



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.bmj.com on March 11, 2010 - Published by group.bmj.com

Model formulation



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}

¹Laboratory of Computer Science, Massachusetts General Hospital, Boston, Massachusetts, USA
²Harvard Medical School, Boston, Massachusetts, USA
³Information Systems, Partners HealthCare System, Inc., Wellesley, Massachusetts, USA
⁴Children's Hospital, Boston, Massachusetts, USA
⁵Brigham and Women's Hospital, Boston, Massachusetts, USA
⁶Department 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'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
	KRAS c.34G>C	p.G12R	TP53 c.400T>C	p.F134L
BRAF c.1782T>G p.D594E	KRAS c.34G>A	p.G12S	TP53 c.734G>A	p.G245D
BRAF c.1801A>G p.K601E	KRAS c.35G>T	p.G12V	TP53 c.388C>G	p.L130V
BRAF c.1799T>A p.V600E	KRAS c.38G>A	p.G13D	TP53 c.524G>A	p.R175H
			TP53 c.817C>T	p.R273C
EGFR Deletion in exon 19	NRAS c.183A>T	p.Q61H	TP53 c.818G>A	p.R273H
EGFR Insertion in exon 20	NRAS c.181C>A	p.Q61K	TP53 c.318C>G	p.S106R
EGFR c.2236G>A p.E746K	NRAS c.182A>T	p.Q61L	TP53 c.659A>G	p.Y220C
EGFR c.2236_2250del15	NRAS c.182A>G	p.Q61R	TP53 c.707A>G	p.Y236C
p.E746_A750delELREA				
EGFR c.2156G>C p.G719A	PIK3CA c.1633G>A	p.E545K	PIK3CA c.3140A>G	p.H1047RC
EGFR c.2155G>T p.G719C	PIK3CA c.3140A>T	p.H1047L	PIK3CA c.1637A>G	p.Q546R
EGFR c.2155G>A p.G719S	PIK3CA c.3140A>G	p.H1047R		
EGFR c.2573T>G p.L858R				
EGFR c.2582T>A p.L861Q	PTEN c.754G>T	p.D252Y		
EGFR c.2303G>T p.S768I	PTEN c.59G>A	p.G20E		
JAK2 c.1849G>T p.V617F	RET rearrangement			
JAK3 c.2164G>A p.V722I	ROS1 rearrangement			
	SMAD4 c.1157G>A	p.G386D		

**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

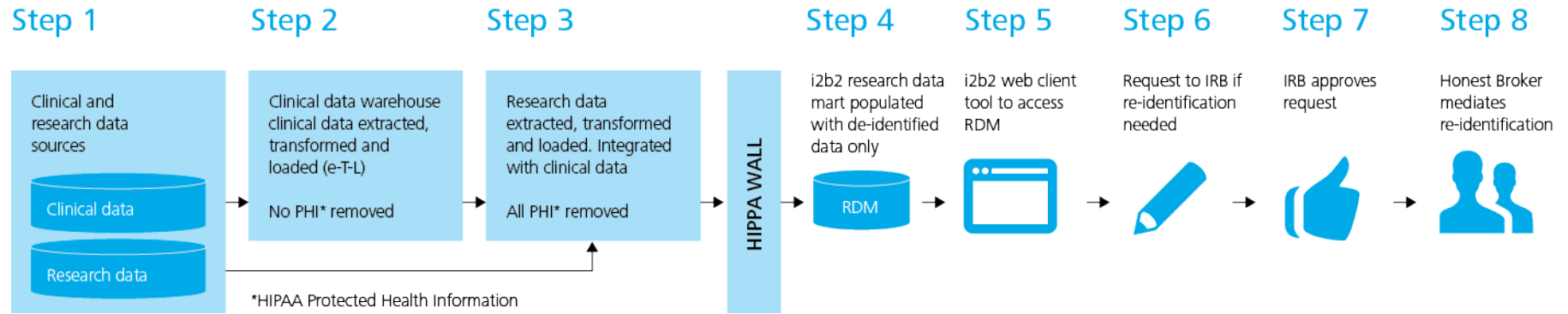
chemotherapy, radiation, surgery, transplant, palliative

