

# i2b2

## A tool for research data analytics

Jack London, PhD  
CI4CC meeting  
March 13, 2016





## 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.



# 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.



# Paper describing i2b2 platform

Downloaded from [jamia.bmjjournals.org](http://jamia.bmjjournals.org) on March 11, 2010 - Published by [group.bmjjournals.org](http://group.bmjjournals.org)

## Model formulation



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

Shawn N Murphy,<sup>1,3</sup> Griffin Weber,<sup>2,6</sup> Michael Mendis,<sup>3</sup> Vivian Gainer,<sup>3</sup> Henry C Chueh,<sup>1</sup> Susanne Churchill,<sup>3</sup> Isaac Kohane<sup>4,5</sup>

<sup>1</sup>Laboratory of Computer Science, Massachusetts General Hospital, Boston, Massachusetts, USA

<sup>2</sup>Harvard Medical School, Boston, Massachusetts, USA

<sup>3</sup>Information Systems, Partners HealthCare System, Inc., Wellesley, Massachusetts, USA

<sup>4</sup>Children's Hospital, Boston, Massachusetts, USA

<sup>5</sup>Brigham and Women's Hospital, Boston, Massachusetts, USA

<sup>6</sup>Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

**Correspondence to**  
Shawn N Murphy, Information Systems, Partner HealthCare System, Inc., One Constitution Center, Charlestown, MA 02129, USA; [murphy.shawn@mgh.harvard.edu](mailto: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.ncbi.ncicb.nih.gov>). 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.<sup>1–5</sup> 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 EMRs is to increase as anticipated,<sup>6</sup> 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. EMRs are typically built to look at data on single patients, not data across combinations of many patients. Attempts to overlay this functionality on existing EMRs demonstrate that the functional and technical requirements of the transactional and analytical systems are in opposition.<sup>7</sup>

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



**Jefferson**™



# 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'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)



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

|      |                      |        |               |          |        |           |           |
|------|----------------------|--------|---------------|----------|--------|-----------|-----------|
| ALK  | rearrangement        | KRAS   | c.35G>C       | p.G12A   | TP53   | c.843C>A  | p.D281E   |
| ALK  | c.4186G>A, p.A1396T  | KRAS   | c.34G>T       | p.G12C   | TP53   | c.811G>T  | p.E271*   |
| ALK  | c.3745G>C, p.D1294H  | KRAS   | c.35G>A       | p.G12D   | TP53   | c.857A>C  | p.E286A   |
| BRAF | c.1782T>G            | BRAF   | c.34G>C       | p.G12R   | TP53   | c.400T>C  | p.F134L   |
| BRAF | c.1801A>G            | BRAF   | c.34G>A       | p.G12S   | TP53   | c.734G>A  | p.G245D   |
| BRAF | c.1799T>A            | BRAF   | c.35G>T       | p.G12V   | TP53   | c.388C>G  | p.L130V   |
|      |                      | BRAF   | c.38G>A       | p.G13D   | TP53   | c.524G>A  | p.R175H   |
| EGFR | Deletion in exon 19  | NRAS   | c.183A>T      | p.Q61H   | TP53   | c.817C>T  | p.R273C   |
| EGFR | Insertion in exon 20 | NRAS   | c.181C>A      | p.Q61K   | TP53   | c.818G>A  | p.R273H   |
| EGFR | c.2236G>A            | NRAS   | c.182A>T      | p.Q61L   | TP53   | c.318C>G  | p.S106R   |
| EGFR | c.2236_2250del15     | NRAS   | c.182A>G      | p.Q61R   | TP53   | c.659A>G  | p.Y220C   |
|      | p.E746_A750delELREA  | PIK3CA | c.1633G>A     | p.E545K  | TP53   | c.707A>G  | p.Y236C   |
| EGFR | c.2156G>C            | PIK3CA | c.3140A>T     | p.H1047L | PIK3CA | c.3140A>G | p.H1047RC |
| EGFR | c.2155G>T            | PIK3CA | c.3140A>G     | p.H1047R | PIK3CA | c.1637A>G | p.Q546R   |
| EGFR | c.2155G>A            | PTEN   | c.754G>T      | p.D252Y  |        |           |           |
| EGFR | c.2573T>G            | PTEN   | c.59G>A       | p.G20E   |        |           |           |
| EGFR | c.2582T>A            | RET    | rearrangement |          |        |           |           |
| EGFR | c.2303G>T            | ROS1   | rearrangement |          |        |           |           |
| JAK2 | c.1849G>T            | SMAD4  | c.1157G>A     | p.G386D  |        |           |           |
| JAK3 | c.2164G>A            |        |               |          |        |           |           |

