Instrumenting the Health Care Enterprise for Discovery in the Course of Clinical Care

Shawn Murphy MD, Ph.D.
Chief Research Information Officer
Harvard Medical School / Mass General Brigham
Personalized Medicine and Genomic technology are critical to managing populations

- Managing a population involves improving health outcomes of the group as a whole by identifying, monitoring and addressing health needs of individuals through:
  - Subpopulation stratification
  - Targeted, evidence-based treatment protocols
  - Predictive analytics

Source: Personalized Medicine Coalition and innovation.org ; Oliver Wyman
Example: PPARγ Pro12Ala and Diabetes

Overall $P$ value = $2 \times 10^{-7}$

Odds ratio = 0.79 (0.72-0.86)

Ala is protective

Courtesy J. Hirschhorn
High Throughput Methods for supporting Translational Research

- Set of patients is selected from medical record data in a high throughput fashion

- Investigators explore phenotypes of these patients using Machine Learning tools and a translational team developed to work specifically with medical record data

- Distributed networks cross institutional boundaries for phenotype selection, public health, and hypothesis testing

- Digital medicine is delivered into clinical care through Digital Twin
Data problems that make working with Electronic Healthcare Data to conduct research difficult

1) There are significant risks of a data breach which will result in very large fines and loss of confidence in the hospitals where the breach occurred.

2) The data are not collected for research purposes, and therefore the data can be poorly structured with significant omissions, biases, and inaccuracies.
Research Patient Data Registry (RPDR) at Mass General Brigham to find patient cohorts and distribute data

1) Queries for aggregate patient numbers
   - Warehouse of in & outpatient clinical data
   - 6.7 million Mass General Brigham patients
   - 2.6 billion diagnoses, medications, genomics, procedures, laboratories, & physical findings coupled to demographic & visit data
   - Authorized use by faculty status
   - Clinicians can construct complex queries
   - Queries cannot identify individuals, internally can produce identifiers for (2)

2) Returns detailed patient data
   - Start with list of specific patients, usually from (1)
   - Authorized use by IRB Protocol
   - Returns contact and PCP information, demographics, providers, visits, diagnoses, medications, procedures, laboratories, microbiology, reports (discharge, LMR, operative, radiology, pathology, cardiology, pulmonary, endoscopy), and images into a Microsoft Access database and text files.
FINDING PATIENTS

Query items

Person who is using tool

Query construction

Results - broken down by number distinct of patients
Theory of Kimball translated to Healthcare Data

Star schema

<table>
<thead>
<tr>
<th>Concept DIMENSION</th>
<th>Patient DIMENSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>concept_key</td>
<td>patient_key</td>
</tr>
<tr>
<td>concept_text</td>
<td>patient_id</td>
</tr>
<tr>
<td>search_hierarchy</td>
<td>(encrypted)</td>
</tr>
<tr>
<td></td>
<td>sex</td>
</tr>
<tr>
<td></td>
<td>age</td>
</tr>
<tr>
<td></td>
<td>birth_date</td>
</tr>
<tr>
<td></td>
<td>race</td>
</tr>
<tr>
<td></td>
<td>deceased</td>
</tr>
<tr>
<td></td>
<td>ZIP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient_Concept FACTS</th>
<th>Practitioner DIMENSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>patient_key</td>
<td>practitioner_key</td>
</tr>
<tr>
<td>concept_key</td>
<td>name</td>
</tr>
<tr>
<td>start_date</td>
<td>service</td>
</tr>
<tr>
<td>end_date</td>
<td></td>
</tr>
<tr>
<td>practitioner_key</td>
<td></td>
</tr>
<tr>
<td>value_type</td>
<td></td>
</tr>
<tr>
<td>numeric_value</td>
<td></td>
</tr>
<tr>
<td>textual_value</td>
<td></td>
</tr>
<tr>
<td>abnormal_flag</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Encounter DIMENSION</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>encounter_key</td>
<td></td>
</tr>
<tr>
<td>encounter_date</td>
<td></td>
</tr>
<tr>
<td>hospital_of_service</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Practitioner DIMENSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>practitioner_key</td>
</tr>
<tr>
<td>name</td>
</tr>
<tr>
<td>service</td>
</tr>
</tbody>
</table>