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:

Group 1			Group 2			Group 3		
Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude
Treat Independently			Treat Independently			Treat Independently		
drop a term on here								

Run Query Clear Print Query 0 Groups New Group

Query Status



Taxonomy for tumor registry data

The image displays three overlapping screenshots of the 'i2b2 Query & Analysis Tool' interface, illustrating the taxonomy for tumor registry data. The tool features two main tabs: 'Navigate Terms' and 'Find Terms'. The 'Navigate Terms' tab is active in all three screenshots, showing a hierarchical tree structure of data categories.

The central screenshot shows the following structure:

- Primary Cancer Diagnosis (ICD-O3)
 - BLOOD, BONE MARROW, HEMATOPOIETIC AND...
 - BONES, JOINTS AND ARTICULAR CARTILAGE C...
 - BRAIN AND OTHER PARTS OF CENTRAL NERV...
 - 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
 - CONNECTIVE, SUBCUTANEOUS AND OTHER SC...
 - DIGESTIVE ORGANS C15-C26
 - ENDOCRINE GLANDS AND RELATED STRUCTU...

The rightmost screenshot shows a more detailed view of the 'BREAST C50' category, including sub-categories like 'Histology' and 'Stage, Grade, Behavior', with further sub-items such as 'AJCC Best Stage - 14368', 'AJCC Clinical Stage - 14368', 'AJCC Pathological Stage - 14368', 'Behavior (benign, malignant, in situ) - 14368', 'Grade (differentiation) - 14368', 'TNM Clinical', 'TNM Pathological', 'Clinical M (metastasis) - 14368', 'Clinical N (nodes) - 14368', 'Clinical T (tumor) - 14368', 'Pathological M (metastasis) - 14368', 'Pathological N (nodes) - 14368', and 'Pathological T (tumor) - 14368'.

Red arrows indicate the flow of navigation: one arrow points from the 'BREAST C50' category in the middle window to the 'BREAST C50' category in the right window, and another arrow points from the 'Stage, Grade, Behavior' category in the right window to the 'Stage, Grade, Behavior' category in the middle window.



Identification of patient cohorts or hypothesis generation

i2b2 Web Client
Project: Jefferson RDM User: Jack London Find Patients | Analysis Tools | Help | Logout

