i2b2

A tool for research data analytics

Jack London, PhD
CI4CC meeting
March 13, 2016

Jefferson
HEALTH IS ALL WE DO

Disclaimer

In addition to my faculty position at Thomas Jefferson University in Philadelphia, I am a consultant for TriNetX Corporation.

Research data analytics use cases

- Hypothesis generation

  Example: An investigator wishes to explore possible links between BRAF mutations and response to treatment for colorectal cancer.

- Cohort identification

  Example: A basic scientist needs to know if there are sufficient tissue specimens from Asian women with “triple negative” breast cancer for a biomarker study.

  Example: A clinical researcher wishes to assess potential patient accrual for a trial under design by obtaining number of patients seen in the recent past that meet the proposed eligibility criteria.

Jefferson

Research Data Marts

- Research data marts (RDM) are patient data warehouses focused on the needs of researchers.
- An RDM can aggregate data from clinical (e.g., EMR) and research-related (e.g., study biobanks) sources into one integrated data repository.
- An RDM can address issues specific to the research domain, such as being HIPAA-compliant by being having only de-identified data, and by providing obfuscated query results when necessary.

i2b2 Research Data Marts

These are research data repositories built on the "informatics for integrating biology and the bedside" (i2b2) framework, developed at the NIH-funded National Center for Biomedical Computing based at Partners HealthCare System (Harvard).

This platform has been deployed at many academic medical centers.

Jefferson

Paper describing i2b2 platform

Model formulation

EDITOR'S CHOICE

Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)

Shawn N Murphy,[1,3] Griffin Weber,[2,6] Michael Mendis,[3] Vivian Gainer,[3] Henry C Chueh,[1] Susanne Churchill,[3] Isaac Kohane[4,5]

[1]Laboratory of Computer Science, Massachusetts General Hospital, Boston, Massachusetts, USA
[2]Harvard Medical School, Boston, Massachusetts, USA
[3]Information Systems, Partners HealthCare System, Inc., Wellesley, Massachusetts, USA
[4]Children's Hospital, Boston, Massachusetts, USA
[5]Brigham and Women's Hospital, Boston, Massachusetts, USA
[6]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

Correspondence to
Shawn N Murphy, Information Systems, Partners HealthCare System, Inc., One Constitution Center, Charlestown, MA 02129, USA; murphy.shawn@mgh.harvard.edu

Received 13 August 2009
Accepted 23 December 2009

ABSTRACT
Informatics for Integrating Biology and the Bedside (i2b2) is one of seven projects sponsored by the NIH Roadmap National Centers for Biomedical Computing (http://www.ncbcs.org). Its mission is to provide clinical investigators with the tools necessary to integrate medical record and clinical research data in the genomics age, a software suite to construct and integrate the modern clinical research chart. i2b2 software may be used by an enterprise's research community to find sets of interesting patients from electronic patient medical record data, while preserving patient privacy through a query tool interface. Project-specific mini-databases ("data marts") can be created from these sets to make highly detailed data available on these specific patients to the investigators on the i2b2 platform, as reviewed and restricted by the Institutional Review Board. The current version of this software has been released into the public domain and is available at the URL: http://www.i2b2.org/software.

INTRODUCTION
Many challenges exist when it comes to repurposing data from an electronic medical record

BACKGROUND
The repurposing of medical record data for clinical research holds high promise.[1 5] Potentially such data are highly useful for research, representing some of the most important everyday clinical events of patients' lives as recorded by trained observers. If the adoption of EMRSs is to increase as anticipated,[6] it is incumbent and opportune to develop methods for providing ways to look at this data across patients. However, this task is much more difficult than would first appear. EMRSs are typically built to look at data on single patients, not data across combinations of many patients. Attempts to overlay this functionality on existing EMRSs demonstrate that the functional and technical requirements of the transactional and analytical systems are in opposition.[7]

Unlike transaction systems that are optimized to show data regarding single patients, a system that supports queries that cut across multiple patients is more dependent on standard descriptors and annotations, queries can be challenging to specify, and these queries have complex implications for the privacy of the patients. Furthermore, attempting to "fit together" medical record data and clinical trial

Significant points about the i2b2 infrastructure

- The i2b2 data model is based on the “star schema”
- The star schema has a central “fact” table where each row represents a single observation about a patient.
- Observations are regarding a specific concept, such as a lab test or disease diagnosis.
- By expressing a concept as an attribute in a row rather than designating it in a column is known as the entity-attribute-value (EAV) model.

=> It is extremely efficient to query data arranged in a star schema represented in an EAV format.

Jefferson

Jefferson's i2b2 Research Data Mart

- Built on "informatics for integrating biology and the bedside" (i2b2) framework.
- RDM data are de-identified. Re-identification possible via an honest broker, who has access to a re-identification application.
- Currently ~ 100 million observations on > 1 million patients. Data refreshed weekly.

Patient data obtained from TJUH EMR

DEMOGRAPHICS

- Age
- Ethnicity
- Gender
- Race
- Vital Status (alive/dead)

DIAGNOSES

- Disease systems --> diseases (organized by ICD9 and ICD10 coding)

CLINICAL LAB RESULTS

- Chemistry
- Coagulation
- Hematology

MEDICATIONS

INPATIENT PROCEDURES

- Diagnostic and Treatment procedures (organized by ICD9 and CPT coding)

Jefferson

Example list of patient mutation data obtained from in-house and Foundation Medicine molecular diagnostic testing

ALK rearrangement
ALK c.4186G>A, p.A1396T
ALK c.3745G>C, p.D1294H

BRAF c.1782T>G p.D594E
BRAF c.1801A>G p.K601E
BRAF c.1799T>A p.V600E

EGFR Deletion in exon 19
EGFR Insertion in exon 20
EGFR c.2236G>A p.E746K
EGFR c.2236_2250del15
p.E746_A750delELREA
EGFR c.2156G>C p.G719A
EGFR c.2155G>T p.G719C
EGFR c.2155G>A p.G719S
EGFR c.2573T>G p.L858R
EGFR c.2582T>A p.L861Q
EGFR c.2303G>T p.S768I

JAK2 c.1849G>T p.V617F

JAK3 c.2164G>A p.V722I

KRAS c.35G>C p.G12A
KRAS c.34G>T p.G12C
KRAS c.35G>A p.G12D
KRAS c.34G>C p.G12R
KRAS c.34G>A p.G12S
KRAS c.35G>T p.G12V
KRAS c.38G>A p.G13D

NRAS c.183A>T p.Q61H
NRAS c.181C>A p.Q61K
NRAS c.182A>T p.Q61L
NRAS c.182A>G p.Q61R

PIK3CA c.1633G>A p.E545K
PIK3CA c.3140A>T p.H1047L
PIK3CA c.3140A>G p.H1047R

PTEN c.754G>T p.D252Y
PTEN c.59G>A p.G20E

RET rearrangement

ROS1 rearrangement

SMAD4 c.1157G>A p.G386D

TP53 c.843C>A p.D281E
TP53 c.811G>T p.E271*
TP53 c.857A>C p.E286A
TP53 c.400T>C p.F134L
TP53 c.734G>A p.G245D
TP53 c.388C>G p.L130V
TP53 c.524G>A p.R175H
TP53 c.817C>T p.R273C
TP53 c.818G>A p.R273H
TP53 c.318C>G p.S106R
TP53 c.659A>G p.Y220C
TP53 c.707A>G p.Y236C

PIK3CA c.3140A>G p.H1047RC
PIK3CA c.1637A>G p.Q546R

As of March 2016, Jefferson omic metadata includes 336 genes with 3,060 mutations

Jefferson

Specimen annotation from campus biobanks

Eight biobanks, including the TJUH paraffin block archive of ~400,000 cases since 1990.