Binary Tree

start search

.22
250
6.7
.04
2600 million
RPDR DETAILED DATA REQUEST WIZARD
Using IRB#mgh-demo-1 (found in the RPDR Identified database) to obtain data from the RPDR
You are logged in as Murphy, Shawn N. in workgroup Shawn Murphy, MD

Select protocol number(s)

Partners IRB (required): mgh-demo-1
Title: RPDR protocol - Demonstration IRB number for Dr. Murphy
Status: Active

Newton Wellesley Hospital IRB: NIWH Demo 1
Title: test
Status: Active

Spaulding Rehabilitation Hospital IRB:

Options for returned set of patients:
- Create a static set of patients from this query that can be used in other RPDR queries
- Rerun the base query shown above to obtain a fresh set of patients
RPDR DETAILED DATA REQUEST WIZARD

Using IRB#mgh-demo-1 (found in the RPDR Identified database) to obtain data from the RPDR
You are logged in as Murphy, Shawn N. in workgroup Shawn Murphy, MD

Please select if you would like a HIPAA-defined (deidentified) limited data set or an identified data set

What's a limited data set?

☐ Limited Data Set
- The files that result from this request will be available in a protected file share with no special encryption.

☐ Identified Data Set
- The text files that result from this request will be encrypted and the Microsoft Access file will be password protected. In order to access the data, a password will be provided.
Select the types of data that should be returned from the RPDR

Only data allowed by your protocol should be chosen

(Identified data sets will always return a set of identified patient medical numbers)
Detailed data is gathered for request and distributed.

Data is gathered from RPDR and other MGB sources.

Output files placed in special directory.

Files include Small Database.
One year's usage of RPDR

- 4526 registered users, 1113 new in just 2019

- 834 teams/year gathering data for research studies

- 4472 detailed patient data sets returned to these teams in 2019, containing data of 24.7 million patient records.

- From a survey of 153 teams
  - Importance of the data received from the RPDR was evaluated in relation to the study it was supporting.
  - Calculated over 4 years (FY15-FY19) the total agreement amounts were $2.27 Billion, making per year consumption critically dependent on RPDR $244 Million.

<table>
<thead>
<tr>
<th>Usefulness of Detailed Data</th>
<th>Total Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical</td>
<td>43%</td>
</tr>
<tr>
<td>Useful</td>
<td>42%</td>
</tr>
<tr>
<td>Not Useful</td>
<td>15%</td>
</tr>
</tbody>
</table>
Rapid investigation of QTc prolongation

- **FDA warning 2011 for Celexa**
  Safety Announcement: [8-24-2011] "should no longer be used at doses greater than 40 mg per day because it can cause abnormal changes in the electrical activity of the heart."

- **But, did NOT include Lexapro (which is active ingredient of Celexa [s-enantiomer])**

- **Shown to be true with RPDR-derived data set with >38,000 EKGs obtained within 14 – 90 day window after medication initiated**

<table>
<thead>
<tr>
<th>Anti-depressant</th>
<th>Adjusted model†</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SSRI</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citalopram (Celexa)</td>
<td>2.85</td>
<td>0.004</td>
</tr>
<tr>
<td>Escitalopram (Lexapro)</td>
<td>3.80</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Fluoxetine (Prozac)</td>
<td>1.44</td>
<td>0.150</td>
</tr>
<tr>
<td>Paroxetine (Paxil)</td>
<td>0.07</td>
<td>0.943</td>
</tr>
<tr>
<td>Sertraline (Zoloft)</td>
<td>0.87</td>
<td>0.383</td>
</tr>
<tr>
<td><strong>Other anti-depressants</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amitriptyline</td>
<td>4.10</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Bupropion</td>
<td>-2.15</td>
<td>0.032</td>
</tr>
<tr>
<td>Duloxetine</td>
<td>0.60</td>
<td>0.547</td>
</tr>
<tr>
<td>Mirtazapine</td>
<td>-1.46</td>
<td>0.145</td>
</tr>
<tr>
<td>Nortriptyline</td>
<td>1.23</td>
<td>0.219</td>
</tr>
<tr>
<td>Venlafaxine</td>
<td>1.15</td>
<td>0.251</td>
</tr>
<tr>
<td><strong>previously known prolonger</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methadone</td>
<td>5.32</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