i2b2 Query & Analysis Tool

Navigate Terms Find Terms

- BRAIN AND CENTRAL NERVOUS SYSTEM (T1000-T1499)
- BREAST (T0400-T0491)
 - Specimen Case Identifier
 - Specimen Class
 - Specimen Identifier
 - Specimen Pathology
 - Specimen Type
 - Areola - 237
 - Both breasts - 86
 - Breast, NOS - 13703
 - [Central portion of breast, NOS - 0]
 - Female breast (T0410-T0403) - 22472
 - Lactiferous duct - 41
 - [Lower inner quadrant of breast, NOS - 0]
 - [Lower outer quadrant of breast, NOS - 0]
 - Male breast (T0404-T0406) - 473
 - Mammary duct - 1355
 - Mammary lobule - 153
 - Nipple - 1021
- CARDIOVASCULAR SYSTEM (T3100-T4953)
- CONNECTIVE, SUBCUTANEOUS AND OTHER SOFT
- DIGESTIVE SYSTEM (T5000-T6X94)
- ENDOCRINE SYSTEM (T9000-T9900)
- EYE AND EAR (TX000-TXY60)
- FEMALE GENITAL ORGANS (T8010-T8Y4)
- HEAD AND NECK REGIONS (TY000-TY062)
- HEMATOPOIETIC AND RETICULOENDOTHELIAL SY
- LYMPH NODES (T080-T095)
- MALE GENITAL ORGANS (T7600-T7940)
- OVERLAPPING LESIONS (D0000-G80A4)
- RESPIRATORY SYSTEM (T2000-T2Y72)
- SKIN (T01000-T02910)
- TRUNK (TY110-TY796)
- URINARY SYSTEM (T70-T75)
- Demographics
- Diagnoses (Primary, Secondary, Admitting, RadOnc)
- Discharge Disposition
- Hospitalization
- Labs, Selected (LOINC)
- Medications, Chemo Orders (RxNorm-Ingredients)
- Procedures, Inpatient (ICD-9 and CPT)
- Research Studies
- Tumor Registry
 - CS Site Specific Factors
 - BREAST
 - CS Site-Specific Factor 01: Estrogen Receptor (ER)
 - CS Site-Specific Factor 02: Progesterone Receptor
 - CS Site-Specific Factor 03: Number of Positive Ipsil
 - CS Site-Specific Factor 04: Immunohistochemistry (
 - CS Site-Specific Factor 05: Molecular (MOL) Studie
 - CS Site-Specific Factor 06: Size of Tumor-Invasive
 - CS Site-Specific Factor 07: Nottingham or Bloom-R
 - CS Site-Specific Factor 08: HER2: Immunohistoche
 - CS Site-Specific Factor 09: HER2: Immunohistoche
 - CS Site-Specific Factor 10: HER2: Fluorescence In
 - CS Site-Specific Factor 11: HER2: Fluorescence In
 - CS Site-Specific Factor 12: HER2: Chromoemic In

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

Temporal Constraint: Treat all groups independently

Group 1	Group 2	Group 3
ER Negative, PR Negative, HER2 Negative (Triple Negative)	Infiltrating duct and lobular carcinoma - 693 Infiltrating duct carcinoma, NOS - 8930 Infiltrating duct mixed with other types of carcinoma - 277 Infiltrating ductular carcinoma - 15 Infiltrating lobular mixed with other types of carcinoma - 17	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 [87.2 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



Molecular Diagnostic data

i2b2 Query & Analysis Tool Project: KC

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



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

Navigate Terms Find Terms

- Race
 - American Indian or Alaska Native - 395
 - Asian - 13108
 - Black or African American - 90913
 - Native Hawaiian or Other Pacific Islander - 47
 - Unknown Race - 65860
 - White - 259586
- Vital Status
- Diagnoses (Primary, Secondary, Admitting, RadOnc)
- Discharge Disposition
- Hospitalization
- Labs, Selected (LOINC)
- Medications, Chemo Orders (RxNorm-Ingredients)
- Omic Data
 - Molecular Diagnostics Lab Results - 549
 - NGS Lab Results - 55
 - Genes
 - ABL1 - 2
 - APC - 13
 - ATM - 4
 - BRAF - 10
 - CSF1R - 2
 - EGFR - 331
 - EGFR Exon 18
 - EGFR Exon 19 deletions
 - EGFR Exon 19 Deletion
 - EGFR Exon 19 Low Amp
 - EGFR Exon 19 No Amp
 - EGFR Exon 19 Wild Type
 - EGFR Exon 20
 - EGFR Exon 20 Codon 768 SER
 - EGFR Exon 20 Codon 790 THR
 - EGFR Exon 20 insertions
 - EGFR Exon 21
 - ERBB4 - 1

Query Tool

Query Name: _____

Temporal Constraint: Treat all groups independently

Group 1			Group 2			Group 3		
Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude
Treat Independently			Treat Independently			Treat Independently		
EGFR = ("PATHOGENIC")			EGFR Exon 19 Deletion			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	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

