CCTSI Colorado Clinical & Translational Sciences Institute
University of Colorado Denver

Married 6-year Olds and Other Diseases of Data

Michael G. Kahn MD, PhD

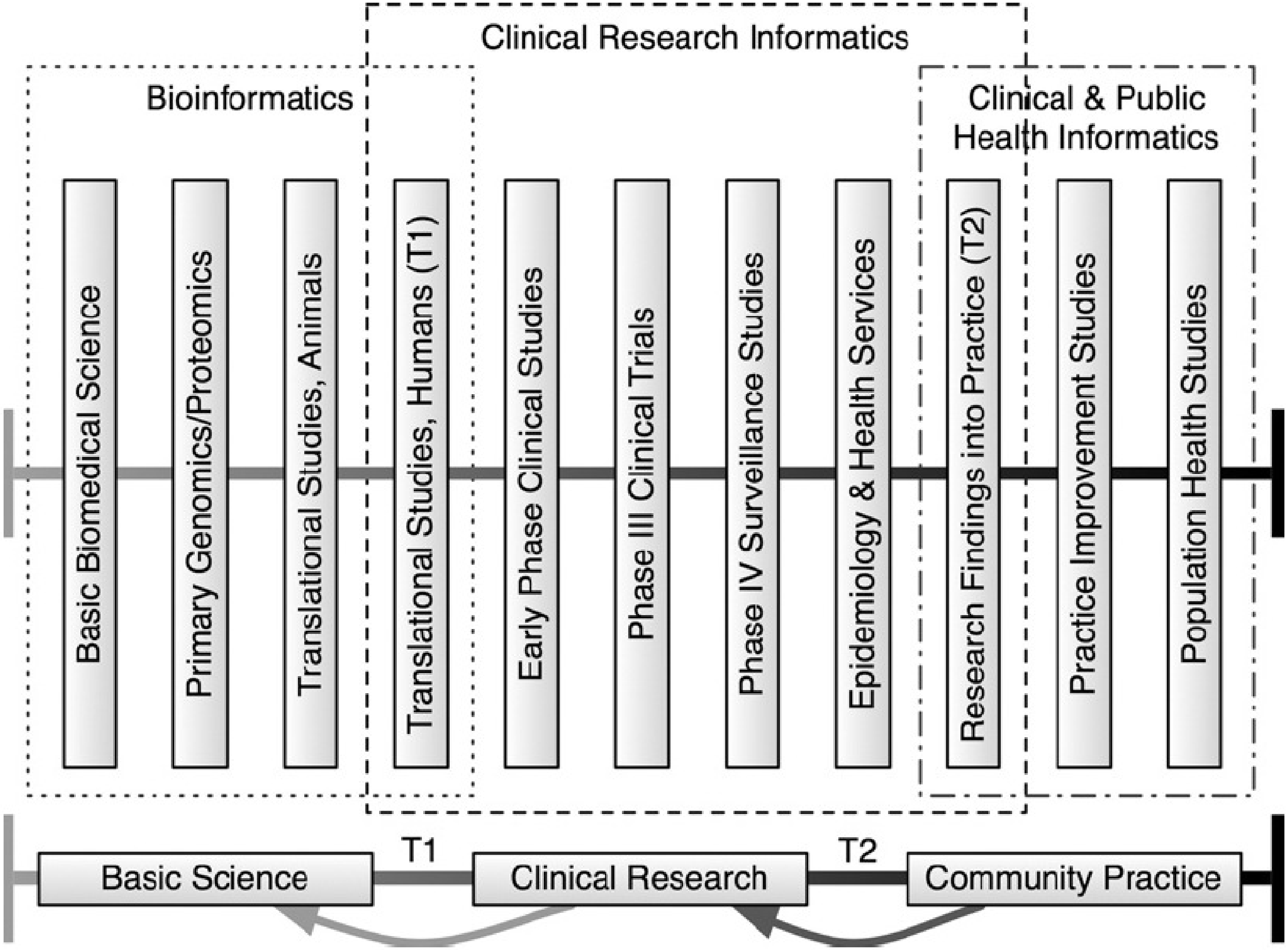
Department of Pediatrics, University of Colorado, Denver
Colorado Clinical and Translational Sciences Institute
Department of Research Informatics, Children’s Hospital Colorado

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

Funding was provided by a contract from AcademyHealth. Additional support was provided by AHRQ 1R01HS019912-01 (Scalable PArtnering Network for CER: Across Lifespan, Conditions, and Settings), AHRQ 1R01HS019908 (Scalable Architecture for Federated Translational Inquiries Network), and NIH/NCRR Colorado CTSI Grant Number UL1 RR025780 (Colorado Clinical and Translational Sciences Institute).