Anatomic origin (SNOMED)

Class (tissue, fluid)

Type (frozen, FFPE)

Pathology (normal, malignant, diseased)

Slide images

Jefferson

Patient data from Jefferson Tumor Registry

Over 100,000 cases since 1990.

Primary Cancer Diagnosis

- Age at diagnosis/date of diagnosis
- Survival (months) from diagnosis
- Tumor histology and behavior
- Stage (AJCC/TNM, clinical and pathological)
- Grade

Recurrence

- local, distant

Treatment

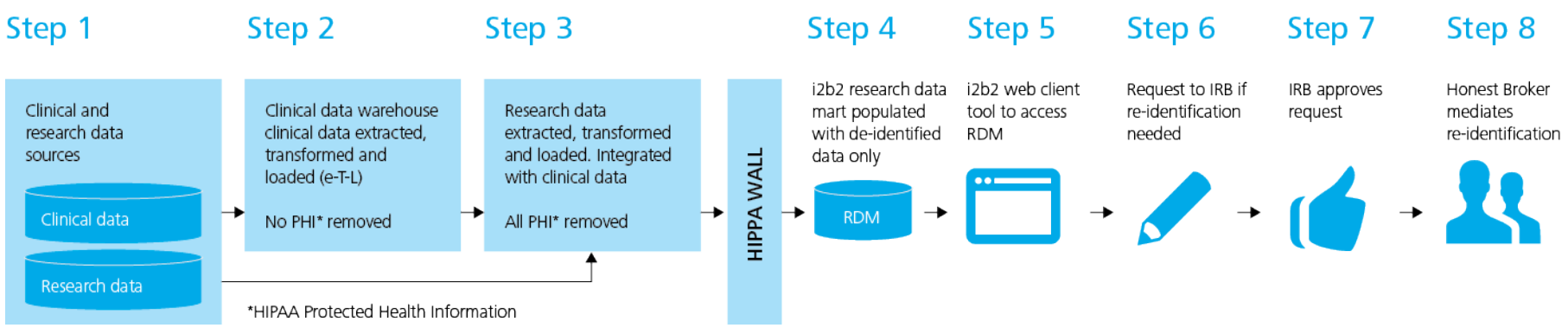
- chemotherapy, radiation, surgery, transplant, palliative

Disease-specific factors

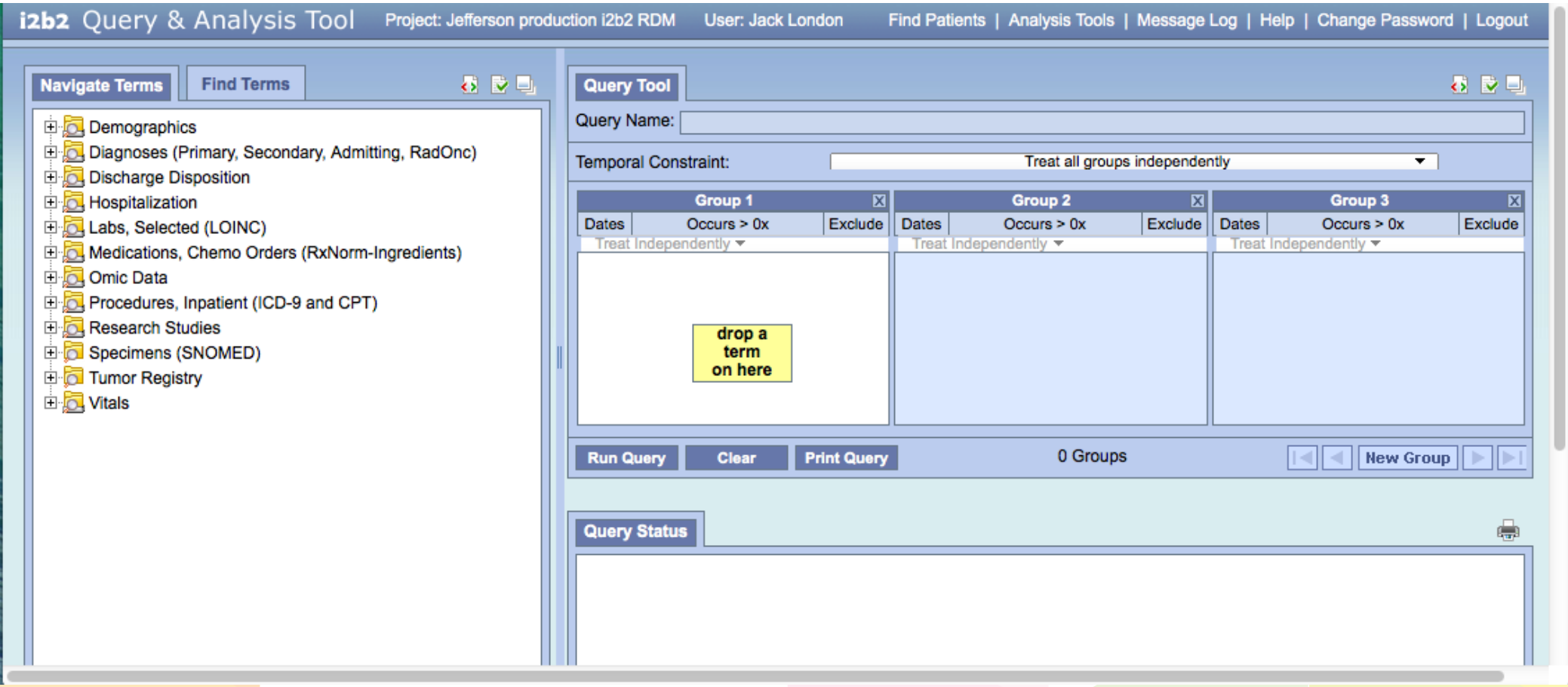
- ex: (prostate --> Gleason score)