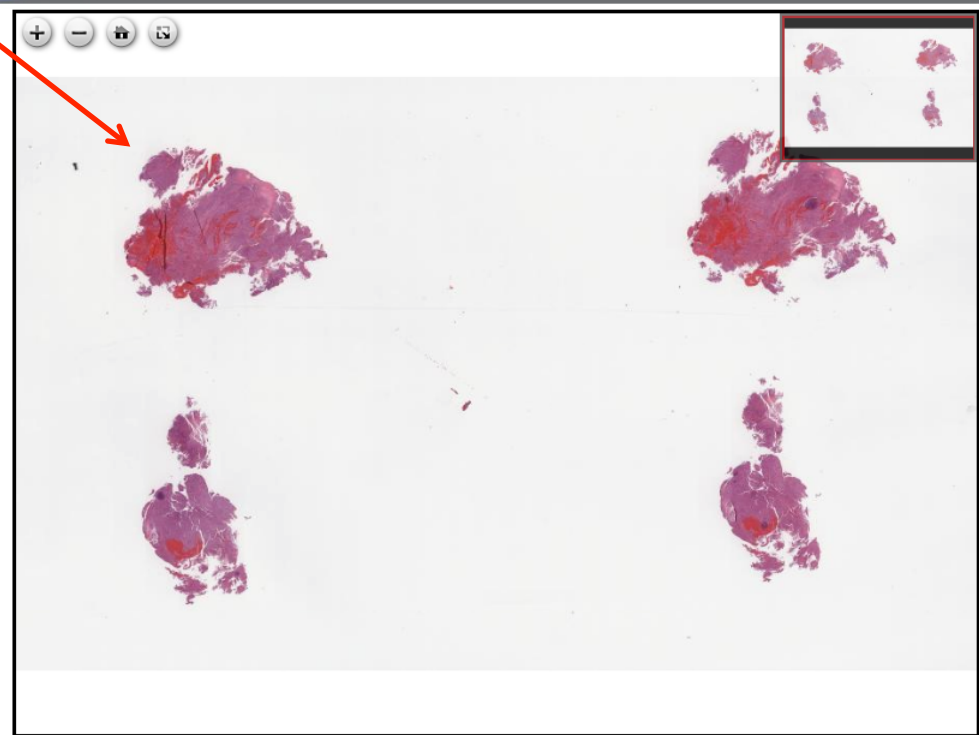


Pathology images are available via *i2b2* query tool

The screenshot shows the i2b2 Web Client interface. On the left, there is a 'Find Terms' panel with a tree view of medical terms including 'Diagnoses', 'Specimens', and 'Bones, Joints, Muscle and Articular Cartilage'. Below this is a 'Workbench' panel with various study names. The main area displays a 'Patient Information' table for patients 1 through 21. The table has columns for 'Patient ID', 'Patient Name', and 'Specimen Images'. An 'Export' button is visible above the table. At the bottom, there is a 'Plugins' section with a 'Detailed List View' and an 'Export' button.

Patient ID	Patient Name	Specimen Images
1	1173995	SP-1782305 = https://blacks.kcc.jhu.edu/...
2	4003991	SP-1781234 = https://blacks.kcc.jhu.edu/...
3	12964390	SP-1782200 = https://blacks.kcc.jhu.edu/...
4	11865511	SP-1782200 = https://blacks.kcc.jhu.edu/...
5	12014073	SP-1782200 = https://blacks.kcc.jhu.edu/...
6	12042334	SP-1782200 = https://blacks.kcc.jhu.edu/...
7	12121040	SP-1782200 = https://blacks.kcc.jhu.edu/...
8	12128356	SP-1782200 = https://blacks.kcc.jhu.edu/...
9	12141573	SP-1782200 = https://blacks.kcc.jhu.edu/...
10	12172202	SP-1782200 = https://blacks.kcc.jhu.edu/...
11	12184442	SP-1782200 = https://blacks.kcc.jhu.edu/...
12	12187041	SP-1782200 = https://blacks.kcc.jhu.edu/...
13	12253423	SP-1782200 = https://blacks.kcc.jhu.edu/...
14	12481589	SP-1782200 = https://blacks.kcc.jhu.edu/...
15	12483595	SP-1786000 = https://blacks.kcc.jhu.edu/...
16	12489849	SP-1782200 = https://blacks.kcc.jhu.edu/...
17	12490026	SP-1782200 = https://blacks.kcc.jhu.edu/...
18	12541287	SP-1782200 = https://blacks.kcc.jhu.edu/...
19	12564726	SP-1782200 = https://blacks.kcc.jhu.edu/...
20	12600992	SP-1782200 = https://blacks.kcc.jhu.edu/...