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





# 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**



# 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

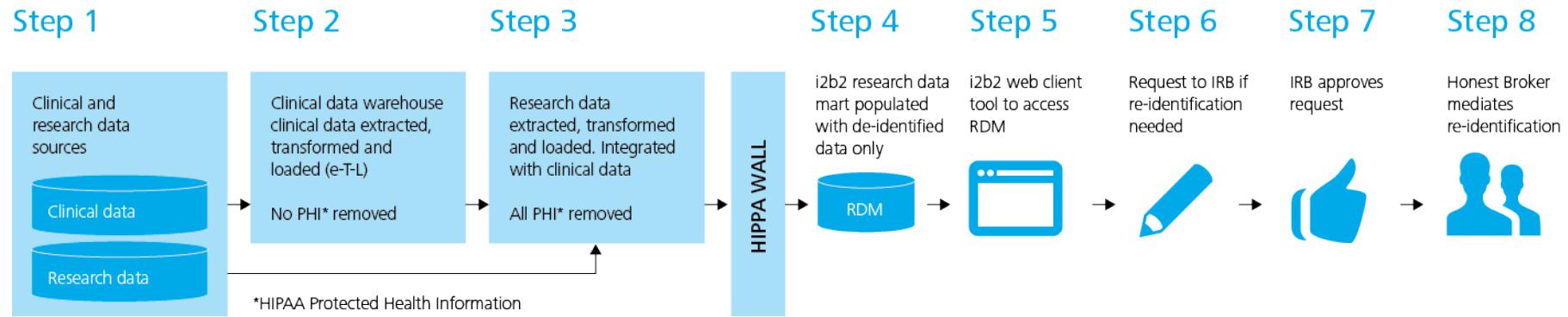
- chemotherapy, radiation, surgery, transplant, palliative

## Disease-specific factors

- ex: (prostate --> Gleason score)



# Research data work flow





# Drag-and-drop i2b2 query tool

i2b2 Query & Analysis Tool   Project: Jefferson production i2b2 RDM   User: Jack London   Find Patients | Analysis Tools | Message Log | Help | Change Password | Logout

**Navigate Terms** **Find Terms**

Demographics  
Diagnoses (Primary, Secondary, Admitting, RadOnc)  
Discharge Disposition  
Hospitalization  
Labs, Selected (LOINC)  
Medications, Chemo Orders (RxNorm-Ingredients)  
Omic Data  
Procedures, Inpatient (ICD-9 and CPT)  
Research Studies  
Specimens (SNOMED)  
Tumor Registry  
Vitals

**Query Tool**

Query Name:

Temporal Constraint:  Treat all groups independently

**Group 1** **Group 2** **Group 3**

Dates Occurs > 0x Exclude  
Treat Independently ▾

drop a term on here

Run Query Clear Print Query 0 Groups New Group

Query Status

The screenshot shows the i2b2 Query & Analysis Tool interface. On the left, there's a sidebar with 'Navigate Terms' and 'Find Terms' buttons, and a tree view of medical data categories like Demographics, Diagnoses, and Procedures. The main area is titled 'Query Tool' with fields for 'Query Name' and 'Temporal Constraint'. It features three 'Group' sections (Group 1, Group 2, Group 3) each with 'Dates', 'Occurs > 0x', and 'Exclude' options, and a 'Treat Independently' dropdown. A yellow callout box with the text 'drop a term on here' points to the first group's input field. At the bottom, there are 'Run Query', 'Clear', and 'Print Query' buttons, along with a '0 Groups' count and 'New Group' button. Below the main area is a 'Query Status' section.



Jefferson™



# Taxonomy for tumor registry data

The image shows two side-by-side windows of the i2b2 Query & Analysis Tool interface, both titled "i2b2 Query & Analysis Tool".

**Left Window (Project: Jefferson F):**

- Left Panel:** A tree view labeled "Navigate Terms" showing a hierarchical structure of tumor categories. A red arrow points to the node "BREAST C50".
  - CS Site Specific Factors
  - Multiple Primary Diagnoses
  - Primary Cancer Diagnosis
    - BLOOD, BONE MARF
    - BONES, JOINTS AND
    - BRAIN AND OTHER P
    - BREAST C50**
  - CONNECTIVE, SUBC
  - DIGESTIVE ORGANS
  - ENDOCRINE GLAND
  - EYE AND ADNEXA C
  - FEMALE GENITAL O
  - LIP, ORAL CAVITY A
  - LYMPH NODES C77
  - MALE GENITAL ORG
  - OTHER AND ILL-DEF
  - PERIPHERAL NERVE
  - RESPIRATORY SYS
  - RETROPERITONEUM
  - SKIN C44
  - URINARY ORGANS C
  - Recurrence (First Only)
  - Treatment
  - Vitals
- Right Panel:** A detailed view of the "BREAST C50" node, also labeled "Navigate Terms".
  - Primary Cancer Diagnosis (ICD-O3)
    - BLOOD, BONE MARROW, HEMATOPOIETIC AND
    - BONES, JOINTS AND ARTICULAR CARTILAGE C
    - BRAIN AND OTHER PARTS OF CENTRAL NERVO
    - BREAST C50**
  - DX Date-Age, Tumor Sequence, Survival, and S
    - BREAST C50 - 14368**
      - Age at diagnosis
      - Date of diagnosis
      - Primary Tumor Sequence
      - Survival (months from date of DX)
      - Survival disease-free (months from date o
      - Year of 1st TJUH contact
      - Axillary tail of breast C506 - 48
      - Breast, NOS C509 - 4124
      - Central portion of breast C501 - 383
      - Lower-inner quadrant of breast C507 - 74
      - Lower-outer quadrant of breast C505 - 96
      - Nipple C500 - 249
      - Overlapping lesion of breast C508 - 2920
      - Upper-inner quadrant of breast C502 - 12
      - Upper-outer quadrant of breast C504 - 43
    - Histology
    - Stage, Grade, Behavior
      - AJCC Best Stage - 14368
      - AJCC Clinical Stage - 14368
      - AJCC Pathological Stage - 14368
      - Behavior (benign, malignant, in situ) - 14368
      - Grade (differentiation) - 14368
    - TNM Clinical
      - Clinical M (metastasis) - 14368
      - Clinical N (nodes) - 14368
      - Clinical T (tumor) - 14368
    - TNM Pathological
      - Pathological M (metastasis) - 14368
      - Pathological N (nodes) - 14368
      - Pathological T (tumor) - 14368
  - Bottom Panel:** A list of other tumor categories:
    - CONNECTIVE, SUBCUTANEOUS AND OTHER SOFT TISSUES
    - DIGESTIVE ORGANS C15-C26
    - ENDOCRINE GLANDS AND RELATED STRUCTURES C73-C75
    - EYE AND ADNEXA C69
    - FEMALE GENITAL ORGANS C51-C58