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



Clinical Research Informatics

Bioinformatics

Clinical & Public Health Informatics

Basic Biomedical Science

Primary Genomics/Proteomics

Translational Studies, Animals

Translational Studies, Humans (T1)

Early Phase Clinical Studies

Phase III Clinical Trials

Phase IV Surveillance Studies

Epidemiology & Health Services

Research Findings into Practice (T2)

Practice Improvement Studies

Population Health Studies

Basic Science

T1

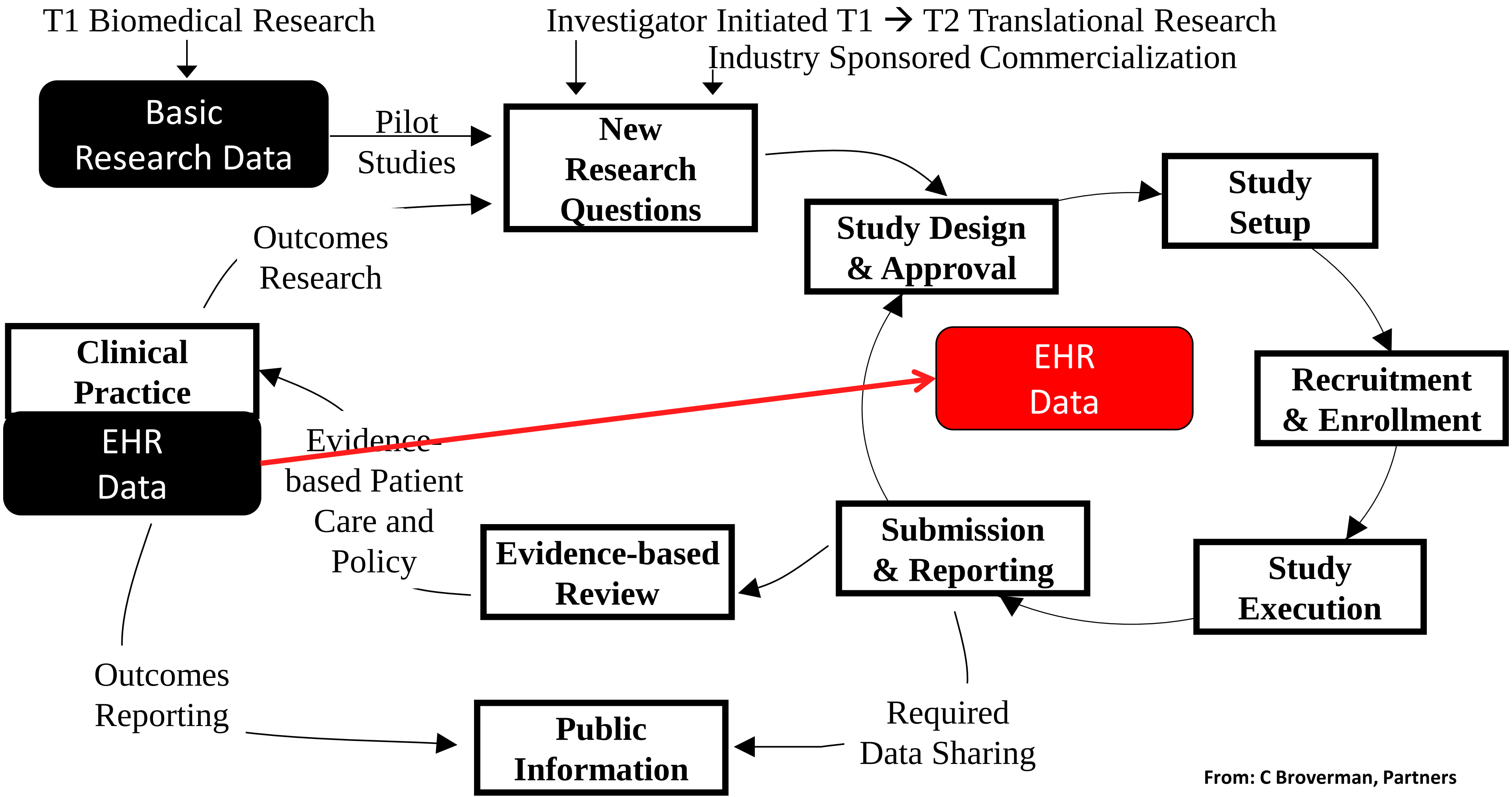
Clinical Research

T2

Community Practice

Embi, Payne: J. Am Med Inform Assoc 16(3) 2009

The Changing View of Clinical Research



T1 Biomedical Research

Investigator Initiated T1 → T2 Translational Research

Industry Sponsored Commercialization

Basic Research Data

Pilot Studies

New Research Questions

Study Design & Approval

Study Setup

Outcomes Research

Clinical Practice

EHR Data

EHR Data

Recruitment & Enrollment

Evidence-based Patient Care and Policy