Sidney Kimmel Cancer Center at Thomas Jefferson University. Navigation links: 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

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](#)



Why data analytics?



Old School

Sabermetrics



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

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 Zhan⁴

¹Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

²Department of Cancer Biology, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

³Jefferson Graduate School of Biomedical Sciences, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

⁴Department 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
 - 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

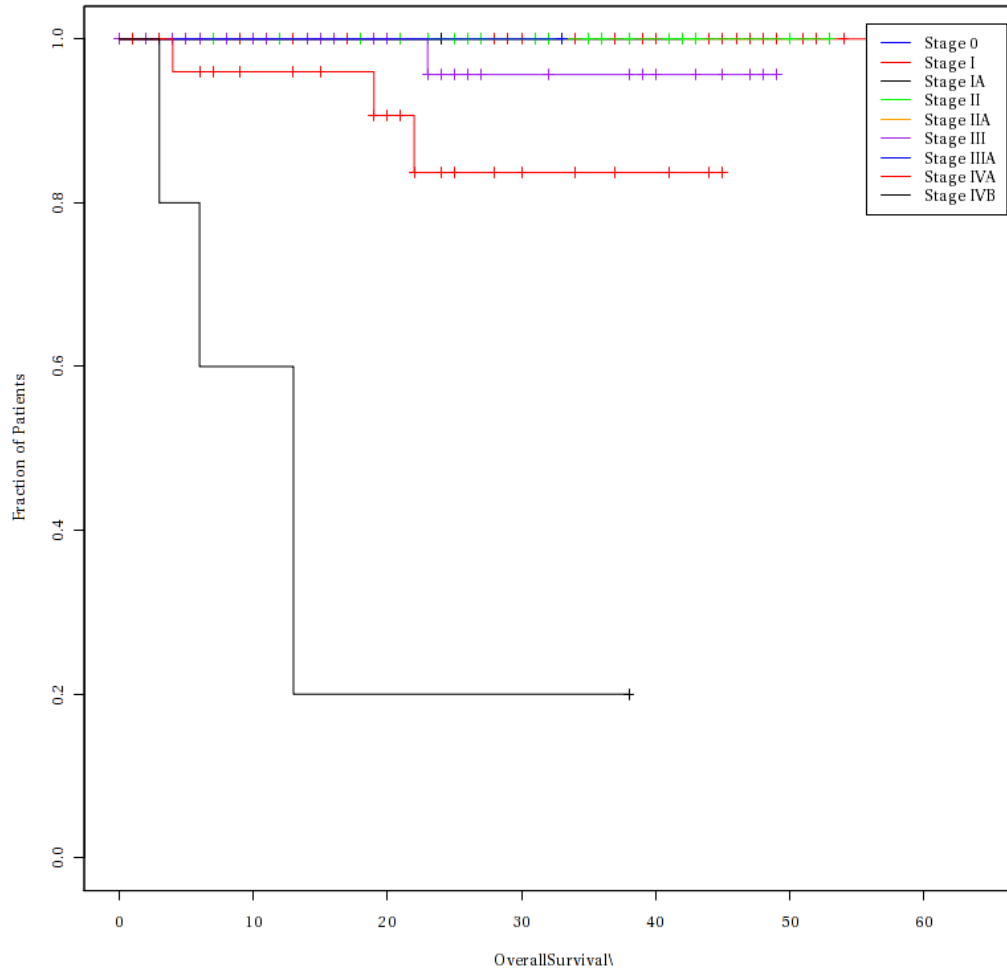
Google Chrome
Transmart_1.2.4_Desktop [Running]
localhost:8080/transmart/datasetExplorer/index
Apps tranSMART v1.2.4
All
Active Filters and [Filter] [Clear]
Navigate Terms
Across Trials
Public Studies
GSE8581 (58)
Endpoints (58)
MRNA (55)
Subjects (58)
Ethnicity (58)
Lung Disease (58)
Organism (58)
Sex (58)
Age (year) (58)
Height (inch) (58)
Analysis
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

