i2b2
Back to the Future

Isaac S. Kohane, MD, PhD
High throughput phenotyping to find rare patients

- Search routine clinical records from 5 major hospitals for:
  - Demographics
  - Diagnosis
  - Medications
  - Lab Results

- Reach N
  - Rare Dx
  - Small Effects

- 10 billion FACTS
- 6 million patients
Finding rare events of interest.

Importance of real-time exploration

Cardiac angiogenic imbalance leads to peripartum cardiomyopathy

samples from subjects with PPCM have been previously described. Patients in both studies were predominantly Caucasian. Retrospective analyses of PPCM and pre-eclampsia in the Harvard teaching hospitals were performed using the Harvard Shared Health Research Information Network (SHRINE), a de-identified repository of aggregate patient information.
Unbiased clustering
Autism or Autisms?
Reproducibility
Reis, et al., BMJ 2009

Jane Q. Doe
ED Profile
June, 1992 – Feb, 1999
Cusum of Hospitalizations due to Myocardial Infarction

- Rofecoxib introduced
- Rofecoxib withdrawn

Care Behavior + Health => Healthcare Dynamics

Join Griffin

Percent of Patients

Patient Age in Years

Deceased
>5 Chronic
2-5 Chronic
1 Chronic
>1 Acute
1 Acute
No Diagnoses
Horizontal blue curves with rectangular labels are five year fact count percentiles. Vertical red curves with hexagonal labels are three year survival curves. Data are from BWH and MGH from 7/28/2001 to 7/27/2009. All patients had at least one fact between 7/28/2005 and 7/27/2006. The “current” age is the patient age on 7/27/2006. This chart represents only male patients.
Normal Physician Behavior

Analysis of WES Protein Ratios HLA-DQB1 for subsets:

Comparison of WES Protein Ratios HLA-DQB1

Histogram of WES Protein Ratios HLA-DQB1

 Subset 1

- WES>Loading
- Mean: 1.05
- Median: -0.01
- IQR: 3.33
- SD: 1.81
- Data Points: 54

Trial
“Hi, honey – I’m home!”
TriNetX Builds off of i2b2

- Supports local data access and formation of ad hoc research networks
- Facilitates translational research for internal and/or sponsored studies
- Contributes to sustainability of your research enterprise
Where To?
Figure. The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health Care
Environment-Wide Association Study (EWAS)

- New associations in 2+ cohorts: gamma tocopherol, heptachlor epoxide, PCB
- Known associations: vitamin D, beta carotene
- Interesting: hepatitis B

Insights From Population Health Data

1. Public
2. Healthcare Workers, Electronic Health Records
3. Researchers, Industry
5. Regulatory Agencies
A Community of Patient-centered Information Commons
Managing the Dance Between Computational Resources and Data Location
Challenge: Data Imputation and Fusion

Data Imputation & Fusion
Algorithm Predicting $D_{RC}$ and/or $D_{PRO}$

- Patient B1
- Patient B$n_B$
- Patient E1
- Patient E$n_E$

Predicted $D_{RC}$
The second $\frac{1}{2}$ of our mission: Afferent and **Efferent** Arms of Healthcare Information
No Small Change for the Health Information Economy

Kenneth D. Mandl, M.D., M.P.H., and Isaac S. Kohane, M.D., Ph.D.

Four Propositions

1. Liquidity of data. The platform and its applications should reduce impediments to the transfer of data, in an agreed-upon form, from one system to another.

2. Substitutability of applications. The system should be sufficiently modular and interoperable so that ... just as consumers may swap out applications on their iPhones, physicians should be able to readily replace one ... system with another.

3. Platform should be built to open standards accommodating both open-source and closed-source software.

4. Platform that supports diverse applications will lead to a robust health information economy ... by allowing competition and "natural selection" for high-value, low-cost products.
State-of-the-Art ... in 2011???
SMART Platform*

• Data Model
  Medical Context, Medical Record Elements, Data Coding

• Application Programming Interface (API)
  Consistent developer data access mechanisms

• User Interface
  Standards-based integration, flexibility

• Authentication
  In-browser, server-to-server
Our vocabulary

• **Data Sources**
  Managed by containers

• **Containers**
  Present data from data sources to apps in a uniform fashion

• **Apps**
  Completely substitutable
Inspired by a

18.12 Issue

Design Challenge
1 Design + 1 Developer + 1 Week
SMART on FHIR® at HIMSS 14
The Effect Arm Opportunity

• Clinical Trials (PCORI & CTSA)
• Direct to/from patient
• Integration with home-monitoring
Methodological Agenda

• Biomedical relevance of data science → expansive view of what constitutes biomedical data.

• The investment in healthcare can be leveraged for data science with incremental $.
  – Challenge: Randomization in retrospective-prospective and other \textit{in silico} trial designs.
  – Linkage across institution and data modalities
  – Generalizing for methods of data-type dependent imputation, de-noising
Federated and Federalist
Federated and Federalist
Federated and Federalist
Federated and Federalist