Married 6-year Olds and Other Diseases of Data

Michael G. Kahn MD, PhD

Department of Pediatrics, University of Colorado, Denver
Colorado Clinical and Translational Sciences Institute
Department of Research Informatics, Children’s Hospital Colorado

National Data Integrity Conference
Enabling Research : New Challenges & Opportunities
8 May 2015
Michael.Kahn@ucdenver.edu
Guide to the Presentation

• The fun stuff
  – What is “clinical and translational” data management?
  – The changing landscape of clinical research
  – Learning health systems
  – National data networks
• The grunt work to do the fun stuff
  – Data harmonization
  – Data quality
  – My database can’t count
The Changing View of Clinical Research

T1 Biomedical Research

Basic Research Data

Pilot Studies

New Research Questions

Study Design & Approval

EHR Data

Recruitment & Enrollment

Study Setup

Study Execution

Public Information

Required Data Sharing

Evidence-based Patient Care and Policy

Outcomes Reporting

Clinical Practice

EHR Data

Outcomes Research

Evidence-based Review

Submission & Reporting

From: C Broverman, Partners
Learning Health Systems:
Every patient contributes knowledge
ImproveCareNow: A Pediatric LHS


**Percentage Of Pediatric Inflammatory Bowel Disease Patients In Remission, 2007–14**

*Source* Data are from the ImproveCareNow pediatric inflammatory bowel disease registry for 2007–14. *Notes* Each blue dot represents the percentage of patients in remission among care centers with more than 75 percent of their patients enrolled in Improve CareNow in a given month. The figure shows the upper and lower confidence limits (dashed red lines in red) and the mean (green solid lines).
Explosive activity in “big data” and “big data analytics” in healthcare

PCORnet: The National Patient-Centered Clinical Research Network
Both researchers and funders now recognize the value in integrating clinical research networks.

- Linking existing networks means clinical research can be conducted more effectively.
- Ensures that patients, providers, and scientists form true "communities of research".
- Creates "interoperability" – networks can share sites and data.
PCORnet embodies a “community of research” by uniting systems, patients & clinicians

PCORnet:
A national infrastructure for patient-centered clinical research
11 CDRN and 18 PPRN awards

This map depicts the number of PCORI-funded Patient-Powered or Clinical Data Research Networks that have coverage in each state.
Guide to the Presentation

• The fun stuff
  – What is “clinical and translational” data management?
  – The changing landscape of clinical research
  – Learning health systems
  – National data networks

• The grunt work to do the fun stuff
  – Data harmonization
  – Data quality
  – My database can’t count
Harmonizing data into a common structure

- **Commercial CDR**
  - Other
  - EHR
- **Local Data Warehouse**
  - Other
  - EHR
- **Existing Clinical Registries**
  - Other
  - EHR

- **Limited Data Set**
- **Common Data Model**
- **Common Terminology**

**Common Query Interface**
Terminology Harmonization – What are we talking about?
Examples of Variations in Platelet (Quantitative) Result Units in Source Data

<table>
<thead>
<tr>
<th>Platelet count original result units‡</th>
<th>FL</th>
<th>K/CMM</th>
<th>TH/UL</th>
<th>THOU/CMM</th>
<th>X10(3)</th>
<th>1000/UL</th>
<th>X10(3)/MCL</th>
<th>X10(3)/UL</th>
<th>X10(6)/MCL</th>
<th>X10*9/L</th>
<th>X10E3/UL</th>
<th>X1000</th>
<th>X10X3</th>
<th>X10^3/UL</th>
<th>x10</th>
<th>X10?3/ul</th>
<th>X10E3/UL</th>
<th>X10E3</th>
<th>K/A?L</th>
<th>K/B5L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/100 W</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/CMM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 3 L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10X3UL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10^3/UL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10*3/uL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10?3/uL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10E3/uL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10e3/uL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10e9/L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E9/L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIL/L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bil/L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CU MM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Examples of Variations in (Qualitative) Pregnancy Result Units in Source Data (aka, how many ways can you spell negative?)