† Adjusted for age, gender, race, type of insurance, history of major depression, history of myocardial infarction and Charlson comorbidity score

Roy Perlis MD, MSc and team
Relevant Cohorts of Patients are Gathered through RPDR and Detailed Data Obtained

- Medication use by individual patients over time
- Patient EKG QTc values at various time points
**Results: QTc interval and medication use**

**Selective serotonin reuptake inhibitor or methadone**

- Citalopram (mg)*
- Escitalopram (mg)*
- Fluoxetine (mg)
- Paroxetine (mg)
- Sertraline (mg)
- Methadone (mg)*

**Other antidepressant or methadone**

- Amitriptyline (mg)*
- Bupropion (mg)*
- Duloxetine (mg)
- Mirtazapine (mg)
- Nortriptyline (mg)
- Venlafaxine (mg)
- Methadone (mg)*

* Dose a significant predictor of QTc in fully adjusted linear models at α=0.05
† QTc at specified dose is significantly different from that at prior dose in fully adjusted linear models at α=0.05

Mean (SD) corrected QT (QTc) interval recorded on electrocardiogram 14–90 days after prescription of antidepressant or methadone, by drug dose
High Throughput Methods for supporting Translational Research

- Set of patients is selected from medical record data in a high throughput fashion

- Investigators explore phenotypes of these patients using Machine Learning tools and a translational team developed to work specifically with medical record data

- Distributed networks cross institutional boundaries for phenotype selection, public health, and hypothesis testing

- Digital medicine is delivered into clinical care through Digital Twin
RPDR Evolved into international “Informatics for Integrating Biology and the Bedside (i2b2)” sponsored by the National Institutes of Health, what is it?

- Software for explicitly organizing and transforming person-oriented clinical data to a way that is optimized for clinical genomics research
  - Allows integration of clinical data, trials data, and genotypic data
- A portable and extensible application framework
  - Software is built in a modular pattern that allows additions without disturbing core parts
  - Available as open source at https://www.i2b2.org
I2b2 Community Software distributed as open source
I2b2 Software adapts through new plugins

- **i2b2 Sponsored Project - i2b2 Web Client**: The i2b2 Web Client is one of several core projects that are directly sponsored by the i2b2 team.

- **i2b2 Sponsored Project - NCBO Ontology Tools**: Tools to extract and organize ontologies from the NCBO, organized by Lori Phillips.

- **Related Project - Clinical Trender**: The Clinical Trender aims to allow researchers to track and visualize certain clinical variables related to a selected patient.

- **Related Project - CRC Tester**: A Workbench plugin that tests the CRC web services by Mike Mendis.

- **Related Project - Crimson**: A project to make specimens available through i2b2 infrastructure, organized by Lynn Bry.

- **Related Project - ExportXLS**: The i2b2 Web Client Plugin that tabulates patient data & applicable specified concepts, and facilitates export to spreadsheet formats.

- **Related Project - FACE caGrid CQL2 Data Source**: A caGrid TRIAD data service that runs i2b2 queries via the RESTful interface to CRC.

- **Related Project - Familial, Associational, & Incidental Relationships (FAIR) Initiative**: A collection of DBA tools and webclient plugins to facilitate the identification of related concepts amongst related patients.

- **Related Project - Federated Query Simulations**: Simulations of federated query tools that return aggregate counts, such as SHRINE, by Griffin Weber.
Genotype Data

https://community.i2b2.org/wiki/display/IGD/Loading+Genomic+VCF+Files+into+i2b2
Use NLP to extract the relevant features from the set of patient notes.