Jefferson

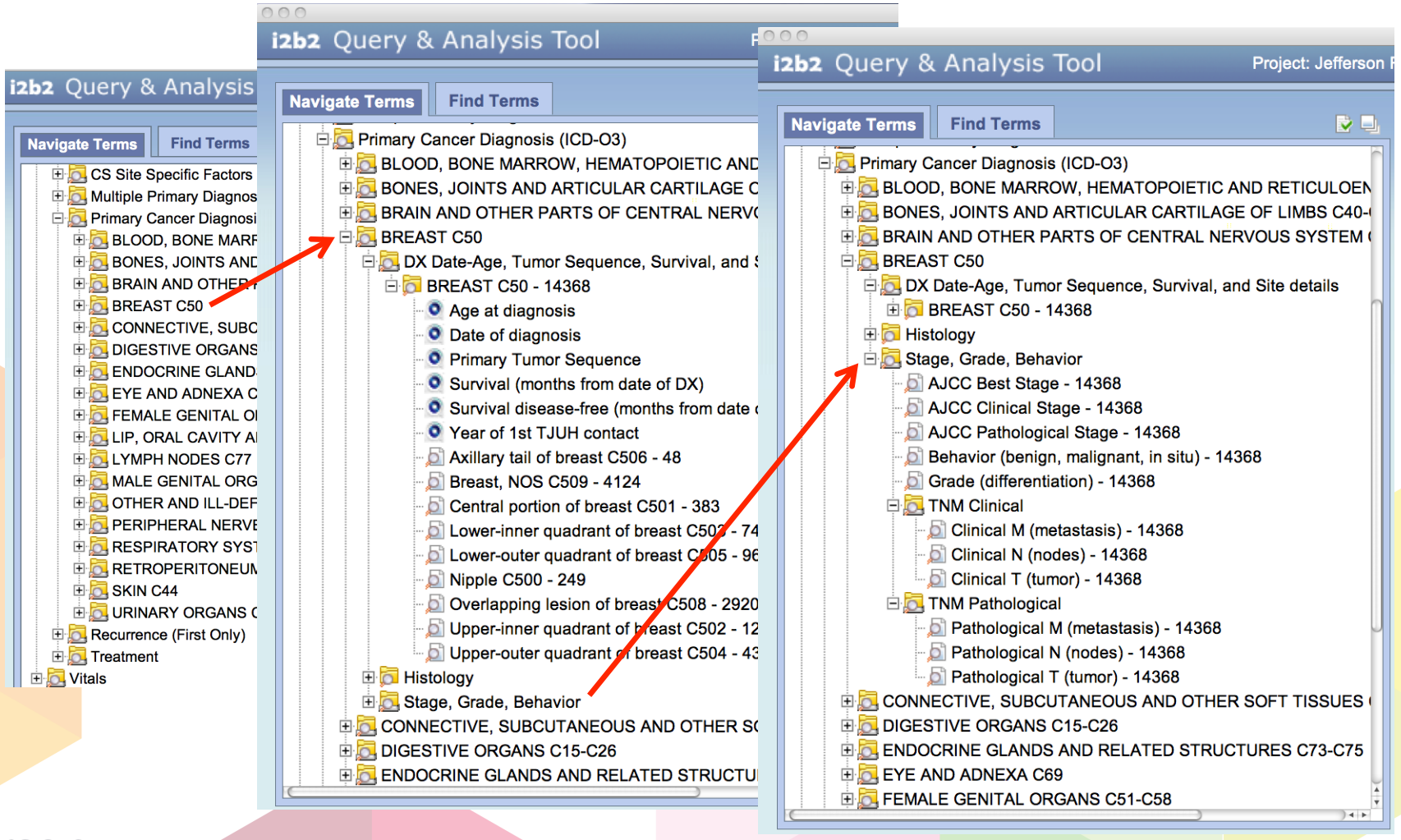
Research data work flow



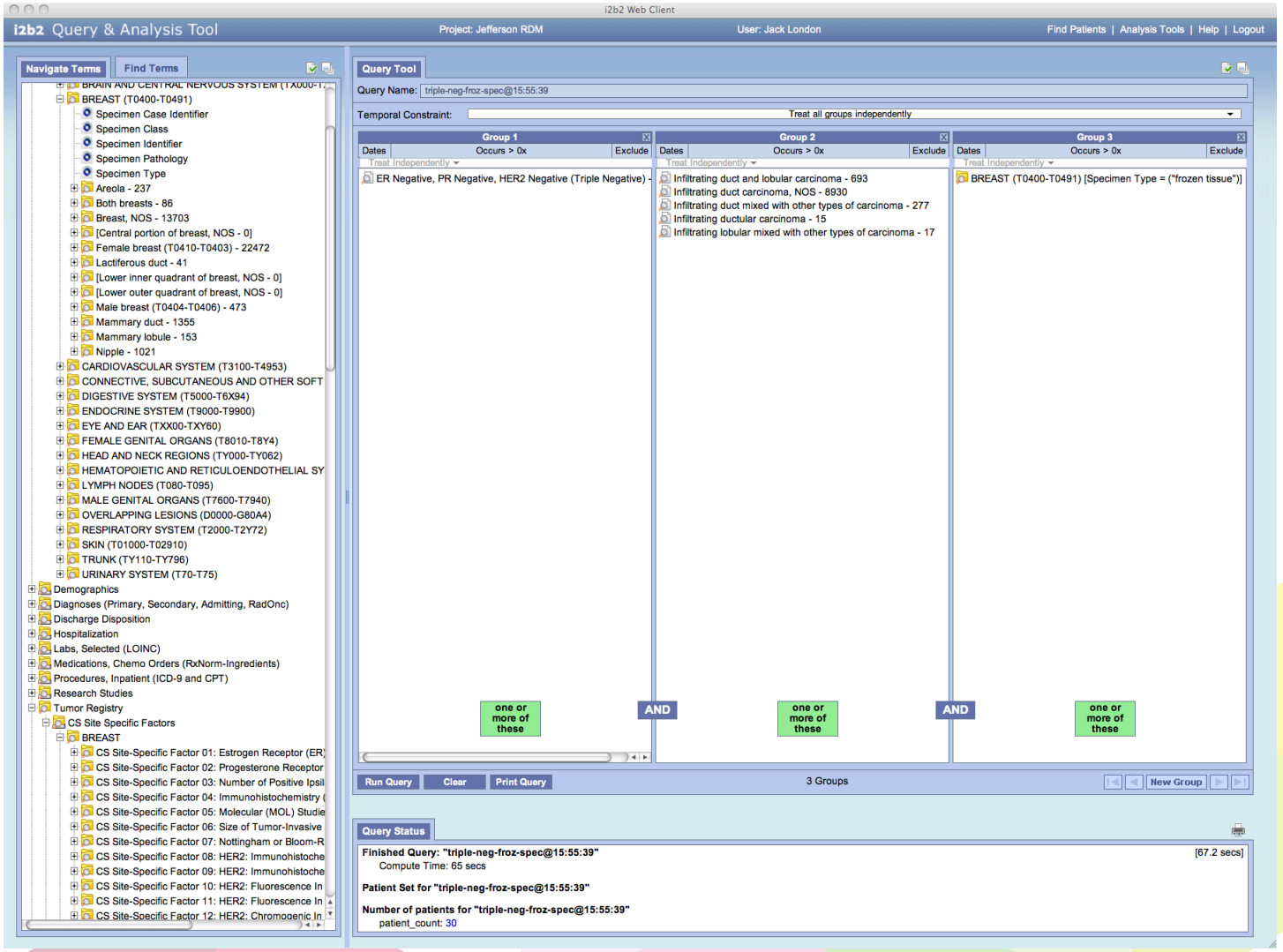
Drag-and-drop i2b2 query tool



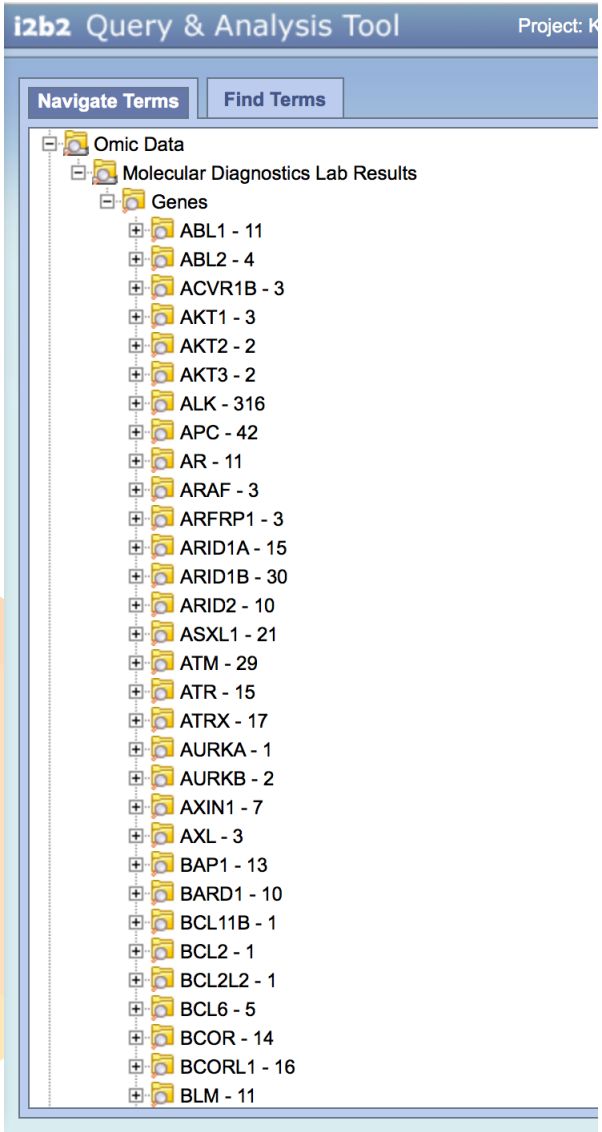
Taxonomy for tumor registry data

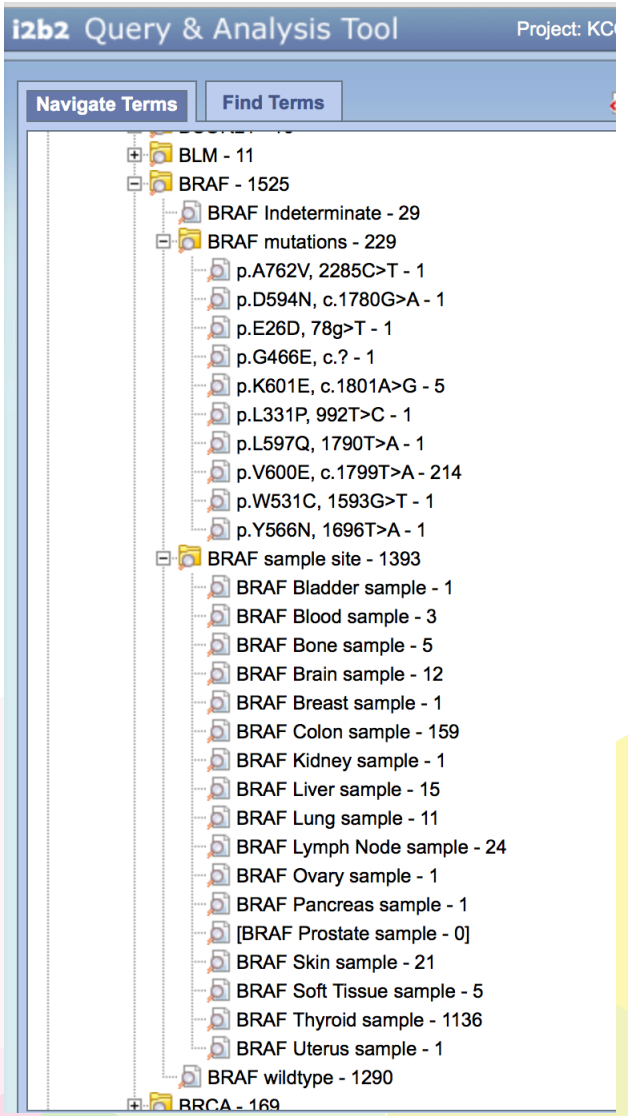


Identification of patient cohorts or hypothesis generation



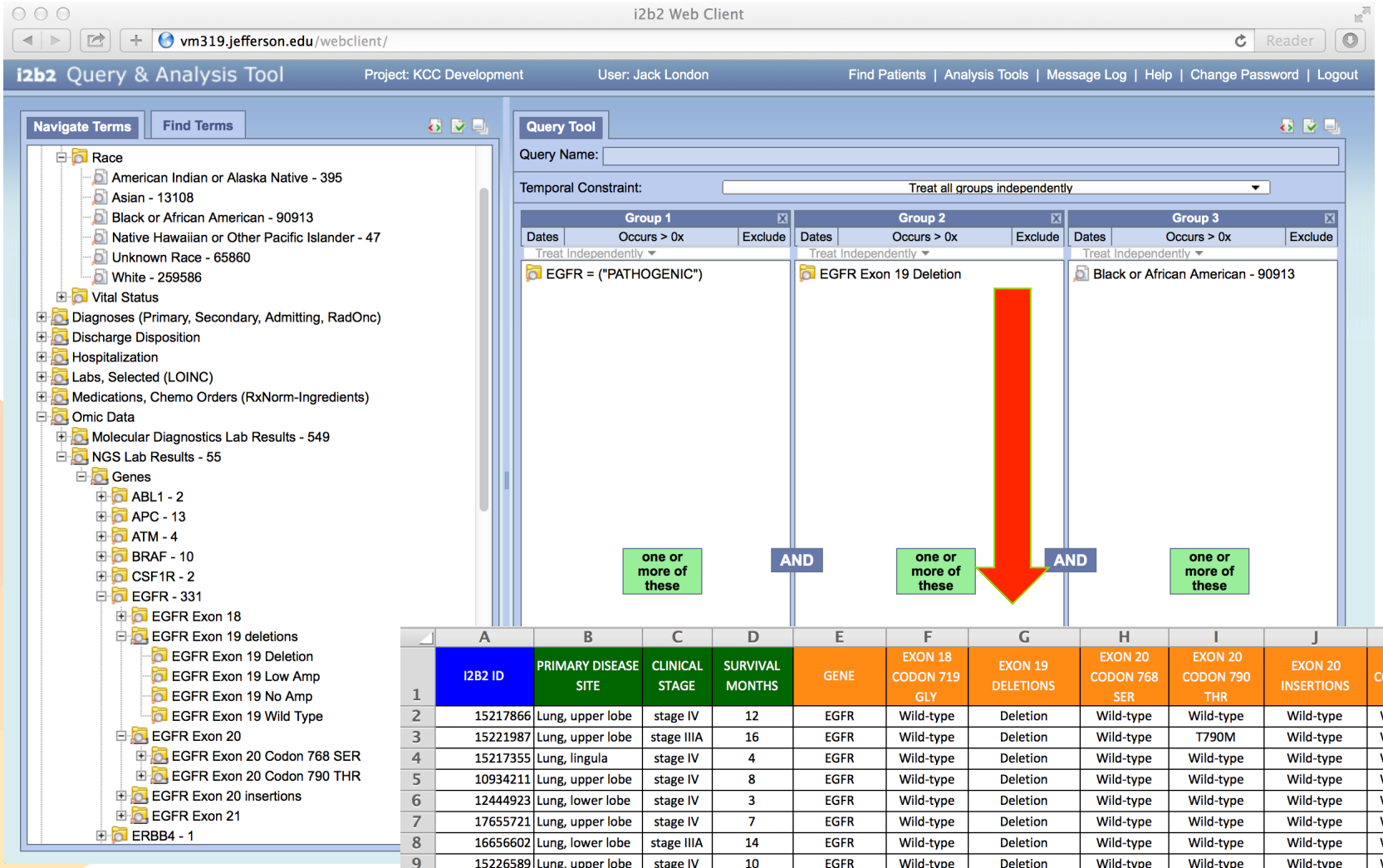
Molecular Diagnostic data





Jefferson

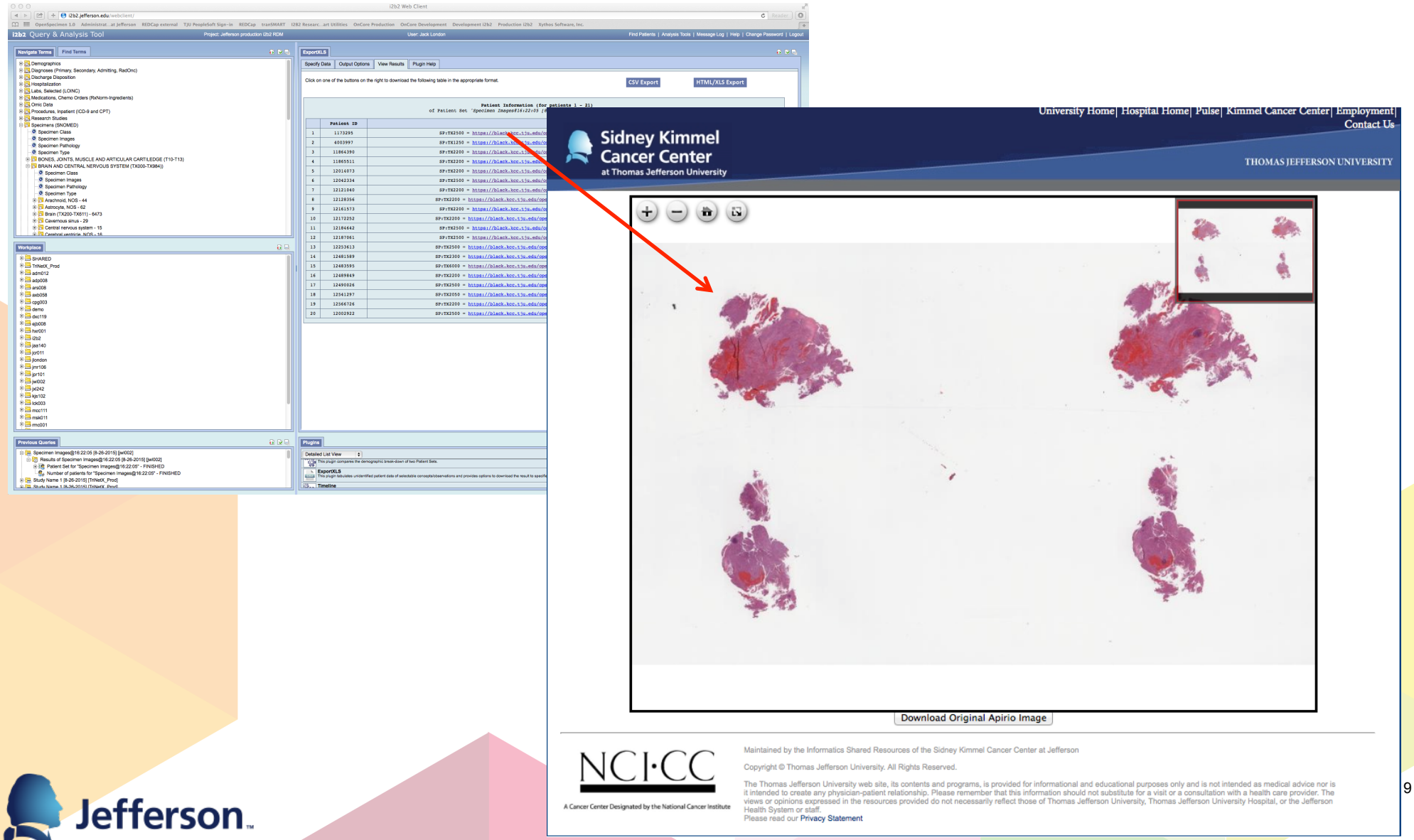
Additional data on selected cohort can be retieved



	A	B	C	D	E	F	G	H	I	J	K	L
1	I2B2 ID	PRIMARY DISEASE SITE	CLINICAL STAGE	SURVIVAL MONTHS	GENE	EXON 18 CODON 719 GLY	EXON 19 DELETIONS	EXON 20 CODON 768 SER	EXON 20 CODON 790 THR	EXON 20 INSERTIONS	EXON 21 CODON 858 LEU	EXON 21 CODON 861 LEU