<table>
<thead>
<tr>
<th>NEGATIVE</th>
<th>POSITIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNDETERMINED</td>
<td>BORDERLINE</td>
</tr>
<tr>
<td>NONE DET</td>
<td></td>
</tr>
</tbody>
</table>

**COMMENT:**
- 160.8
- 0.5
- 1.2
- 1000
- 122
- 14
- 140
- 15
- 2
- 2
- 2.1
- 203
- 252.3
- 278
- 28
- 3178.2
- 345
- 38.1
- 400
- 5
- Int
- 5272.4
- 642.2
- 670
- 697.7

**DETECTED INDETERM**

<table>
<thead>
<tr>
<th>N</th>
<th>NOT DETE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neg</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>Negativ</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPRCS</td>
<td></td>
</tr>
<tr>
<td>TNP</td>
<td></td>
</tr>
</tbody>
</table>

**n**

| neg                | negative          |

**COMMENT:**
- 820
- 840
- 1615
- ABNORMAL
- BOLDERL
- BODERLIN
- CANCELLE
- DUPLICAT
- EQUIVOCAL
- EQUIVOCA
- HIRABAYA

**NE-CHECK**

<table>
<thead>
<tr>
<th>NEAGTIVE</th>
<th>NEGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGA T I</td>
<td>NEGA TIV</td>
</tr>
<tr>
<td>NGAT IV</td>
<td>NGATAIV</td>
</tr>
<tr>
<td>NEGATIBE</td>
<td>NEGATIAY</td>
</tr>
<tr>
<td>NEGATIVIE</td>
<td>NEGATIVIE</td>
</tr>
<tr>
<td>NEGATIVE</td>
<td>NEGITIVE</td>
</tr>
<tr>
<td>NEGITIVE</td>
<td>NETGATIV</td>
</tr>
</tbody>
</table>

**NORM**

<table>
<thead>
<tr>
<th>NORMAL</th>
<th>POA</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSTIV</td>
<td>POSTIVF</td>
</tr>
<tr>
<td>POSITTIVE</td>
<td>POSITIVE</td>
</tr>
<tr>
<td>POSITIVE</td>
<td>PSOITIVE</td>
</tr>
</tbody>
</table>

**REPEAT**

| STAT              | URINE             |

**COMMENT:**
- 17
# Integrated child-life stages for NICHD Pediatric Terminology as mapped to existing medical terminologies

<table>
<thead>
<tr>
<th>Age Span</th>
<th>In Utero</th>
<th>Preterm</th>
<th>0-1 month</th>
<th>1-12 months</th>
<th>1-2 years</th>
<th>2-5 years</th>
<th>5-11 yrs</th>
<th>12-21 yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNOMED</strong></td>
<td></td>
<td></td>
<td>infants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CDC</strong></td>
<td></td>
<td></td>
<td>infants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AAP</strong></td>
<td></td>
<td></td>
<td>infants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CDISC</strong></td>
<td></td>
<td></td>
<td>infants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ICH-E11</strong></td>
<td></td>
<td></td>
<td>infants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EPA</strong></td>
<td></td>
<td></td>
<td>infants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>From: Steven Hirschfeld MD, PhD</strong></td>
<td></td>
<td></td>
<td>infants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Infancy Stage**
- Infancy: Prenatal - 1 year
- Infancy Stage (birth – 12 months)
- Toddlers (1-2 yrs)
- Toddlers (2-3 yrs)
- PreSchooler (3-5 yrs)
- Middle childhood (6-8 yrs)
- Middle childhood (9-11 yrs)
- Early Adolescence (12 to 14 yrs)

**Fetal Stage**
- 0-<1 month
- 1-<3 months
- 3-<6 months
- 6-<12 months
- 1-<2 years
- 2-<3 years
- 3-<6 years
- 6<11 years
- 11<16 years
- 16<21 years