# Identification of patient cohorts or hypothesis generation

i2b2 Query & Analysis Tool

Project: Jefferson RDM User: Jack London

Find Patients | Analysis Tools | Help | Logout

**Query Tool**

Query Name: triple-neg-froz-spec@15:55:39

Temporal Constraint: Treat all groups independently

**Group 1**

Dates Occurs > 0x Exclude  
Treat Independently

ER Negative, PR Negative, HER2 Negative (Triple Negative) -

**Group 2**

Dates Occurs > 0x Exclude  
Treat Independently

Infiltrating duct and lobular carcinoma - 693  
Infiltrating duct carcinoma, NOS - 8930  
Infiltrating duct mixed with other types of carcinoma - 277  
Infiltrating ductal carcinoma - 15  
Infiltrating lobular mixed with other types of carcinoma - 17

**Group 3**

Dates Occurs > 0x Exclude  
Treat Independently

BREAST (T0400-T0491) [Specimen Type = ("frozen tissue")]

one or more of these AND one or more of these AND one or more of these

Run Query Clear Print Query 3 Groups New Group

**Query Status**

Finished Query: "triple-neg-froz-spec@15:55:39"  
Compute Time: 65 secs

Patient Set for "triple-neg-froz-spec@15:55:39"

Number of patients for "triple-neg-froz-spec@15:55:39"  
patient\_count: 30

The screenshot shows the i2b2 Query & Analysis Tool interface. On the left, there's a navigation tree with categories like BRAIN AND CENTRAL NERVOUS SYSTEM, BREAST, CARDIOVASCULAR SYSTEM, etc. A specific node under BREAST is selected. The main area contains three query groups. Group 1 filters for 'ER Negative, PR Negative, HER2 Negative (Triple Negative)'. Group 2 filters for various types of carcinomas. Group 3 filters for 'BREAST (T0400-T0491)' where 'Specimen Type' is 'frozen tissue'. The bottom status bar shows the query was finished at 15:55:39 with a compute time of 65 seconds, resulting in a patient count of 30.



Jefferson™



# Molecular Diagnostic data

i2b2 Query & Analysis Tool Project: K

Navigate Terms Find Terms

Omic Data

- Molecular Diagnostics Lab Results
  - Genes
    - ABL1 - 11
    - ABL2 - 4
    - ACVR1B - 3
    - AKT1 - 3
    - AKT2 - 2
    - AKT3 - 2
    - ALK - 316
    - APC - 42
    - AR - 11
    - ARAF - 3
    - ARFRP1 - 3
    - ARID1A - 15
    - ARID1B - 30
    - ARID2 - 10
    - ASXL1 - 21
    - ATM - 29
    - ATR - 15
    - ATRX - 17
    - AURKA - 1
    - AURKB - 2
    - AXIN1 - 7
    - AXL - 3
    - BAP1 - 13
    - BARD1 - 10
    - BCL11B - 1
    - BCL2 - 1
    - BCL2L2 - 1
    - BCL6 - 5
    - BCOR - 14
    - BCORL1 - 16
    - BLM - 11



i2b2 Query & Analysis Tool Project: KC

Navigate Terms Find Terms

- BLM - 11
- BRAF - 1525
  - BRAF Indeterminate - 29
  - BRAF mutations - 229
    - p.A762V, 2285C>T - 1
    - p.D594N, c.1780G>A - 1
    - p.E26D, 78g>T - 1
    - p.G466E, c.? - 1
    - p.K601E, c.1801A>G - 5
    - p.L331P, 992T>C - 1
    - p.L597Q, 1790T>A - 1
    - p.V600E, c.1799T>A - 214
    - p.W531C, 1593G>T - 1
    - p.Y566N, 1696T>A - 1
  - BRAF sample site - 1393
    - BRAF Bladder sample - 1
    - BRAF Blood sample - 3
    - BRAF Bone sample - 5
    - BRAF Brain sample - 12
    - BRAF Breast sample - 1
    - BRAF Colon sample - 159
    - BRAF Kidney sample - 1
    - BRAF Liver sample - 15
    - BRAF Lung sample - 11
    - BRAF Lymph Node sample - 24
    - BRAF Ovary sample - 1
    - BRAF Pancreas sample - 1
    - [BRAF Prostate sample - 0]
    - BRAF Skin sample - 21
    - BRAF Soft Tissue sample - 5
    - BRAF Thyroid sample - 1136
    - BRAF Uterus sample - 1
  - BRAF wildtype - 1290
- BRCA - 169