2	15217866	Lung, upper lobe	stage IV	12	EGFR	Wild-type	Deletion	Wild-type	Wild-type	Wild-type	Wild-type	Wild-type
3	15221987	Lung, upper lobe	stage IIIA	16	EGFR	Wild-type	Deletion	Wild-type	T790M	Wild-type	Wild-type	Wild-type
4	15217355	Lung, lingula	stage IV	4	EGFR	Wild-type	Deletion	Wild-type	Wild-type	Wild-type	Wild-type	Wild-type
5	10934211	Lung, upper lobe	stage IV	8	EGFR	Wild-type	Deletion	Wild-type	Wild-type	Wild-type	Wild-type	Wild-type
6	12444923	Lung, lower lobe	stage IV	3	EGFR	Wild-type	Deletion	Wild-type	Wild-type	Wild-type	Wild-type	Wild-type
7	17655721	Lung, upper lobe	stage IV	7	EGFR	Wild-type	Deletion	Wild-type	Wild-type	Wild-type	Wild-type	Wild-type
8	16656602	Lung, lower lobe	stage IIIA	14	EGFR	Wild-type	Deletion	Wild-type	Wild-type	Wild-type	Wild-type	Wild-type
9	15226589	Lung, upper lobe	stage IV	10	EGFR	Wild-type	Deletion	Wild-type	Wild-type	Wild-type	Wild-type	Wild-type
10	18660745	Lung, upper lobe	stage IIIA	19	EGFR	Wild-type	Deletion	Wild-type	Wild-type	Wild-type	Wild-type	Wild-type
11	1966663	Lung, lower lobe	stage IV	9	EGFR	Wild-type	Deletion	Wild-type	Wild-type	Wild-type	Wild-type	Wild-type

Jefferson

Pathology images are available via *i2b2* query tool



Jefferson

Why data analytics?