Evidence-based Review

Submission & Reporting

Study Execution

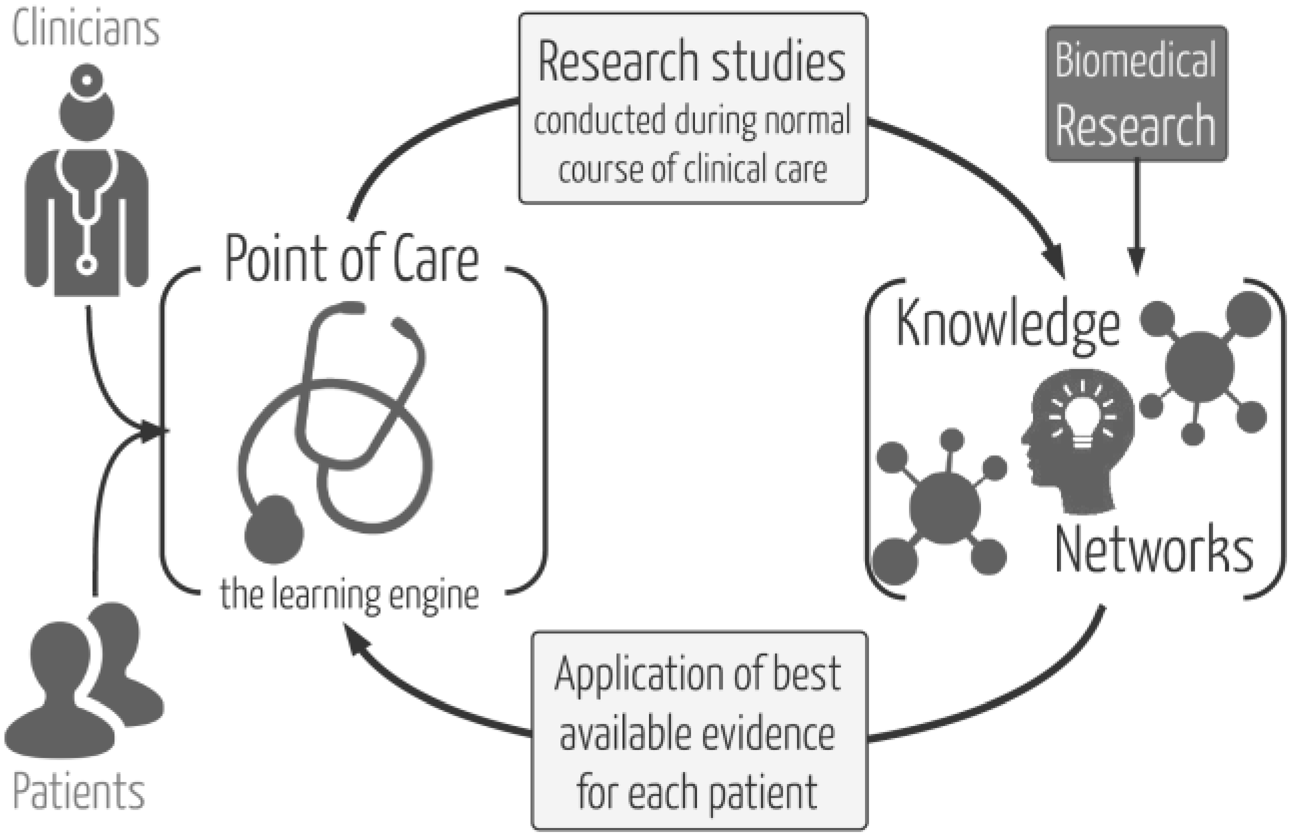
Outcomes Reporting

Public Information

Required Data Sharing

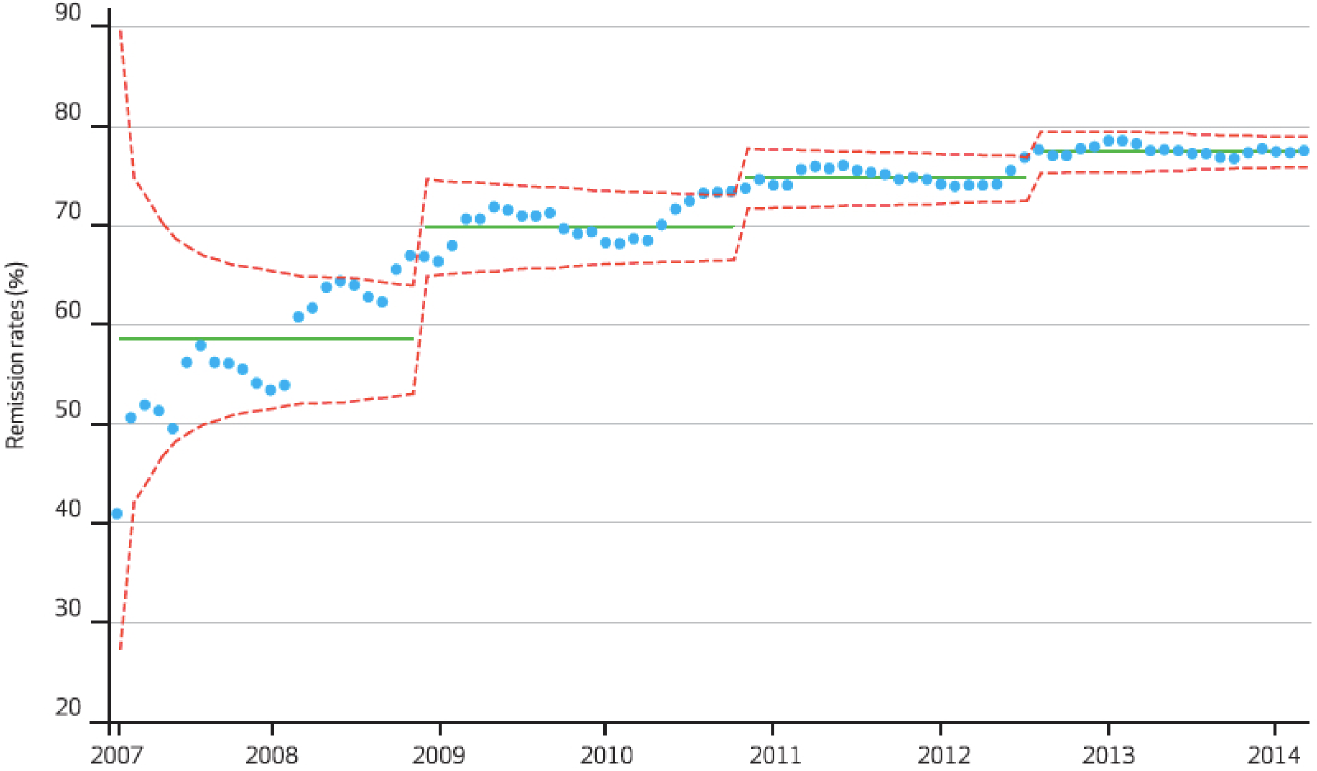
From: C Broverman, Partners

Learning Health Systems:
Every patient contributes knowledge


Clinicians
Research studies
conducted during normal course of clinical care
Biomedical Research
Point of Care
the learning engine
Knowledge
Networks
Application of best available evidence for each patient
Patients