**Term Neonatal Stage**
- Term Neonatal Stage
- Term newborn infants 0-27 days

**Preterm Neonatal Stage**
- Preterm newborn infants
- Term newborn infants 0-27 days

**Early Childhood**
- Early Childhood: 1 year - 4 years
- Childhood: 5 - 10 years
- Adolescence: 11 – 21 years

**Children**
- Children (2-11 Years)
- Adolescents (12 to 16-18 years)

**Adolescent Stage**
- Adolescents (12 to 16-18 years)
- Late Adolescence (19-21 years)

**Infancy Stage**
- Infancy: Prenatal - 1 year
- Infancy Stage (birth – 12 months)
- Toddlers (1-2 yrs)
- Toddlers (2-3 yrs)
- PreSchooler (3-5 yrs)
- Middle childhood (6-8 yrs)
- Middle childhood (9-11 yrs)
- Early Adolescence (12 to 14 yrs)

**Fetal Stage**
- 0-<1 month
- 1-<3 months
- 3-<6 months
- 6-<12 months
- 1-<2 years
- 2-<3 years
- 3-<6 years
- 6<11 years
- 11<16 years
- 16<21 years

**Term Neonatal Stage**
- Term Neonatal Stage
- Term newborn infants 0-27 days

**Preterm Neonatal Stage**
- Preterm newborn infants
- Term newborn infants 0-27 days

**Early Childhood**
- Early Childhood: 1 year - 4 years
- Childhood: 5 - 10 years
- Adolescence: 11 – 21 years

**Children**
- Children (2-11 Years)
- Adolescents (12 to 16-18 years)

**Adolescent Stage**
- Adolescents (12 to 16-18 years)
- Late Adolescence (19-21 years)
Aligning Terminologies
SNOMED CT: A “mandated” clinical standard

- Sign and symptoms of attention deficit hyperactivity disorder
- ADHD
- attention deficit
- hyperactivity
- ADD

- loss of scalp hair
- scalp hair loss

- Epistaxis
- nosebleeds

From: Rachel Richesson PhD
Standardizing terminologies to accommodate disparate observational data sources

**Standardizing conditions:**
- System Organ Class (Level 5)
- High Level Group Terms (Level 4)
- High Level Terms (Level 3)
- Preferred Terms (Level 2)
- Low-level Terms (Level 1)

**Source codes:**
- MedDRA
- SNOMED-CT
- ICD-9-CM
- Read
- Oxmis

**Mapping:**
- Existing
- De Novo
- Derived

**Top-level classification (Level 3)**
- SNOMED-CT

**Higher-level classifications (Level 2)**
- SNOMED-CT

**Low-level concepts (Level 1)**
- SNOMED-CT

**Standardizing drugs:**
- Top-level concepts (Level 4)
- Classifications (Level 3)
- Ingredients (Level 2)
- Low-level drugs (Level 1)

**Source codes:**
- NDC
- GPI
- Multilex
- VA-NDF
- Multum
- ICD-9-PCS*
- CPT-4*

**Mapping:**
- NDF-RT
- ATC
- ETC
- Indications and Cl
Data Quality in Electronic Health Records

- Data collection tools optimized for efficiency
  - Text templates
  - Copy/paste

- Minimal data validation checks
  - Min/Max limits
  - Pick lists
  - Required fields

- Even “simple” stuff has problems
## A trivial example: Martial Status by Age

<table>
<thead>
<tr>
<th>Total</th>
<th>Divorced</th>
<th>Legally Separat</th>
<th>Married</th>
<th>Significant other</th>
<th>Single</th>
<th>Unknown</th>
<th>Widowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>-16.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-15.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-8.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A trivial example: Martial Status by Age

Would these results be worrisome?