Old School

Sabermetrics

Cohort definition via i2b2 can be used to predict accrual for proposed clinical trials

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

Overall result of this study

Our results show that the methodology, while having an excellent positive predictive value (95.8%, predicted failure for 23 of the 24 trials that actually failed), is not good at predicting failed accrual (39.7%, 23/58 trials).

In other words: if the methodology predicts "failed accrual," then we should trust this prediction and should not proceed to open the trial with its current eligibility criteria.

However, a prediction of accrual success using this method is no guarantee that target goals will be met, since other factors (e.g., competing trials) exist in addition to patient population considerations.

Jefferson SKCC experience with i2b2

- Cancer center initial deployment of i2b2 was in 2010
  - Hospital had contracted for a proprietary clinical data warehouse whose vendor supported i2b2 data mart deployment
  - Open source preferable to proprietary solutions
    - interoperablity with other academic centers
    - cost effective
- Support through the i2b2 Academic Users Group has been outstanding
- Major drawback to i2b2 query tool is the lack of data visualization capability.

Jefferson

Extensions of i2b2 database use

- tranSMART
- TriNetX

tranSMART

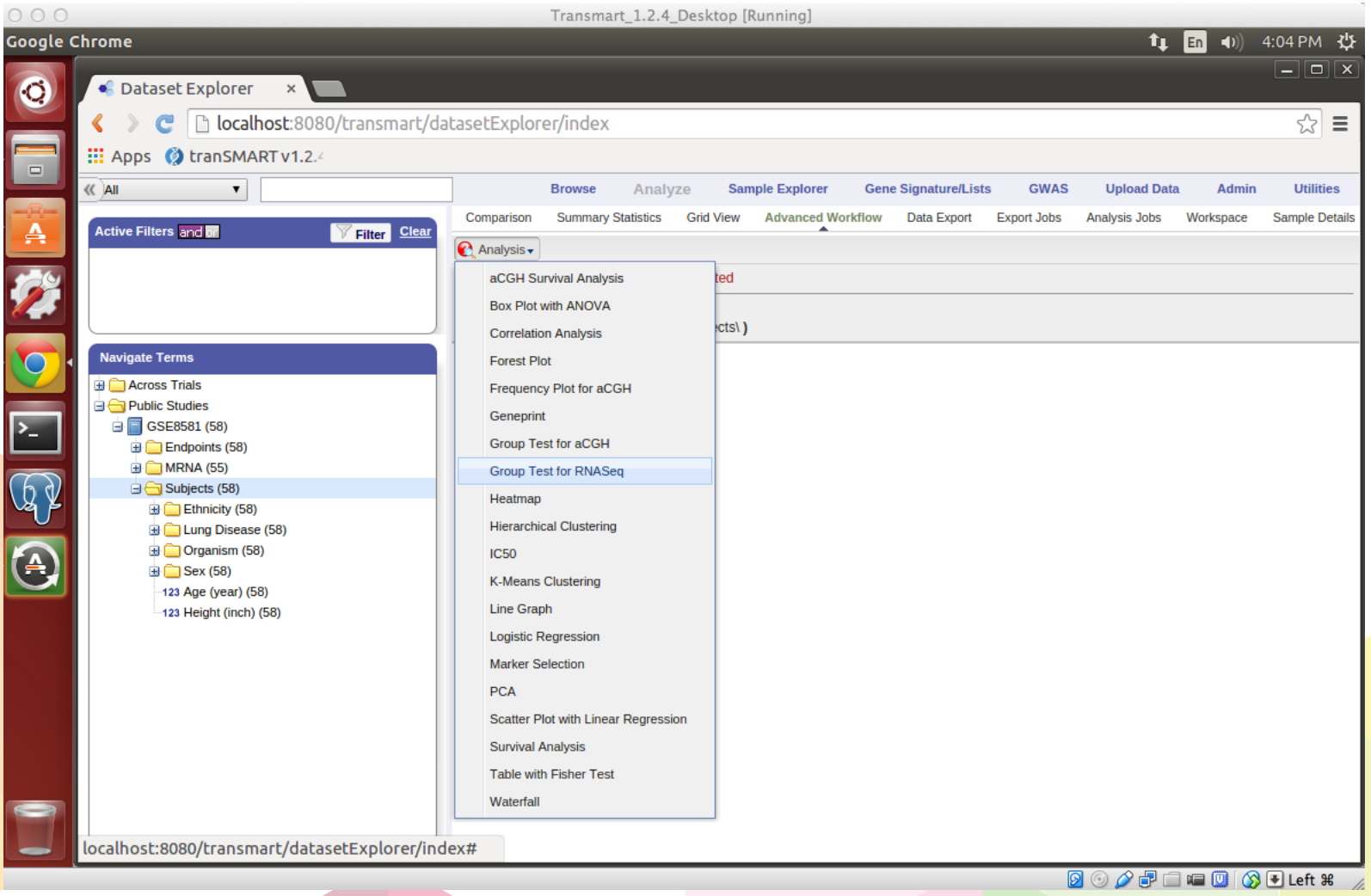
- tranSMART is a knowledge management platform, built on i2b2, that has statistical analysis pipeline capabilities, as well as an IGV pipeline for high dimensional data.

- The initial version of tranSMART's data management system was developed in 2009 by scientists at Johnson & Johnson and Recombinant Data Corporation.

- Established in 2013, the tranSMART Foundation is a public-private partnership – the result of collaborations between scientists in the United States and the European Union. Founding partners include the University of Michigan, the Pistoia Alliance and Imperial College London.