ImproveCareNow: A Pediatric LHS

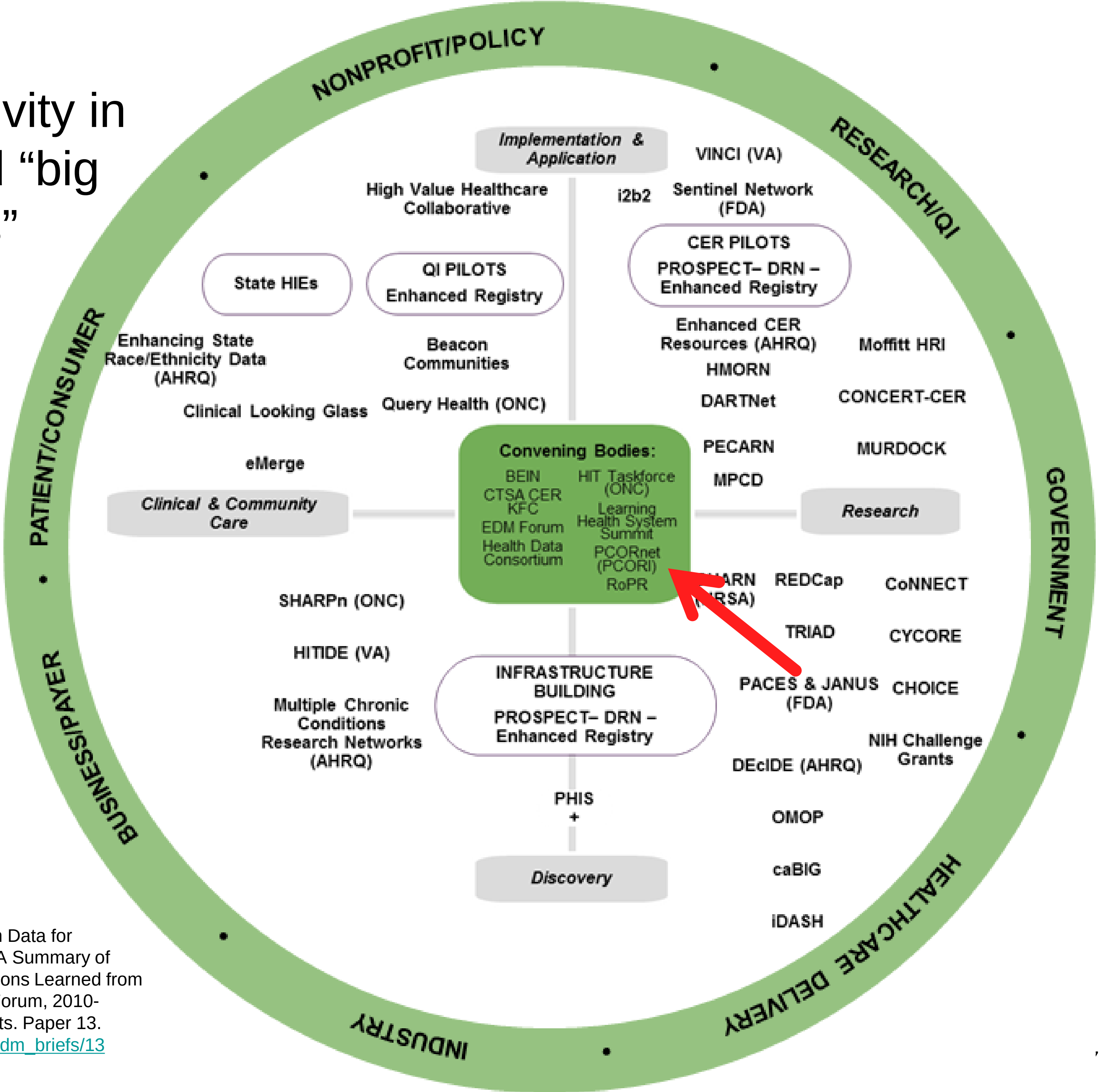
Percentage Of Pediatric Inflammatory Bowel Disease Patients In Remission, 2007–14



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).

Forrest CB, Margolis P, Seid M, *et al.* PEDSnet: How A Prototype Pediatric Learning Health System Is Being Expanded Into A National Network. *Health Aff* 2014;**33**:1171–7. doi:10.1377/hlthaff.2014.0127

Explosive activity in “big data” and “big data analytics” in healthcare



EDM Forum, "Using Electronic Health Data for Research and Quality Improvement: A Summary of Innovations, Achievements, and Lessons Learned from the Electronic Data Methods (EDM) Forum, 2010-2013" (2014). Issue Briefs and Reports. Paper 13. http://repository.academyhealth.org/edm_briefs/13

PCORnet:

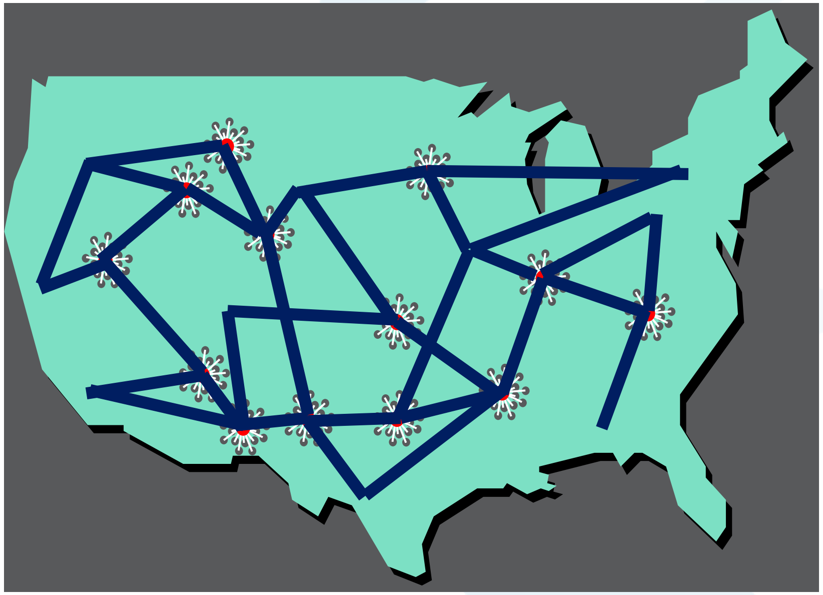
The National Patient-Centered Clinical Research Network

pcornet

The National Patient-Centered Clinical Research Network

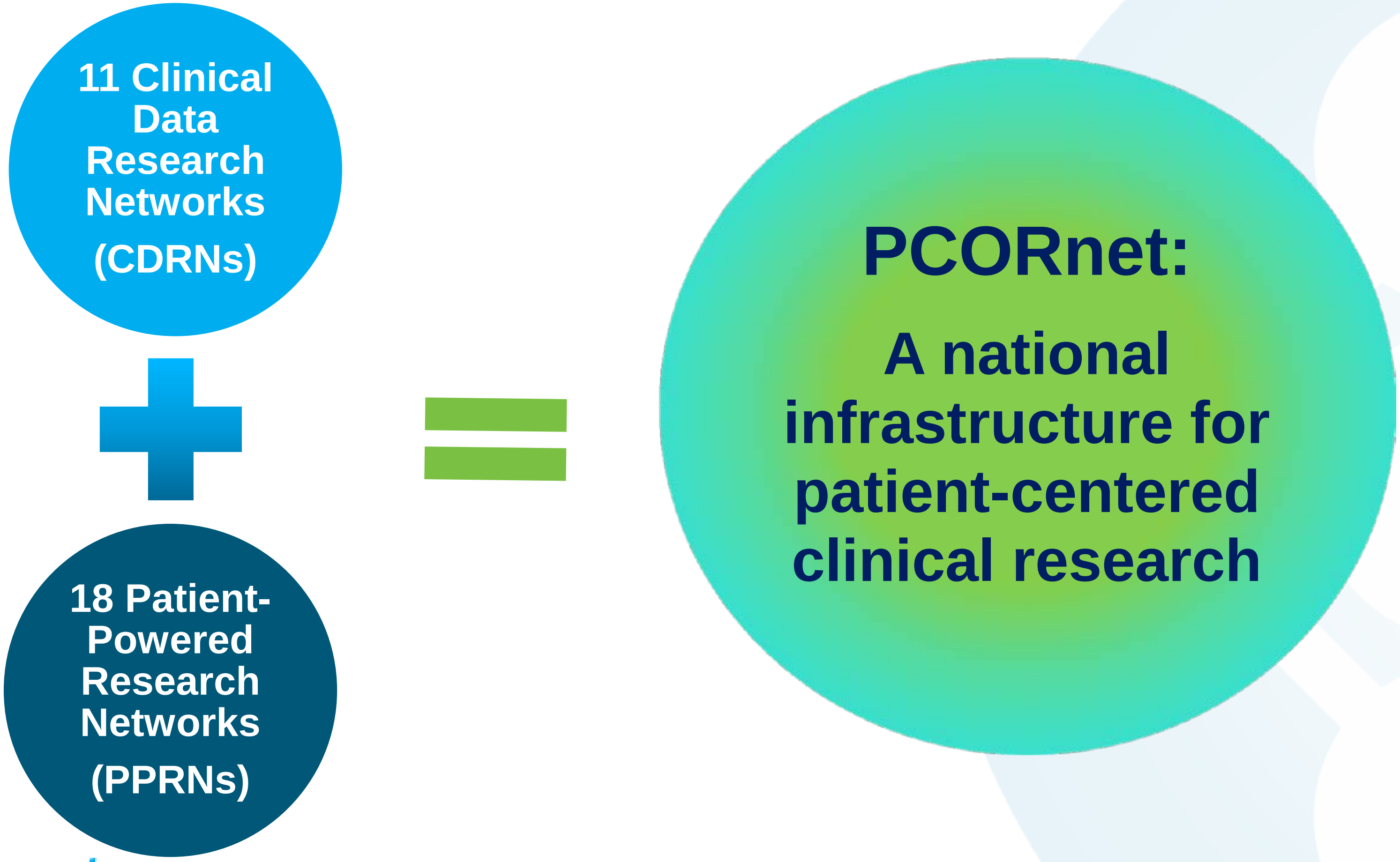
Both researchers and funders now recognize the value in integrating clinical research networks