Kaplan-Meier estimator





Jefferson – TriNetX project

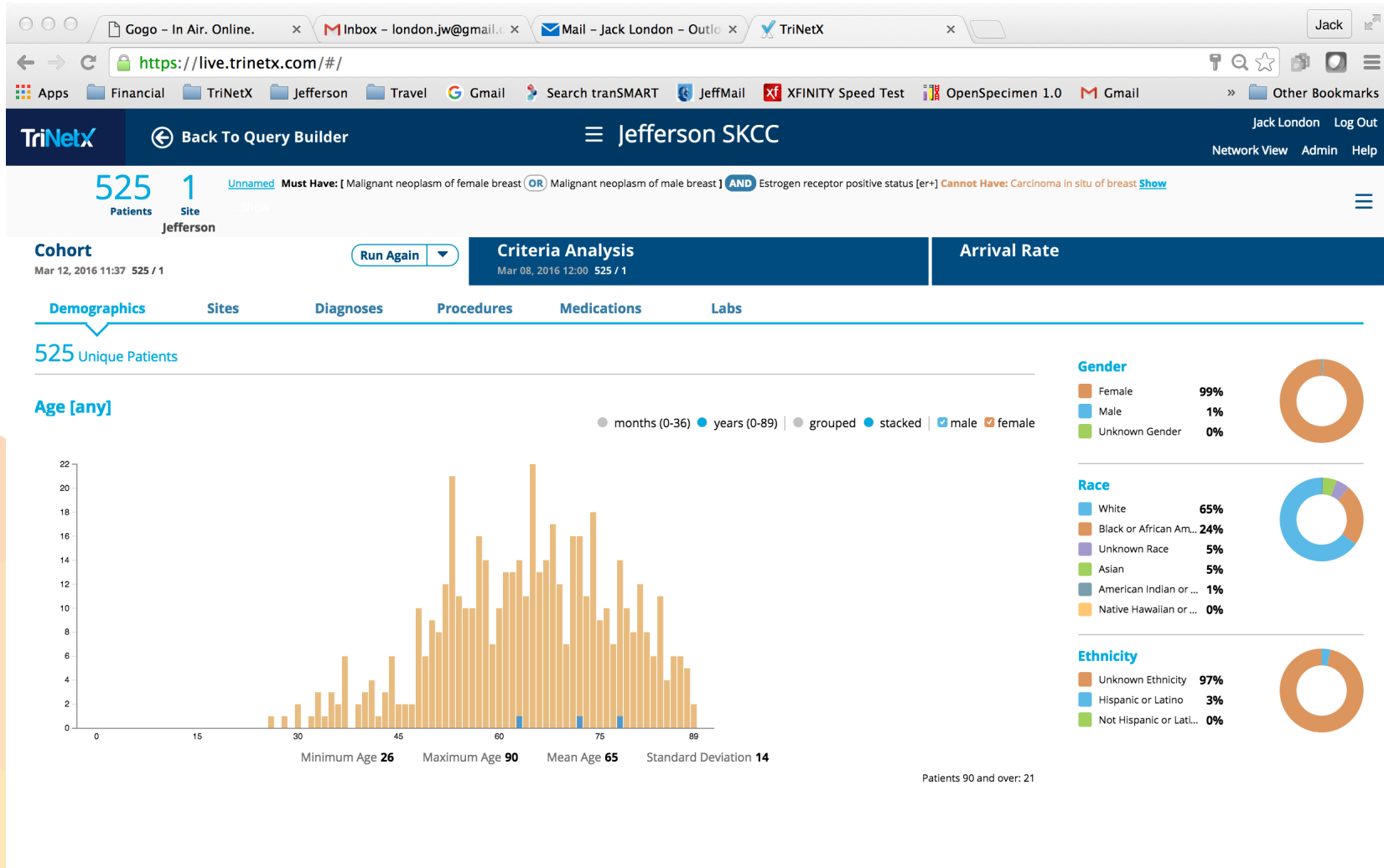
TriNetX facilitates clinical trial collaboration between pharmaceutical companies and academic medical center data providers by providing access to aggregate data from academic members of the TriNetX network.