- **Social History**: The patient is a **non-smoker**. No alcohol.
- **Non-Smoker**
- **Social History**: Negative for tobacco, alcohol, and IV drug abuse.
- **Past Smoker**
- **Social History**: The patient lives in rehab, married. Information from admission notes...
- **Unclear smoking history**
- **Hospital Course**: … It was recommended that she receive … We also added Lactinax, oral form of Lactobacillus … acidophilus to repopulate her gut.
- **Hard to pick**
- **Hospital Course**: SH: widow, lives alone, 2 children, no **tob/alcohol**.
- **Hard to pick**
LMM Enhanced interaction with Patient Representation
Medical conditions supported by description in chart

Emily Alsentzer et al
Zero-shot Interpretable Phenotyping of Postpartum Hemorrhage Using Large Language Models
medRxiv preprint doi:
https://doi.org/10.1101/2023.05.31.23290753
Data Integration in Big Data Commons

Electronic Medical Record (EMR) Data
- RPDR
  - Coded Data
    - Demographics
    - Diagnoses
    - Lab Results
  - Text Data (Notes/Reports)
    - Medications
    - Procedures
    - Visits
    - Physician Notes
    - Imaging Reports
    - Pathology Reports
    - Surgery Notes

Informatics Tools
- Calculated Controls (Charlson Index)
- Data Visualization
- Annotation
- Extract Data
- Natural Language Processing

Additional Data
- Survey Data

Genetic Data
- GWAS

Biobank Data
- DNA
- Serum
- Plasma

Consent Status
- Recontact
- Consent Status

Validated Phenotypes
- Type II Diabetes
- Coronary Artery Disease
- Congestive Heart Failure
- Multiple Sclerosis
- Rheumatoid Arthritis
- IBD
- Bipolar Disorder

Research
Curating a Disease Algorithm

1. **Create a gold standard training set.**

2. **Create a comprehensive list of features from patient’s electronic data that describe the disease of interest.**

3. **Develop the classification algorithm.** Using the data analysis file and the training set from step 1, assess the frequency of each variable. Remove variables with low prevalence. Apply adaptive LASSO penalized logistic regression to identify highly predictive variables for the algorithm.

4. **Apply the algorithm to all subjects** in the superset and assign each subject a probability of having the phenotype.
## Biobank Portal | Curated Diseases

<table>
<thead>
<tr>
<th>Validated Phenotype</th>
<th>Count*</th>
<th>Predictive Positive Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bipolar Disease</td>
<td>71</td>
<td>89%</td>
</tr>
<tr>
<td>Congestive Heart Failure</td>
<td>387</td>
<td>90%</td>
</tr>
<tr>
<td>Coronary Artery Disease</td>
<td>2,420</td>
<td>97%</td>
</tr>
<tr>
<td>Crohn’s Disease</td>
<td>453</td>
<td>90%</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>94</td>
<td>90%</td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>550</td>
<td>90%</td>
</tr>
<tr>
<td>Type 2 Diabetes Mellitus</td>
<td>1,887</td>
<td>97%</td>
</tr>
<tr>
<td>Ulcerative Colitis</td>
<td>330</td>
<td>90%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Healthy Controls based on Charlson Index</th>
<th>Count**</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 10-year survival probability is &gt;98.3%</td>
<td>2,206</td>
</tr>
<tr>
<td>1 – 10-year survival probability is &gt;95.87%</td>
<td>4,343</td>
</tr>
<tr>
<td>2 – 10-year survival probability is &gt;90.15%</td>
<td>6,545</td>
</tr>
</tbody>
</table>

* Based on 15,880 patients
** Based on 21,300 patients
Automated Learning Algorithms enabled in RPDR such as PheNorm Algorithm.
Machine Learned Phenotypes