<table>
<thead>
<tr>
<th>Age</th>
<th>Total</th>
<th>Divorced</th>
<th>Legally Separate</th>
<th>Married</th>
<th>Significant other</th>
<th>Single</th>
<th>Unknown</th>
<th>Widowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.00</td>
<td>36,932</td>
<td>33</td>
<td>3</td>
<td>1,806</td>
<td>81</td>
<td>420,944</td>
<td>70</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>768</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>13,660</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>21,350</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>24,960</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>27,861</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>29,933</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>30,932</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>27,861</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>24,198</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>22,522</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>20,283</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>18,705</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>17,340</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>16,510</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>15,792</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>15,354</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>15,474</td>
<td>2</td>
<td>0</td>
<td>19</td>
<td>1</td>
<td>15,439</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>15,208</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>15,181</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>16,000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>17,000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Should we be worried?

• No
  – Large numbers will swamp out effect of anomalous data or use trimmed data
  – Simulation techniques are insensitive to small errors

• Yes
  – Observed site variation may be driven by differences in data quality, not clinical practices
  – Genomic associations look for small signals (small differences in risks) amongst populations
“Big Data” and “Big Data Analytics”

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren’t seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, “It was like arriving at a conference reception and realizing you don’t know anyone. So you just stand in the corner sipping your drink—and you probably leave early.”

Harvard Business Review October 2012
Exploring prevalence of all diseases

Treemap displays 3 dimensions:
- **Size of rectangle**: Standardized prevalence
- **Color**: ‘Prevalence ratio’, comparing source prevalence with the overall community average
- **Hierarchy**: Aggregate SNOMED-CT concepts by MedDRA System Organ Class
Exploring prevalence of disease with standardize databases:
ex: Acute Myocardial Infarction
The Tale of A Trivial Data Request

• The original data request:

“For an upcoming grant application, how many patients were seen recently with neurofibromatosis-1 (NF-1) and scoliosis?”
The Tale of A Trivial Data Query

- Getting more specificity:
  - “Recently seen” = an encounter of any type since 1/1/2012
  - NF-1: ICD-9 code starts with “237.7”
  - Scoliosis: ICD-9 code starts with “737.3”
The Tale of A Simple Data Query

- First query result: N = 15

Clinical investigator did not believe this result even though we used her definitions.
The Tale of A Simple Data Query

• Drilling down:
  – This query required both diagnoses to be coded on the same encounter (event).

![Diagram]

- N(Pt)
- Encounter
  - Dx1 = NF-1
  - Dx2 = Scoliosis
- 1/1/2012 - today
The Tale of A Simple Data Query

- Second query:
  - NF-1 and Scoliosis diagnoses can be coded on different encounters, both within time window
  - N = 28

Diagram:
- \( N(Pt) \)
- \( Dx1 = NF-1 \)
- \( Dx2 = Scoliosis \)
- 1/1/2012 - today

Investigator still did not like the answer!
Table 1: Ten graphical diagrams representing the question: "How many ambulatory patients did I ("Provider = Kahn") see with diabetes mellitus (ICD-9 = 250.xx) and essential hypertension (ICD-9 = 401.xx) between January 1, 2009 and December 31, 2009?" Each diagram, when converted into a database query, returns a different result. \( N(Pt) \) = number of patients.
Guide to the Presentation

• The fun stuff
  – What is “clinical and translational” data management?
  – The changing landscape of clinical research
  – Learning health systems
  – National data networks

• The grunt work to do the fun stuff
  – Data harmonization
  – Data quality
  – My database can’t count

It is a wonderful time in this field!
The fun stuff >>> The grunt work
(And even the grunt work ain’t bad!)
Explosive activity in “big data” and “big data analytics” in healthcare

Married 6-year Olds and Other Diseases of Data

Michael G. Kahn MD, PhD

Department of Pediatrics, University of Colorado, Denver
Colorado Clinical and Translational Sciences Institute
Department of Research Informatics, Children’s Hospital Colorado

National Data Integrity Conference
Enabling Research : New Challenges & Opportunities
8 May 2015
Michael.Kahn@ucdenver.edu