Why and How You Get the Whole Country’s Healthcare Harnessed for Biomedical Science and Care.

Isaac S. Kohane, MD, PhD

8 year old initiative

- Center for Biomedical Informatics
- Countway Library
- Why?

Jane Q. Doe
ED Profile
June, 1992 – Feb, 1999

Reis, et al., BMJ 2009
Narrative data (NLP text extractions)

Codified data (ICD9 codes, etc)

HEGP transSMART

Cohorts

Category

Time

Select a variable on which you would like to work and drag it into the box. For example, "Stage" or "Age". If the variable is categorical (e.g., "Stage"), drag it into the box and then "bin" using the option below. This variable is not required.

Variable Selection

Analysis: Survival Analysis

Sample Explorer

Gene Signature/Lists

Search to Select

Navigate Terms

Across Tools

Comparison

Advanced Workflow

Results/Analysis

Data View

Data Export

Exp. Info

ирранSMART
**Common-Rare: Weak-Strong Spectrum**

- **R** module in tranSMART
- Published figure in JCO

**SHRINE: Governance over Technology**

- 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

**Needs**

- Rare & COMMON
- BIG DATA
Finding rare events of interest.

Cardiac angiogenic imbalance leads to peripartum cardiomyopathy

17 MAY 2012 | VOL 485 | NATURE | 333

samples from subjects with PPCM have been previously described. Patients in both studies were predominantly Caucasian. Retrospective analysis of PPCM and pre-eclampsia in the Harvard teaching hospitals were performed using the Harvard Shared Health Research Information Network (HSRN) is a de-identified repository of aggregate patient information.

Importance of real-time exploration

Multisite-multidrug detection

i2b2 network

Ad hoc, bottom up and top down

Clin Pharmacol Ther. 2011 Jul; 90(1): 133–1
Results

Figure 2c, a similar pattern exists for other tests, such as high-density lipoprotein cholesterol (HDLc) and prostate-specific antigen (PSA), with a large variance for most tests. However, the repeat intervals can be decreasing at larger or smaller values. As seen in Figure 2b and 2c, the repeat intervals for WBC, HDLc, and PSA are larger within the hospital reference ranges (indicated by markers on the horizontal axis) than outside. However, it is not a binary response. Rather, the repeat intervals vary with different initial values. Although WBC is highly dependent on the initial value of the test as well as the patient population and clinical setting. The next three sections describe this relationship by testing three hypotheses.

Identifying subpopulations

To determine if we can automatically identify the various factors that can influence physician behavior, such as patient demographics and clinical settings, we calculated the median repeat intervals for WBC for different pediatric age groups and clinical settings. In the first part of this study, we used repeat intervals to examine normality in laboratory tests. Whereas laboratory test results can be divided into normal and abnormal, we hypothesized that repeat intervals might be detectable. Although we are not arguing that this method should replace physician group intelligence, it is interesting to examine normality in laboratory tests. Whereas laboratory test results can be divided into normal and abnormal, we hypothesized that repeat intervals might be detectable.

Identifying subpopulations

To determine if we can automatically identify the various factors that can influence physician behavior, such as patient demographics and clinical settings, we calculated the median repeat intervals for WBC for different pediatric age groups and clinical settings. In the first part of this study, we used repeat intervals to examine normality in laboratory tests. Whereas laboratory test results can be divided into normal and abnormal, we hypothesized that repeat intervals might be detectable. Although we are not arguing that this method should replace physician group intelligence, it is interesting to examine normality in laboratory tests. Whereas laboratory test results can be divided into normal and abnormal, we hypothesized that repeat intervals might be detectable.

Identifying subpopulations

To determine if we can automatically identify the various factors that can influence physician behavior, such as patient demographics and clinical settings, we calculated the median repeat intervals for WBC for different pediatric age groups and clinical settings. In the first part of this study, we used repeat intervals to examine normality in laboratory tests. Whereas laboratory test results can be divided into normal and abnormal, we hypothesized that repeat intervals might be detectable. Although we are not arguing that this method should replace physician group intelligence, it is interesting to examine normality in laboratory tests. Whereas laboratory test results can be divided into normal and abnormal, we hypothesized that repeat intervals might be detectable.
At a glance

- Patients: 13750 (with basic demographics)
- ~0.5% hospital population
- M:F (5:1)
- Diagnoses: 5627
- Laboratory measurements: 3,158,234 on 3581 lab measurement types
- Medications: > 800,000 Rx’s

Unbiased clustering

Prevalence relative to hospital population

![Graph showing prevalence relative to hospital population](image)

Autism or Autisms?

![Graph showing autism prevalence over years](image)
Age differences

Pneumonia/Influenza Mortality (CDC)
Influenza like illness (CDC)
Adult ED
Pediatric ED

Disease covered (n=179)
iPhone Submissions vs CDC sentinel surveillance

Bringing AHC and Patients Back Together

Science 2007
How to consider the environmental causes of disease?

- US CDC NHANES: four independent cross-sectional cohorts
- Blood and urine tests on 100s to 1000s
- Consider environmental causes of diseases like genetic causes
- Association between each environmental factor and T2DM?