- Abdominal hernia
- Acute bronchitis and bronchiolitis
- Acute pancreatitis
- Alcoholism
- Alzheimer's disease
- Aortic aneurysm
- Aplastic anemia
- Atrial fibrillation
- Atrioventricular block
- Autism spectrum disorders
- Basal cell carcinoma
- Bipolar Disease
- Bladder cancer
- Brain cancer
- Breast cancer
- Cerebral aneurysm
- Cholelithiasis
- Chronic pancreatitis
- Chronic sinusitis
- Coronary atherosclerosis
- Crohn's disease
- Deep vein thrombosis
- Depression
- Diverticulosis and diverticulitis
- Eating disorder
- Epilepsy
- Gastroesophageal reflux disease
- Gout
- Heart valve disorders
- Hyperlipidemia
- Hyperparathyroidism
- Hypertension
- Hypothyroidism
- Insomnia
- Intracranial hemorrhage
- Ischemic stroke
- Leukemia
- Lung cancer
- Melanoma
- Migraine headache
- Multiple sclerosis
- Myocardial infarction
- Neutropenia
- Non-Hodgkin lymphoma
- Obesity
- Obsessive compulsive disorder
- Obstructive sleep apnea
- Ovarian cancer
- Pancreatic cancer
- Parkinson's disease
- Peripheral vascular disease
- Pneumonia
- Polycystic ovaries
- Prostate cancer
- Pulmonary heart disease
- Renal cancer
- Renal failure
- Schizophrenia
- Substance addiction
- Suicidal ideation
- Suicide attempt or self-inflicted injury
- Thyroid cancer
- Tobacco use disorder
- Type 1 diabetes
- Type 2 diabetes
- Ulcerative colitis
- Urinary calculus
- Uterine cancer
### Phenotype Automation: Phenotype Quality Dashboard

#### Table:

<table>
<thead>
<tr>
<th>Category</th>
<th>PhEWS_code</th>
<th>abbr</th>
<th>PhEWS_name</th>
<th>model</th>
<th>ICD_PPV</th>
<th>ICD_AUC</th>
<th>AUC</th>
<th>PPV</th>
<th>TPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONC</td>
<td>PheWAS:189.21</td>
<td>BLCA</td>
<td>Bladder cancer</td>
<td>PheNorm_ICD</td>
<td>0.80</td>
<td>0.903</td>
<td>1.00</td>
<td>1.00</td>
<td>0.42</td>
</tr>
<tr>
<td>ONC</td>
<td>PheWAS:204</td>
<td>LEUK</td>
<td>Leukemia</td>
<td>PheNorm_ICD</td>
<td>0.73</td>
<td>1.000</td>
<td>1.00</td>
<td>1.00</td>
<td>0.91</td>
</tr>
<tr>
<td>PSYCH</td>
<td>PheWAS:297.1</td>
<td>SI</td>
<td>Suicidal ideation</td>
<td>PheNorm_ICDNLNP</td>
<td>0.93</td>
<td>0.786</td>
<td>1.00</td>
<td>1.00</td>
<td>0.43</td>
</tr>
<tr>
<td>PSYCH</td>
<td>PheWAS:305.2</td>
<td>EATD</td>
<td>Eating disorder</td>
<td>PheNorm_ICDNLNP</td>
<td>0.53</td>
<td>0.482</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>NEURO</td>
<td>PheWAS:327.4</td>
<td>INSOM</td>
<td>Insomnia</td>
<td>PheNorm_ICDNLNP</td>
<td>0.93</td>
<td>0.821</td>
<td>1.00</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>CARDIO</td>
<td>PheWAS:452.2</td>
<td>DVT</td>
<td>Deep vein thrombosis</td>
<td>PheNorm_ICDNLNP</td>
<td>0.87</td>
<td>0.692</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>NEURO</td>
<td>PheWAS:817</td>
<td>CONC</td>
<td>Concussion</td>
<td>PheNorm_NLP</td>
<td>0.73</td>
<td>0.682</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>METAB</td>
<td>PheWAS:250.1</td>
<td>TIDM</td>
<td>Type 1 diabetes</td>
<td>PheNorm_ICD</td>
<td>0.17</td>
<td>0.882</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>ONC</td>
<td>PheWAS:184.11</td>
<td>OVCA</td>
<td>Ovarian cancer</td>
<td>PheNorm_ICDNLNP</td>
<td>0.60</td>
<td>0.926</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>ONC</td>
<td>PheWAS:182</td>
<td>UTCA</td>
<td>Uterine cancer</td>
<td>PheNorm_ICD</td>
<td>0.50</td>
<td>0.867</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>GI</td>
<td>PheWAS:555.1</td>
<td>CD</td>
<td>Crohn’s disease</td>
<td>PheNorm_mean</td>
<td>0.54</td>
<td>0.961</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