Jefferson™



# Additional data on selected cohort can be retrieved

i2b2 Web Client  
vm319.jefferson.edu/webclient/

i2b2 Query & Analysis Tool Project: KCC Development User: Jack London Find Patients | Analysis Tools | Message Log | Help | Change Password | Logout

**Query Tool**

Query Name: Temporal Constraint: Treat all groups independently

**Group 1** Dates Occurs > 0x Exclude Treat Independently  
EGFR = ("PATHOGENIC")

**Group 2** Dates Occurs > 0x Exclude Treat Independently  
EGFR Exon 19 Deletion

**Group 3** Dates Occurs > 0x Exclude Treat Independently  
Black or African American - 90913

one or more of these AND one or more of these AND one or more of these

| A       | B                    | C                | D               | E    | F                     | G                 | H                     | I                     | J                  | K                     | L                     |
|---------|----------------------|------------------|-----------------|------|-----------------------|-------------------|-----------------------|-----------------------|--------------------|-----------------------|-----------------------|
| 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 |
| 1       |                      |                  |                 |      |                       |                   |                       |                       |                    |                       |                       |
| 2       | 15217866             | Lung, upper lobe | stage IV        | 12   | EGFR                  | Wild-type         | Deletion              | 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             |
| 4       | 15217355             | Lung, lingua     | stage IV        | 4    | EGFR                  | Wild-type         | Deletion              | 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             |
| 6       | 12444923             | Lung, lower lobe | stage IV        | 3    | EGFR                  | Wild-type         | Deletion              | 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             |
| 8       | 16656602             | Lung, lower lobe | stage IIIA      | 14   | EGFR                  | Wild-type         | Deletion              | 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             |
| 10      | 18660745             | Lung, upper lobe | stage IIIA      | 19   | EGFR                  | Wild-type         | Deletion              | Wild-type             | Wild-type          | Wild-type             | Wild-type             |
| 11      | 19666663             | Lung, lower lobe | stage IV        | 9    | EGFR                  | Wild-type         | Deletion              | Wild-type             | Wild-type          | Wild-type             | Wild-type             |

**Jefferson**



Pathology images are available via *i2b2* query tool

The screenshot shows a dual-pane interface. The left pane is a web-based clinical information system (I2b2) displaying patient data and specimen details. The right pane is the Sidney Kimmel Cancer Center website, featuring a 3D reconstruction of a tumor and several histological specimen images.

**I2b2 Web Client**

**Query & Analysis Tool**

**Project: Jefferson production (I2b2 ROM)**

**Navigation Terms**

- Demographics
- Diagnoses Primary, Secondary, Admitting, Rad(On)
- Examinations
- Hospitalization
- Labs, Selected (LOINC)
- Measurements, Chemo Orders (RxNorm-Ingredients)
- Oral Rx
- Procedures, Inpatient (ICD-9 and CPT)
- Research Studies
- Specimens (Specimen MED)

  - Specimen Class
  - Specimen Images
  - Specimen Metadata
  - Specimen Type

- BONES, JOINTS, MUSCLE AND ARTICULAR CARTILAGE (T10-T13)
- ENTITLED: SPINAL AND NERVOUS SYSTEM (T000-T084)

  - Specimen Class
  - Specimen Images
  - Specimen Metadata
  - Specimen Type

- Arachnoid, NOS - 44
- Arteriovenous Malformation - 6473
- Cavernous sinus - 29
- Central nervous system - 15
- Spinal canal/bone - 44

**Workplace**

- SHARED
- admin001
- admin012
- admin008
- admin008
- admin008
- cpg003
- demo
- dec119
- hx0008
- hx001
- hx001
- hx002
- jeff140
- jeff101
- jordon
- lmm106
- lmm107
- hx002
- hx042
- hx0102
- hx0003
- hx0011
- msa011
- msa001

**Previous Queries**

- Specimen Images@16:22:05 [hx002]
- Results of Specimen Images@16:22:05 [hx002]
- Patient Set for 'Specimen Images@16:22:05' - FINISHED
- Study Name 1@16:22:05 [TInNMR\_Prod]
- Study Name 1@16:22:05 [TInNMR\_Prod]

**ExportXLS**

Specify Data Output Options View Results Plugin Help

Click on one of the buttons on the right to download the following table in the appropriate format.