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.

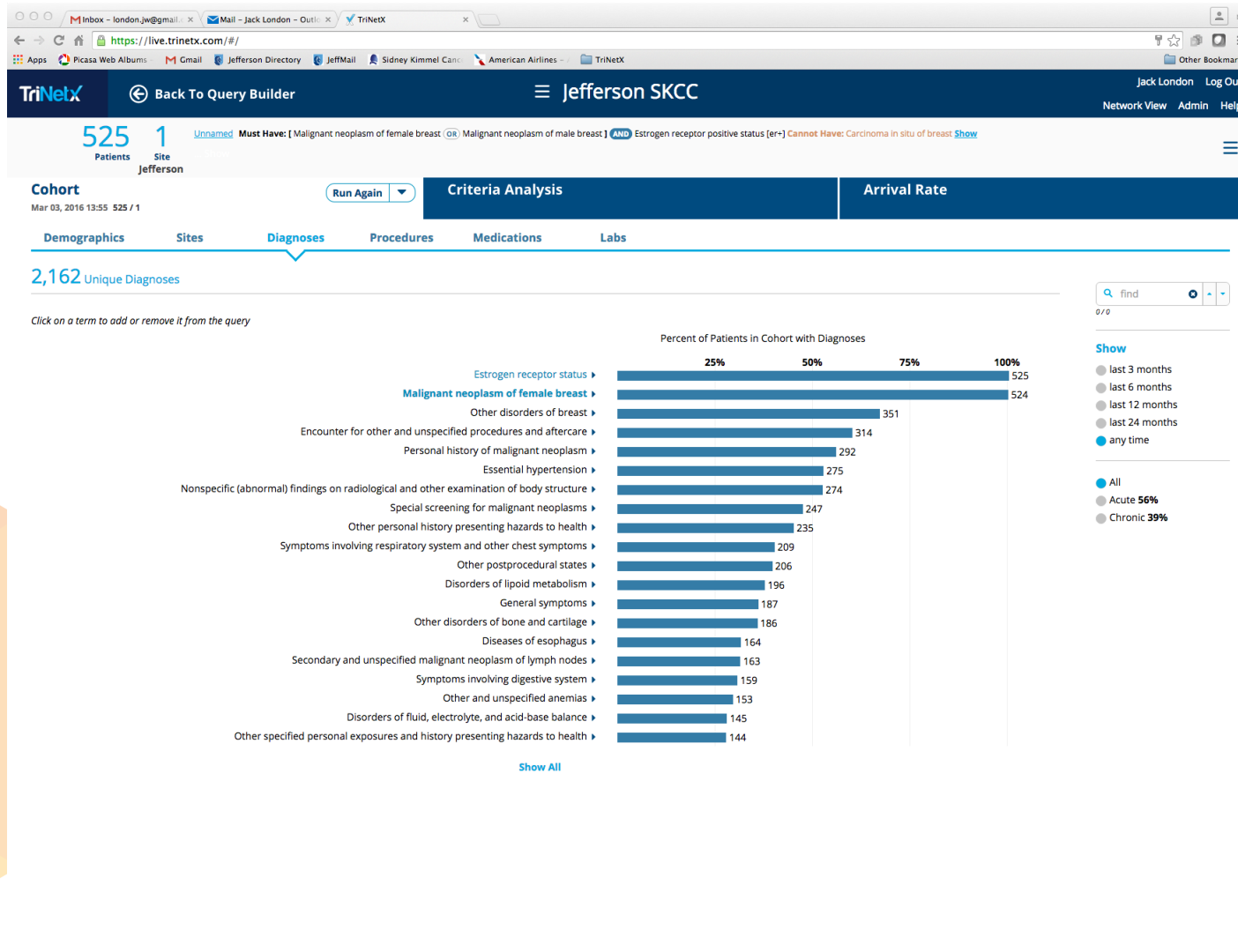


TriNetX display of cohort demographics