#### Diagram:

- [Cardiology](#)
  - Aortic aneurysm (AA)
  - Atrial fibrillation (AFIB)
  - Atrioventricular block (AVB)
  - Coronary atherosclerosis (CAD)
  - Deep vein thrombosis (DVT)
  - Heart valve disorders (HVD)
  - Hypertension (HTN)
  - Myocardial infarction (MI)
  - Peripheral vascular disease (PVD)
  - Pulmonary heart disease (PHD)
High Quality Phenotypes for Research Studies

Image of a computer interface showing a query tool for selecting patient phenotypes. The interface includes menus for navigating and querying biobank data, with specific conditions such as 'Primary dilated cardiomyopathy - 4002' and 'CHF - current or past history (PPV 0.90) - 700'. The query results show that 70 patients meet the specified criteria.
Combined with Generative AI can produce Digital Twin of Patient

Digital Twin – Abigail Test

- EHR
- Genomics
- IoT
- Billing
- Imaging

Study Specific Conditions and Common Data Elements

Core Conditions
High Throughput Methods for supporting Translational Research

- Set of patients is selected from medical record data in a high throughput fashion

- Investigators explore phenotypes of these patients using Machine Learning tools and a translational team developed to work specifically with medical record data

- Distributed networks cross institutional boundaries for phenotype selection, public health, and hypothesis testing

- Digital medicine is delivered into clinical care through Digital Twin
I2b2 Implementations
>250 across the USA and Internationally, some illustrated below:
Federated Queries

Mass General Brigham

Boston Children’s Hospital

BIDMC

Boston Health Net (BMC and Community Health Centers)

Columbia U. Medical Center and New York Presbyterian Hospital

University of California, Davis

Wake Forest Baptist Medical Center

Morehouse/Grady/RCMI

Washington University in St. Louis

U Texas Health Science Center/Houston
Drive Pragmatic Clinical Studies
RECOVER Study Data Harmonization

RECOVER: Researching COVID to Enhance Recovery

The National Institutes of Health (NIH) created the RECOVER Initiative to learn about the long-term effects of COVID.

The goal of RECOVER is to rapidly improve our understanding of and ability to predict, treat, and prevent PASC (post-acute sequelae of SARS-CoV-2), including Long COVID.

LEARN MORE ABOUT LONG COVID

https://recovercovid.org
Data harmonized within i2b2 star schema

Data Sources
- Acute and Post-Acute COVID-19 recovery cohort trials & autopsy cohort studies
- Biospecimen data
- EHR / FHIR data
- Mobile & digital health data platform
- Other PASC-related studies

i2b2 Format
- Common Data Elements

Core i2b2 Data Index
- ETL and updating
- UUID management
- Ontology management
- Payload management
- Provenance management
- Query Tool for Cohorts

Researcher Workspaces
- Analytical Tools
  - Jupyter, R, SAS, etc.
- Exporting data for analysis
- Create Datasets

Data Resource Core

RECOVER
Researching COVID to Enhance Recovery
Concepts in database available in harmonized ontology

Ontology-driven normalization of source data
New i2b2 Query Tool to be released:

https://i2b2transmart.org/2023-i2b2-symposium/2023-symposium-recordings-slides/
I2B2 AI

User asks a question

I2b2 Query & Analysis Tool

I want to find all patients with a diagnosis of acute respiratory infections who have also been prescribed bronchodilators.

Result rendered in web client
AI-ENABLED QUERY BUILDER:
(I.E. INSTRUCTION-TUNED POC)

Data Prep

- Data Cleaning
- Randomize
- Tokenize

Fine Tuning

- Code Llama built on Llama 2, trained on 500B tokens of code
- LORA 4-Bit Quant
- nmitchko/i2b2-querybuilder-34b-merged

Training Sandbox (@Mitchko Labs)

i2b2

- Export Query Log (XML)

i2b2 API

i2b2 XML Query

VAL

Validation

XML INSTRUCT

- Web UI
  - Query Builder
- Inference
  - Code Llama (LLaMa2)
  - LORA 4-Bit Quant

i2b2 LLM POC Cell (@MGB)