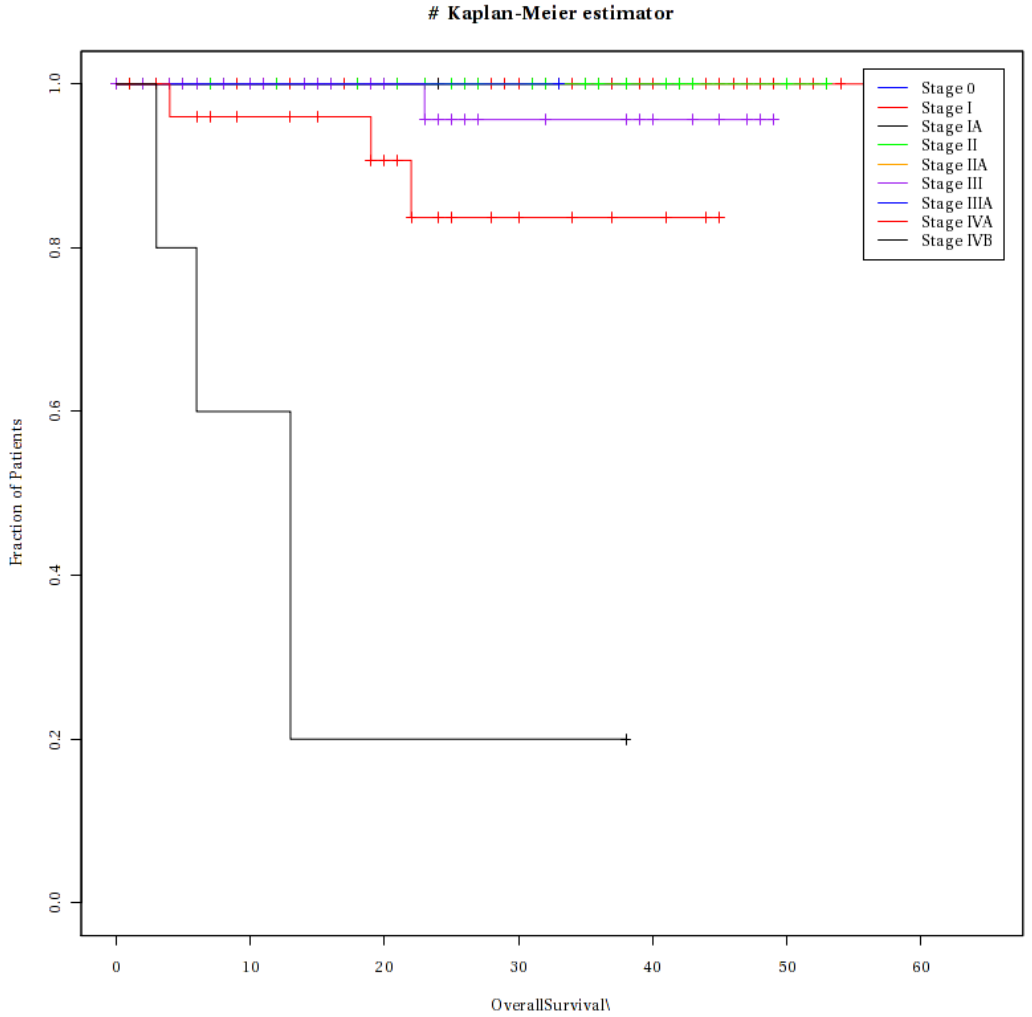
Jefferson

tranSMART statistical analyses



Example of Kaplan-Meier plot from tranSMART

SURVIVAL CURVE (STAGE) OF TJUH PATIENTS WITH THYROID SPECIMENS



Jefferson

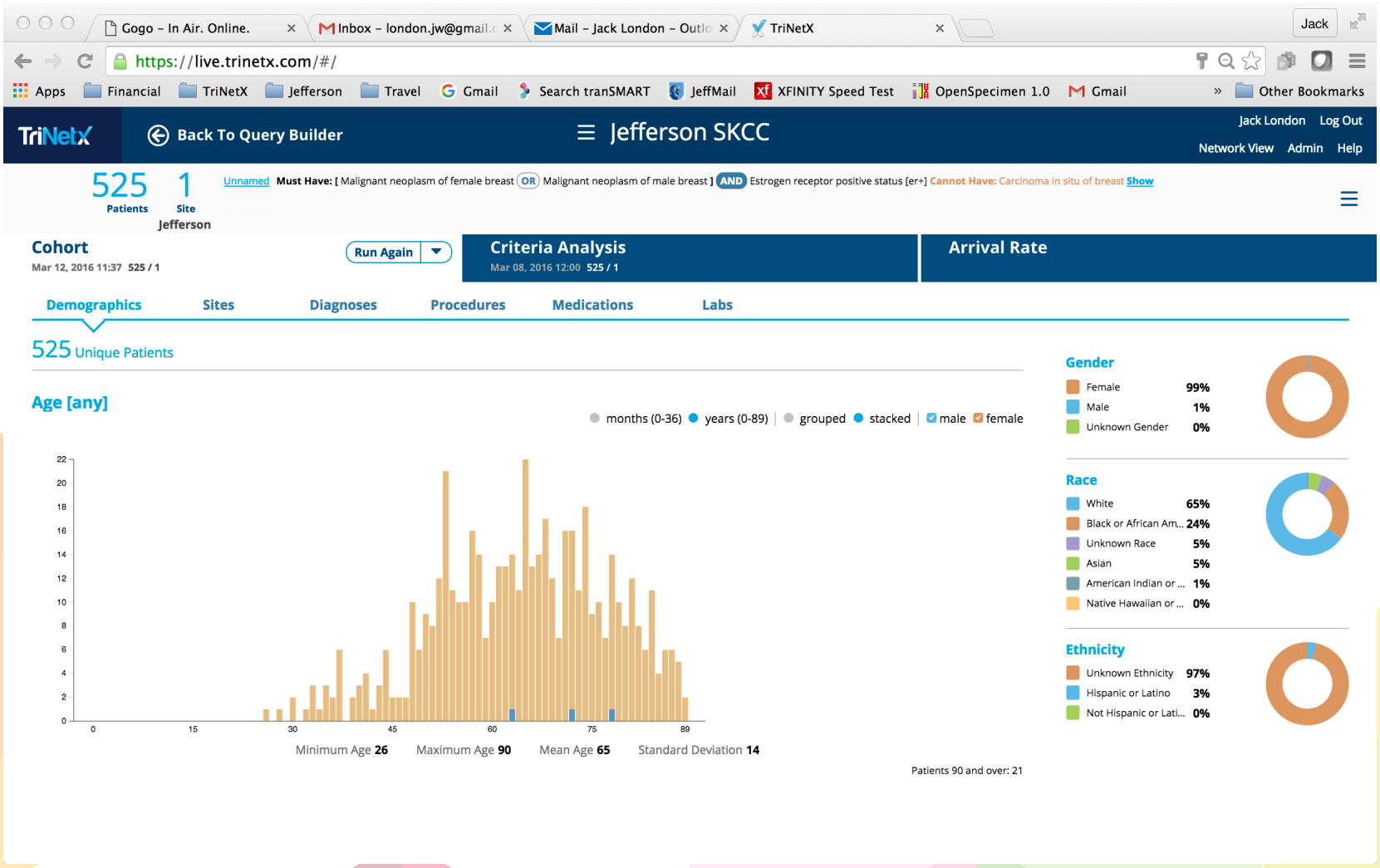
Jefferson – TriNetX project

TriNetX facilitates clinical trial collaboration between pharmaceutical companies and academic medical center data providers by providing access to aggregate data from academic members of the TriNetX network.