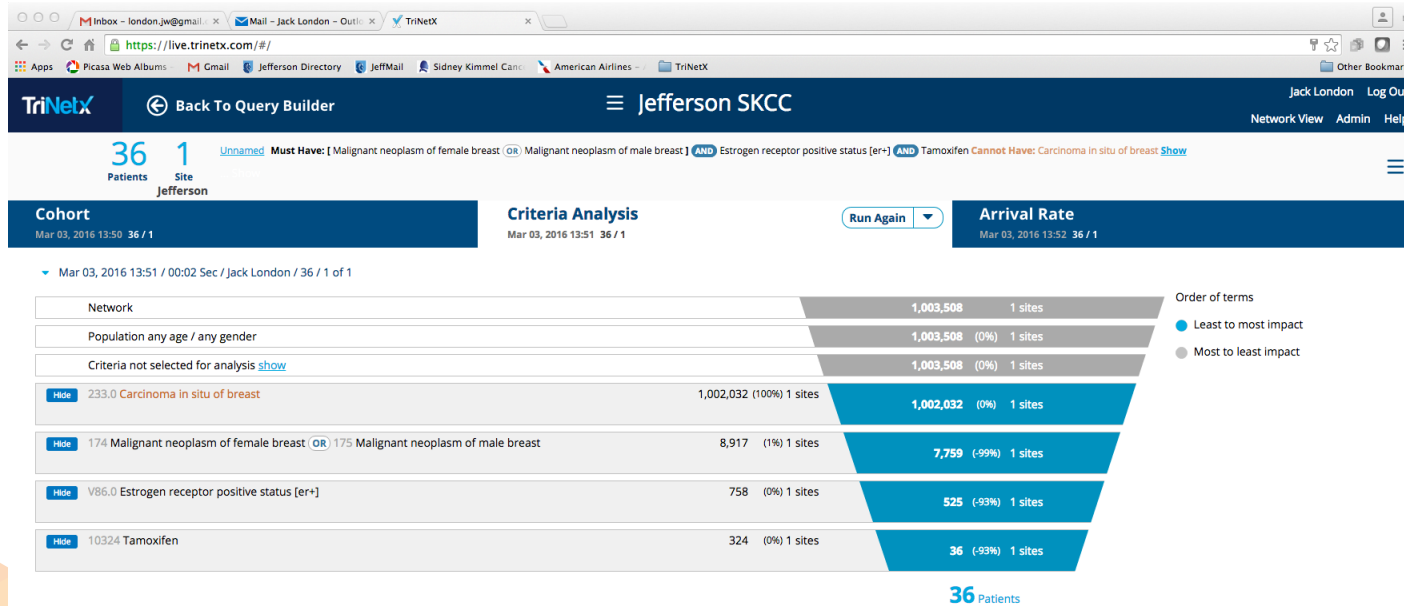


TriNetX display of cohort co-morbidities





TriNetX display of cohort criteria analysis





Questions?