CSV Export HTML/XLS Export

**Sidney Kimmel Cancer Center**  
at Thomas Jefferson University

**Patient Information (for patients 1 - 21)**

| Patient ID |          |
|------------|----------|
| 1          | 1173295  |
| 2          | 4033997  |
| 3          | 11864390 |
| 4          | 11865511 |
| 5          | 12014073 |
| 6          | 12042334 |
| 7          | 12121040 |
| 8          | 12128356 |
| 9          | 12161573 |
| 10         | 12172252 |
| 11         | 12184442 |
| 12         | 12187061 |
| 13         | 12235613 |
| 14         | 12401589 |
| 15         | 12409395 |
| 16         | 12409849 |
| 17         | 12409926 |
| 18         | 12541397 |
| 19         | 12566726 |
| 20         | 12002922 |

**Timeline**

Download Original Aprio Image

Maintained by the Informatics Shared Resources of the Sidney Kimmel Cancer Center at Jefferson  
Copyright © Thomas Jefferson University. All Rights Reserved.  
The Thomas Jefferson University web site, its contents and programs, is provided for informational and educational purposes only and is not intended as medical advice nor is it intended to create any physician-patient relationship. Please remember that this information should not substitute for a visit or a consultation with a health care provider. The views or opinions expressed in the resources provided do not necessarily reflect those of Thomas Jefferson University, Thomas Jefferson University Hospital, or the Jefferson Health System or staff.  
Please read our Privacy Statement

**University Home | Hospital Home | Pulse | Kimmel Cancer Center | Employment | Contact Us**

**THOMAS JEFFERSON UNIVERSITY**

**NCI-CC**  
A Cancer Center Designated by the National Cancer Institute



# Why data analytics?



Old School



Sabermetrics



**Jefferson**™



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

Downloaded from [jamia.bmjjournals.org](http://jamia.bmjjournals.org) on September 16, 2013 - Published by [group.bmjjournals.org](http://group.bmjjournals.org)

## Research and applications

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

Jack W London,<sup>1,2</sup> Luanne Balestrucci,<sup>3</sup> Devjani Chatterjee,<sup>1</sup> Tingting Zhan<sup>4</sup>

<sup>1</sup>Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

<sup>2</sup>Department of Cancer Biology, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

<sup>3</sup>Jefferson Graduate School of Biomedical Sciences, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

<sup>4</sup>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](mailto: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.<sup>1</sup> 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.<sup>2</sup> A number of studies have investigated barriers to clinical trial accrual, and reported various physician-related and patient-related obstacles.<sup>3–9</sup> 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*<sup>10</sup> 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.



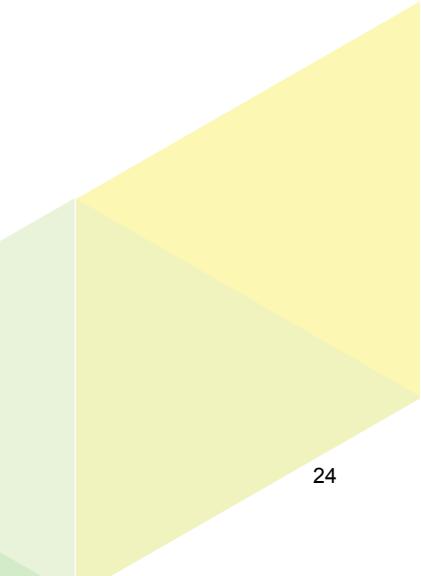
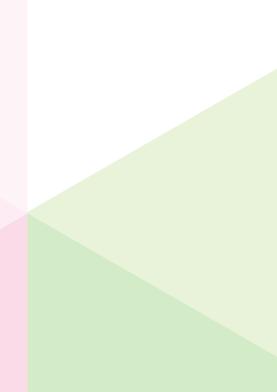
## Jefferson SKCC experience with i2b2

- 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
    - interoperability 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.