The TriNetX application provides advanced visualization of the data in the institution's i2b2 database.

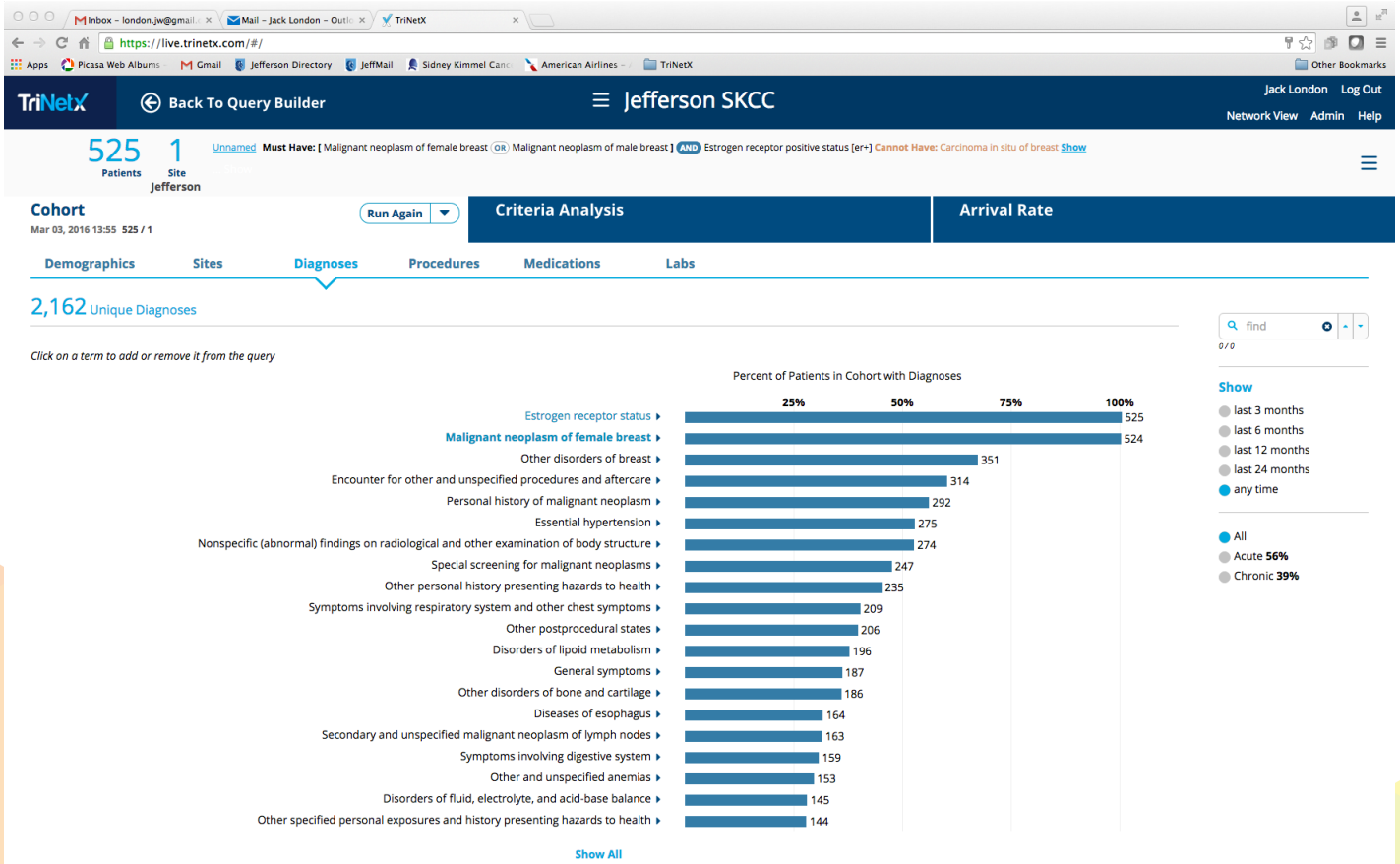
Data sharing between academic members is facilitated since data harmonization to the TriNetX model has already been done.

TriNetX display of cohort demographics

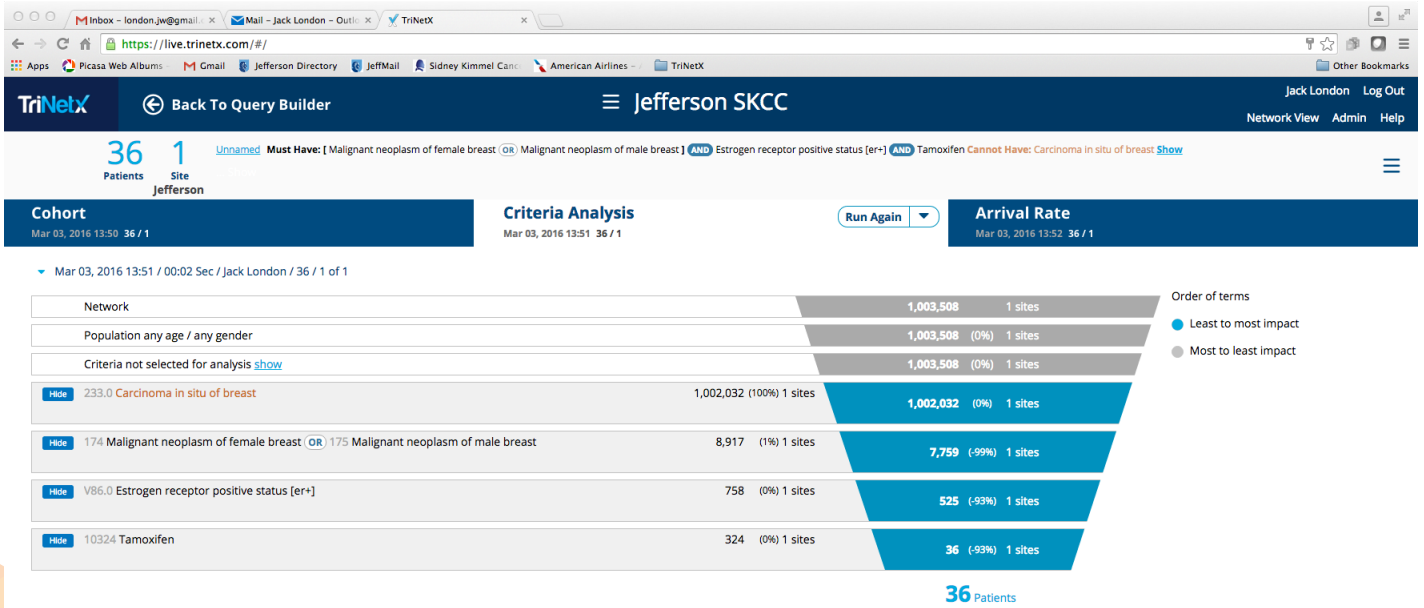


Jefferson

TriNetX display of cohort co-morbidities



TriNetX display of cohort criteria analysis



Questions?

Jefferson