Environment-Wide Association Study (EWAS)

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


NHANES: National Health and Nutrition Examination Survey

- Since the 1960s: 50+ years
- Now biennial: 1999 onwards
- 10,000 participants per cohort
- Disease prevalence estimates: T2D, obesity, cardiovascular disease
- Growth charts for development: WHO standard
- Environment: Elimination of lead -- 70% decline since ’70s
- Use NHANES as a pilot to survey environmental factors for disease

Do known disease-associated genetic variants interact with environmental factors found in EWAS?

Example: Type 2 Diabetes

known variants X top EWAS hits 

TCF7L2 CDKAL1 HHEX PPARG
tocopherol heptachlor PCBs

Candidate Gene Studies

GWAS

EWAS
Do known disease-associated genetic variants interact with environmental factors found in EWAS?

<table>
<thead>
<tr>
<th>rs1236634(SLC30A8)</th>
<th>OR (95% CI)</th>
<th>rs2237895(KCNQ1)</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>trans-β-carotene (low-1SD)</td>
<td>1.1 (0.8, 1.5)</td>
<td>trans-β-carotene (mean)</td>
<td>0.9 (0.8, 1.1)</td>
</tr>
<tr>
<td>trans-β-carotene (high+1SD)</td>
<td>2.1 (1.3, 3.2)</td>
<td>p-value (FDR)</td>
<td>0.0085</td>
</tr>
<tr>
<td>N(cases): 1000 (66)</td>
<td>p-value: 0.0085</td>
<td>N(cases): 1000 (66)</td>
<td>p-value: 0.0085</td>
</tr>
</tbody>
</table>

Disatisfaction breeds alternative models.

Business as Usual Not Working

- Clinical trials are not better than 20 years ago.
- 1/3 clinical trials by 20 largest US companies solely outside US
  - Glickman et al. NEJM 2009
- Gold-standard is expensive but increasingly irrelevant.
- 100’s of 1000’s patients required for common weak effects and rare events
- Lack of nuance in policy.

In era of big data, well-characterized patients are the only non-commodity

Storage/compute cycles are expensive but can be purchased.

1,000,000 arrays x 10000 genes

2500 individuals, 130 TB
In era of big data, well-characterized patients are the only non-commodity.

- WGS is no more expensive than electrolyte panel.
- BLUE button liquifies all data in any healthcare institution on behalf of the patient.
- Lifetime record is maintained by third party bonded data agency.
  - Includes everything we currently think of as health related but only the patient or guardian can provide access.

2023, more healthdata generated in 1 home than an ICU bed in 2013

What are the questions/challenges?

- Who is going to curate these data?
- What is the vector of FHx and genomic data?
- Is there a role for the healthcare system?
- How do we create a larger market for the information commons?
- What are we going to do to uniquely identify patients.
<table>
<thead>
<tr>
<th>Genomics</th>
<th>Systems Informatics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lou Kunkel</td>
<td>Susanne Churchill</td>
</tr>
<tr>
<td>Alal Eran</td>
<td>Rachel Ramoni</td>
</tr>
<tr>
<td></td>
<td>Shawn Murphy</td>
</tr>
<tr>
<td></td>
<td>Griffin Weber</td>
</tr>
<tr>
<td></td>
<td>Doug MacFadden</td>
</tr>
<tr>
<td></td>
<td>Bill Simons</td>
</tr>
<tr>
<td></td>
<td>Mike Mendis</td>
</tr>
<tr>
<td></td>
<td>Nich Wattanasin</td>
</tr>
<tr>
<td></td>
<td>Stan Shaw</td>
</tr>
<tr>
<td></td>
<td>Peter Szolovits</td>
</tr>
<tr>
<td></td>
<td>Kat Liao</td>
</tr>
<tr>
<td></td>
<td>Robert Plenge</td>
</tr>
<tr>
<td></td>
<td>Tianxi Cai</td>
</tr>
<tr>
<td></td>
<td>Finale Doshi</td>
</tr>
<tr>
<td></td>
<td>Roy Perlis</td>
</tr>
<tr>
<td></td>
<td>Jordan Smoller</td>
</tr>
<tr>
<td></td>
<td>David Kreda</td>
</tr>
<tr>
<td></td>
<td>Josh Mandel</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>Developmental Medicine</td>
</tr>
<tr>
<td>Nathan Palmer</td>
<td>Lenny Rappaport</td>
</tr>
<tr>
<td>Patrick Schmid</td>
<td>Chuck Nelson</td>
</tr>
<tr>
<td>Sek Won Kong</td>
<td>Stephanie Brewster</td>
</tr>
<tr>
<td>Luke Hutchinson</td>
<td></td>
</tr>
<tr>
<td>Bonnie Berger</td>
<td></td>
</tr>
<tr>
<td>Populat Science</td>
<td>Biotechnology</td>
</tr>
<tr>
<td>Ken Mandl</td>
<td>Daria Prilutsky</td>
</tr>
<tr>
<td>Ben Reis</td>
<td>David Margulies</td>
</tr>
<tr>
<td>John Brownstein</td>
<td></td>
</tr>
</tbody>
</table>