# Extensions of i2b2 database use

- tranSMART
- TriNetX



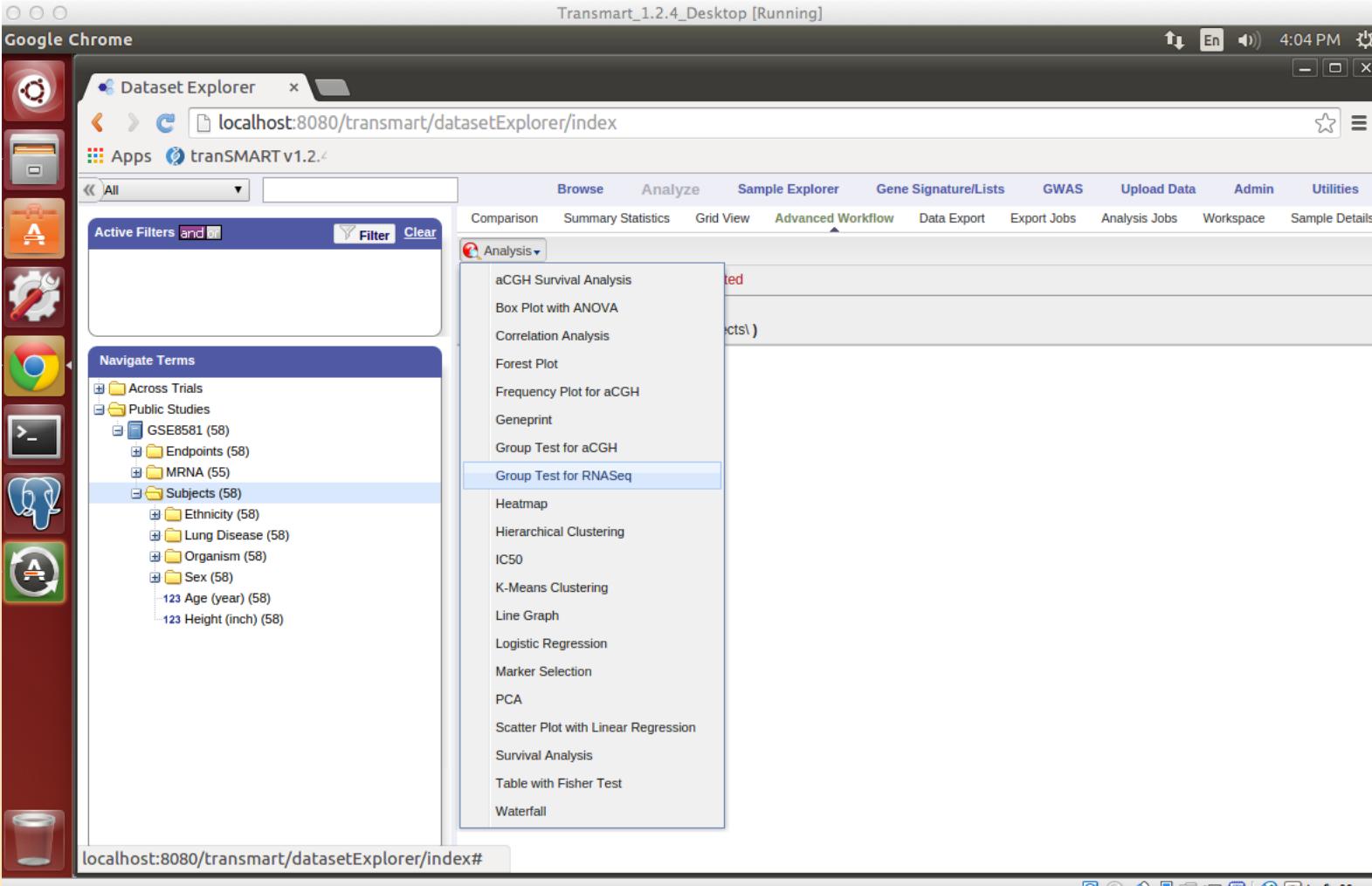


# tranSMART

- 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.



# tranSMART statistical analyses



The screenshot shows the Google Chrome browser window for the tranSMART Dataset Explorer at [localhost:8080/transmart/datasetExplorer/index](http://localhost:8080/transmart/datasetExplorer/index). The interface includes a navigation sidebar with icons for various applications like R, Python, and MySQL, and a main panel with tabs for Browse, Analyze, Sample Explorer, etc. A dropdown menu under 'Analysis' is open, listing various statistical methods:

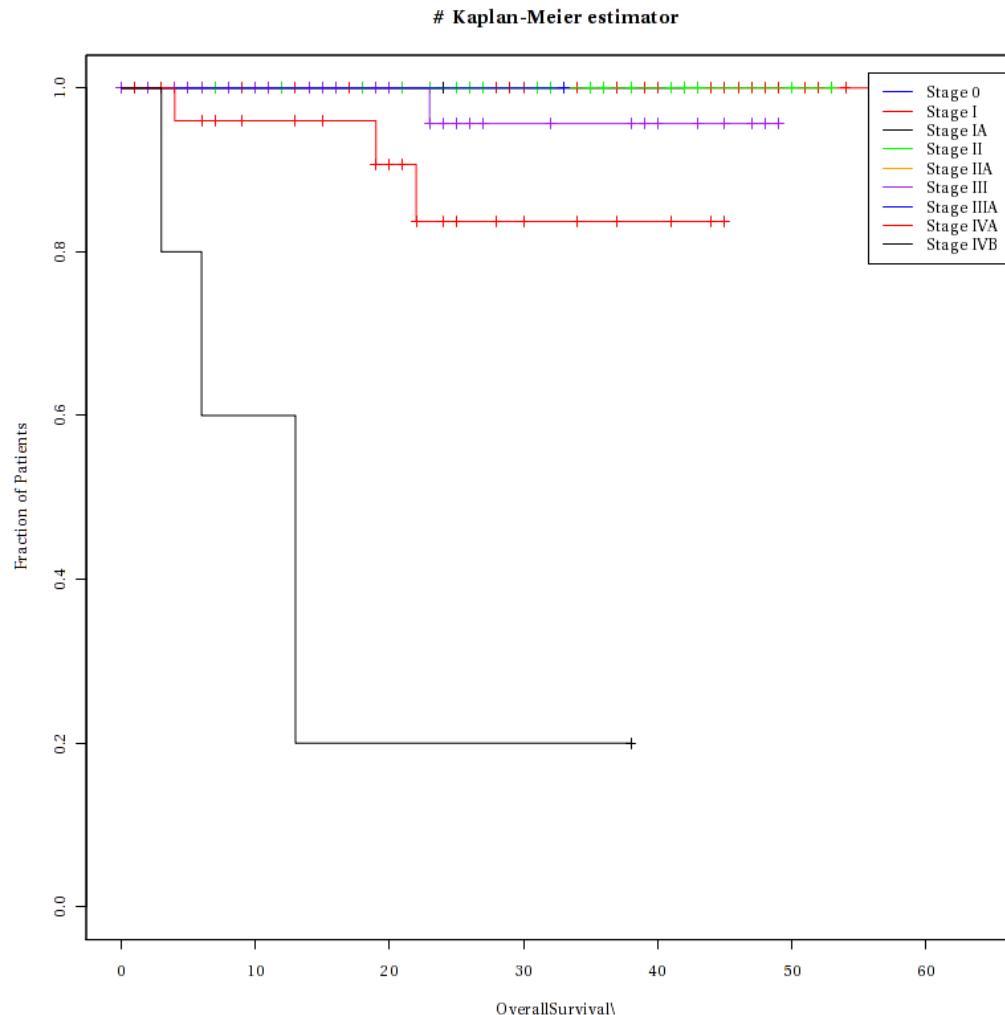
- aCGH Survival Analysis
- Box Plot with ANOVA
- Correlation Analysis
- Forest Plot
- Frequency Plot for aCGH
- Geneprint
- Group Test for aCGH
- Group Test for RNASeq
- Heatmap
- Hierarchical Clustering
- IC50
- K-Means Clustering
- Line Graph
- Logistic Regression
- Marker Selection
- PCA
- Scatter Plot with Linear Regression
- Survival Analysis
- Table with Fisher Test
- Waterfall





