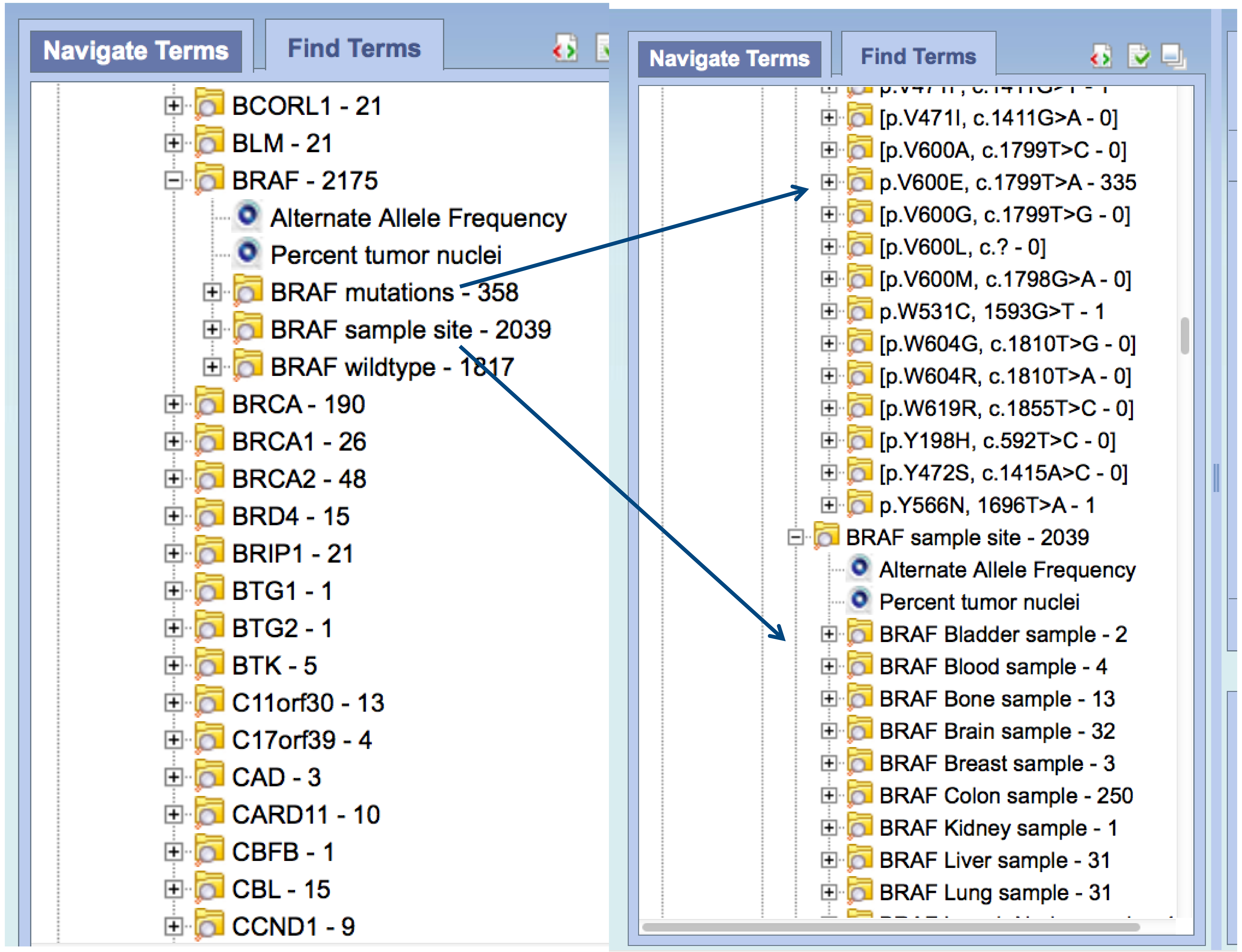
For **molecular diagnostic targeted gene panel data**, **used for clinical research cohort identification**, the necessary data elements are:

- tissue sampled
- gene tested
- mutations expressed (using standard mutation nomenclature based on coding DNA reference sequences and protein-level amino acid sequences – “c.” and “p.”), or wild type).

Such as, **colon tumor, BRAF, p.V600E, c.1799T>A**

These data elements correspond to clinical trial eligibility criteria, and are usually sufficient for mutation identification in databases/literature so that further information can be obtained.

SKCC’s i2b2 “omic” ontology



BCORL1 - 21
BLM - 21
BRAF - 2175
Alternate Allele Frequency
Percent tumor nuclei
BRAF mutations - 358
BRAF sample site - 2039
BRAF wildtype - 1817
BRCA - 190
BRCA1 - 26
BRCA2 - 48
BRD4 - 15
BRIP1 - 21
BTG1 - 1
BTG2 - 1
BTK - 5
C11orf30 - 13
C17orf39 - 4
CAD - 3
CARD11 - 10
CBFB - 1
CBL - 15
CCND1 - 9

[p.V471I, c.1411G>A - 0]
[p.V600A, c.1799T>C - 0]
p.V600E, c.1799T>A - 335
[p.V600G, c.1799T>G - 0]
[p.V600L, c.? - 0]
[p.V600M, c.1798G>A - 0]
p.W531C, 1593G>T - 1
[p.W604G, c.1810T>G - 0]
[p.W604R, c.1810T>A - 0]
[p.W619R, c.1855T>C - 0]
[p.Y198H, c.592T>C - 0]
[p.Y472S, c.1415A>C - 0]
p.Y566N, 1696T>A - 1
BRAF sample site - 2039
Alternate Allele Frequency
Percent tumor nuclei
BRAF Bladder sample - 2
BRAF Blood sample - 4
BRAF Bone sample - 13
BRAF Brain sample - 32
BRAF Breast sample - 3
BRAF Colon sample - 250
BRAF Kidney sample - 1
BRAF Liver sample - 31
BRAF Lung sample - 31

Other genomic data considerations

- When extracting data from some molecular diagnostic sources, thresholds need to be set for the alternate allele frequency and percent tumor of valid variant results.

- The genes tested for an assay need to be known, since all genes that are wild type are not necessarily reported. (Pertinent negatives are sometimes reported; also results for ordered genes may only be reported.)

- The molecular diagnostic ontology (i.e., taxonomy) is not static. Variants not previously reported may occur at any time; genes may be added to test panels at any time (ideally you would be alerted in advance to test panel changes).

Integration of genomic data with other patient data

For **hypothesis generation** – perhaps for **biomarker discovery** – further data capabilities may also be needed.

- inclusion of high-dimensional data
- statistical pipeline (“R”) and other analysis tools (genomic data viewer)
- additional of public data sets (TCGA)

This type of environment found on tranSMART.

Targeted therapies also a driver for data sharing

- About half of therapeutic oncology trials investigate targeted approaches.

- The small patient sub-populations needed for these studies increase the likelihood that multi-institutional trials are required (and that networked, inter-operative data analytics platforms are available).

Jefferson University's Evolution with Genomics Data Analytics

Work now "in progess"

Since last summer, we are working on how to process and integrate whole exome sequencing data, particularly from a biomarker discovery perspective.

The Jefferson i2b2 RDM team

Jack London, PhD

SKCC Informatics Director

Stephen Peiper, MD

Chair, Department of Pathology

Robert Stapp, MD

Jefferson Department of Pathology

Chirayu Goswami, MS

Bioinformatician

John Reber

Systems Administrator/Programmer

Funding: NCI/NIH grant # P30CA056036