- 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

PCORnet embodies a “community of research” by uniting systems, patients & clinicians



11 Clinical Data Research Networks

(CDRNs)

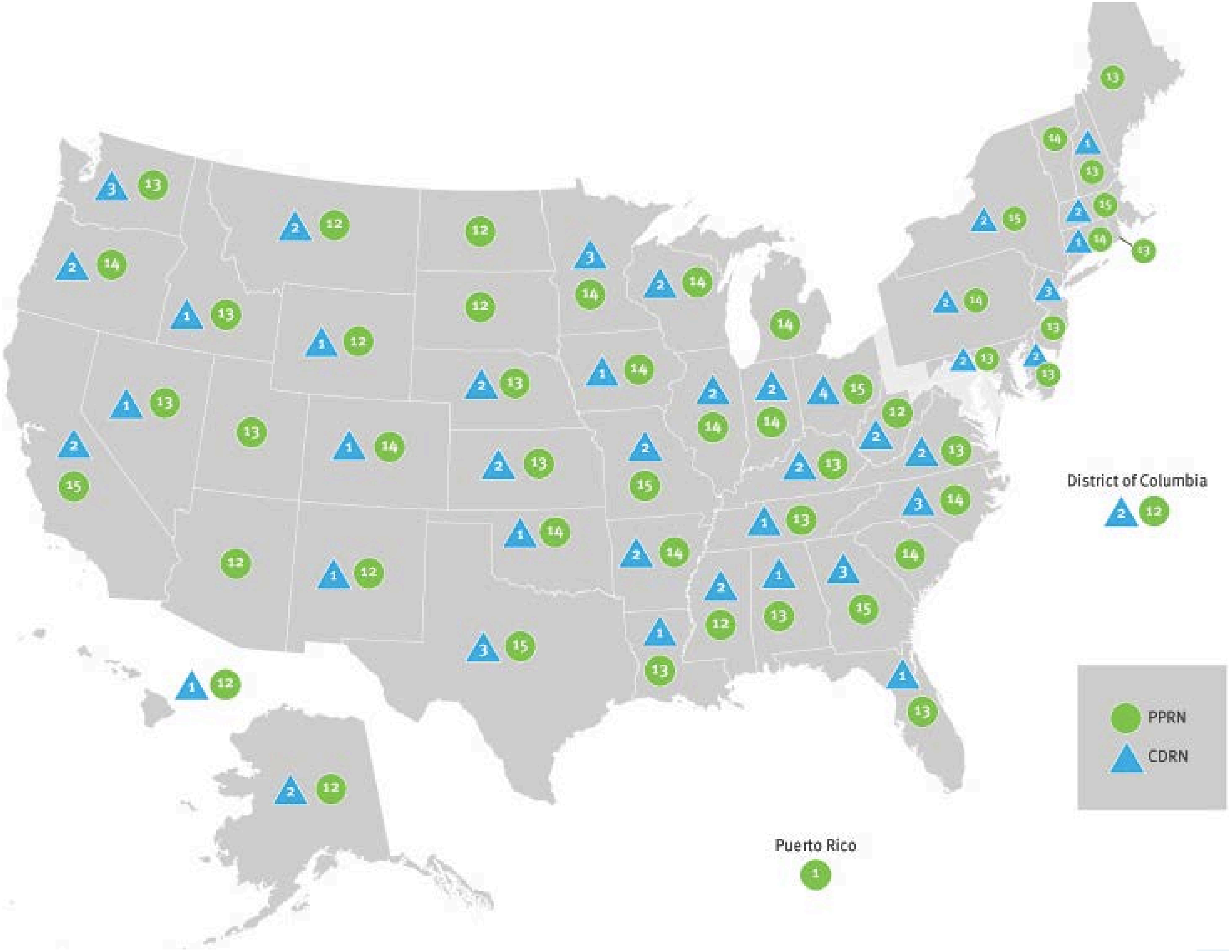
18 Patient-Powered Research Networks

(PPRNs)