# Example of Kaplan-Meier plot from tranSMART

## SURVIVAL CURVE (STAGE) OF TJUH PATIENTS WITH THYROID SPECIMENS





## 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

Screenshot of the TriNetX interface showing cohort demographics for Jefferson SKCC.

**Header:** Gogo – In Air. Online. × M Inbox – london.jw@gmail.com × Mail – Jack London – Outlook × TriNetX × Jack

**Breadcrumbs:** Apps Financial TriNetX Jefferson Travel Gmail Search transSMART JeffMail XFINITY Speed Test OpenSpecimen 1.0 Gmail Other Bookmarks

**User:** Jack London Log Out Network View Admin Help

**Summary:** 525 Patients, 1 Site (Jefferson)

**Criteria Analysis:** Must Have: [ Malignant neoplasm of female breast OR Malignant neoplasm of male breast ] AND Estrogen receptor positive status [er+] Cannot Have: Carcinoma in situ of breast Show

**Cohort:** Mar 12, 2016 11:37 525 / 1

**Run Again**

**Arrival Rate:** Mar 08, 2016 12:00 525 / 1

**Demographics** (selected) | **Sites** | **Diagnoses** | **Procedures** | **Medications** | **Labs**

**525 Unique Patients**

**Age [any]**

Legend: ● months (0-36) ● years (0-89) | ● grouped ● stacked | ♂ male ♀ female

Bar chart showing Age distribution: Minimum Age 26, Maximum Age 90, Mean Age 65, Standard Deviation 14.

Patients 90 and over: 21

**Gender:**

| Gender         | Percentage |
|----------------|------------|
| Female         | 99%        |
| Male           | 1%         |
| Unknown Gender | 0%         |

**Race:**

| Race                   | Percentage |
|------------------------|------------|
| White                  | 65%        |
| Black or African Am... | 24%        |
| Unknown Race           | 5%         |
| Asian                  | 5%         |
| American Indian or ... | 1%         |
| Native Hawaiian or ... | 0%         |