Code Llama built on Llama 2, trained on 500B tokens of code
High Throughput Methods for supporting Translational Research

- Set of patients is selected from medical record data in a high throughput fashion

- Investigators explore phenotypes of these patients using Machine Learning tools and a translational team developed to work specifically with medical record data

- Distributed networks cross institutional boundaries for phenotype selection, public health, and hypothesis testing

- Digital medicine is delivered into clinical care through Digital Twin
Congestive Heart Failure

- Affects 2% of the adult population
- Risk of death first year after diagnosis: 35%
- Inpatient hospital costs in 2011: $10.5B which is a small fraction of all heart failure related care
Early Detection of Worsening or Improving Anemia

**Background and Methods**
- Anemia is one of the strongest predictors of morbidity and mortality in CHF.
- Increasing or decreasing HGB is a further strong predictor, but there is no good way to determine whether a patient’s HGB is on its way up or down (Circulation. 2005;112:1121-1127)

**Results and Conclusions**
- A novel mathematical model of the RBC lifecycle enables estimation of patient-specific rates of RBC maturation and turnover from a routine CBC.

**Applications**
1. CHF patients most likely to have decreasing HGB may benefit from altered treatment or longer hospitalization to avoid readmission.
2. CHF patients most likely to have increasing HGB may be responding well to treatment and benefit from earlier discharge or maintenance of current therapy.
Creatinine Prediction: Hypothetical Application

- Hypothetical analysis of creatinine times series where possible treatments are introduced into the model.

- The model hypothetically provide a future trajectory conditioned on each treatment.
Population Based Predictive Analytics to Support Improved Decision Making

Exploring Integration of MGH Path renal predictive model

Heart Failure Physiology Tool

John Doe

Date of Birth: January 1, 1940
MRN: 1234567890

Actions Under Consideration
Prescribe 30mg / day HCT
Discharge Patient

Longitudinal Data

Predicted Events
Length of Stay: 1 days -> 0 days
30 Day Readmission: 30% -> 50%

PROJEXED RESULTS OF INTERVENTION

Notes:
Bringing Big Data into Clinical Care with Open App Development

Clinician

Analytic Calculation Engine

Core Integration Database

SMART App embedded in Epic

Epic Data Repository

FHIR interface for real time updates

DATA
GeneInsight, mHealth, ePath, Medical Images, 25 years of Legacy electronic data, and Other External Systems

Laboratory Personnel

SMART App in Lab

Non-EHR Users View Standalone App

Smart App in Lab

Analytics have direct access to repository
Transforming Care in the Digital Age

Digital and IoT devices continuously output Patient Data

Digital Twin of patient enables continuous assessment of patient with Real Time Algorithms

Navigator Model dramatically increases Frequency and Convenience for Patient Communication

System drives Pragmatic Clinical Trials Leading to Continuous Process Improvement
MGB Data Enclave Overview

Network Boundary of MGB Data Enclave

Investigator
Specific Calculations

Digital Twin
Large Language
Model Training

Clinical Data Hub
Current EHR Data
Epic

Legacy EHR Data

I2b2 Vocabulary
Harmonization

OMOP CDM v5.3

I2b2 CDM

Loyalty Cohort

Computed Phenotypes

Edge Computing

DIGITAL TWIN

Indexed Notes Repository

Unstructured Data
Hub

Indexed VCF

Genomics Data Hub

Indexed Quantitative
Imaging

Image Data Hub

Clinical Trial Data

Clinical Trial Data Hub

I2b2 index Visualization

REST API

REST API

ETL

ETL

ETL

REST API

REST API

Git Lab

Investigator
Specific Calculations

I2b2 index Visualization

REST API

REST API

ETL

ETL

ETL

REST API

REST API

ETL

ETL

REST API

REST API
I2b2 tranSMART Software

i2b2 Homepage (https://www.i2b2.org)
i2b2 Software (https://www.i2b2.org/software)
i2b2 Community Site (https://community.i2b2.org)
https://i2b2transmart.org/2023-i2b2-symposium/2023-symposium-recordings-slides/