**Ethnicity:**

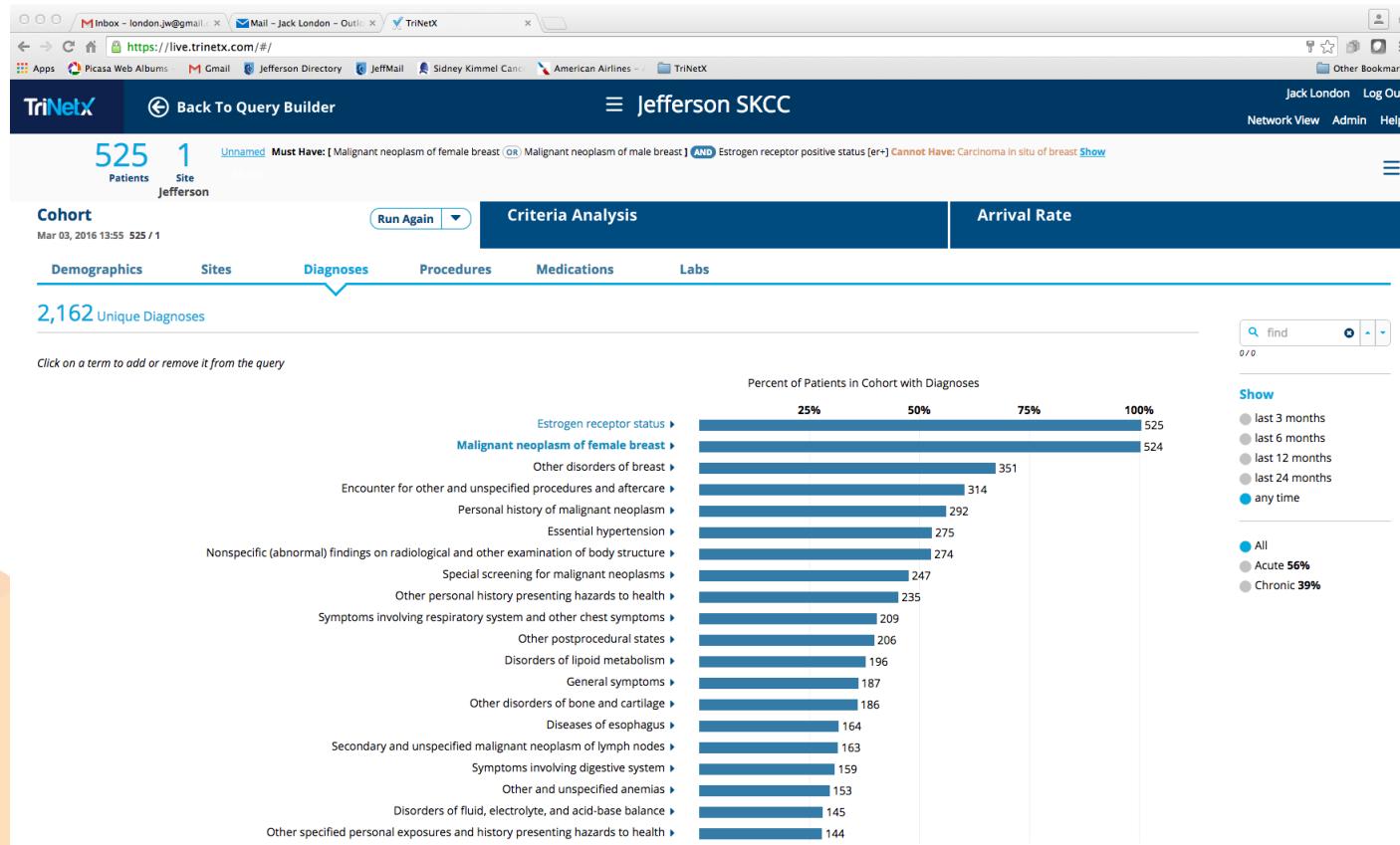
| Ethnicity                    | Percentage |
|------------------------------|------------|
| Unknown Ethnicity            | 97%        |
| Hispanic or Latino           | 3%         |
| Not Hispanic or Lat...<br>er | 0%         |



Jefferson™



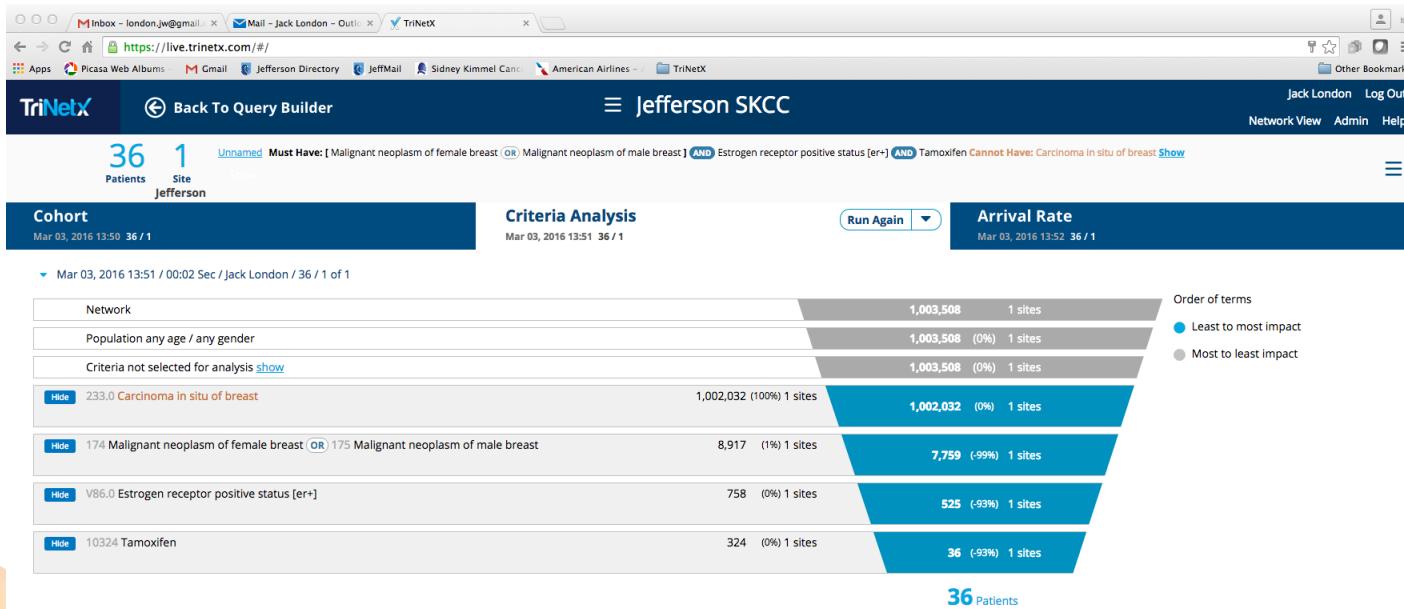
# TriNetX display of cohort co-morbidities



Jefferson™



# TriNetX display of cohort criteria analysis





# Questions?



**Jefferson**™