PCORnet:

A national infrastructure for patient-centered clinical research

11 CDRN and 18 PPRN awards



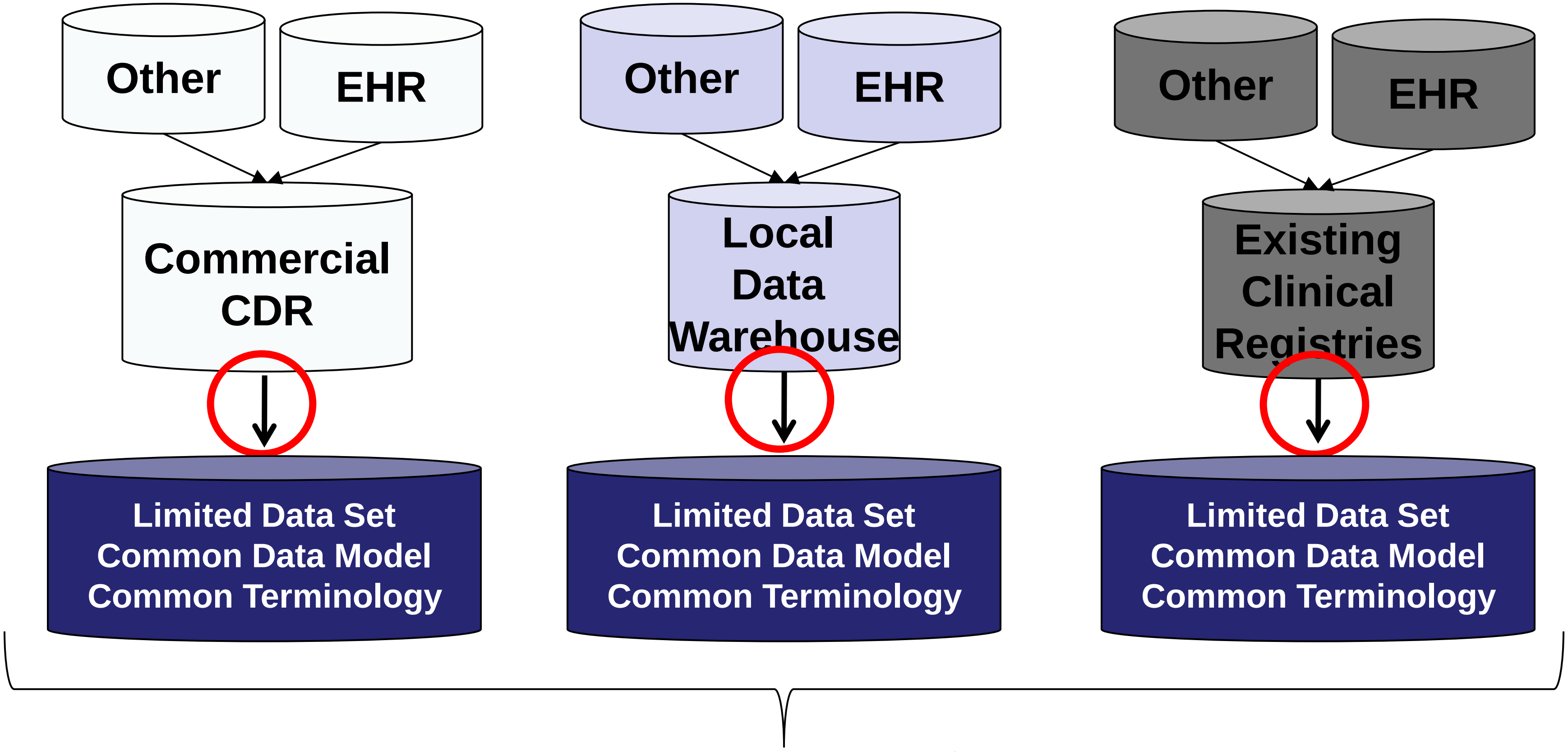
This map depicts the number of PCORI-funded Patient-Powered or Clinical Data Research Networks that have coverage in each state.

PEDSNet
CDRN

Guide to the Presentation

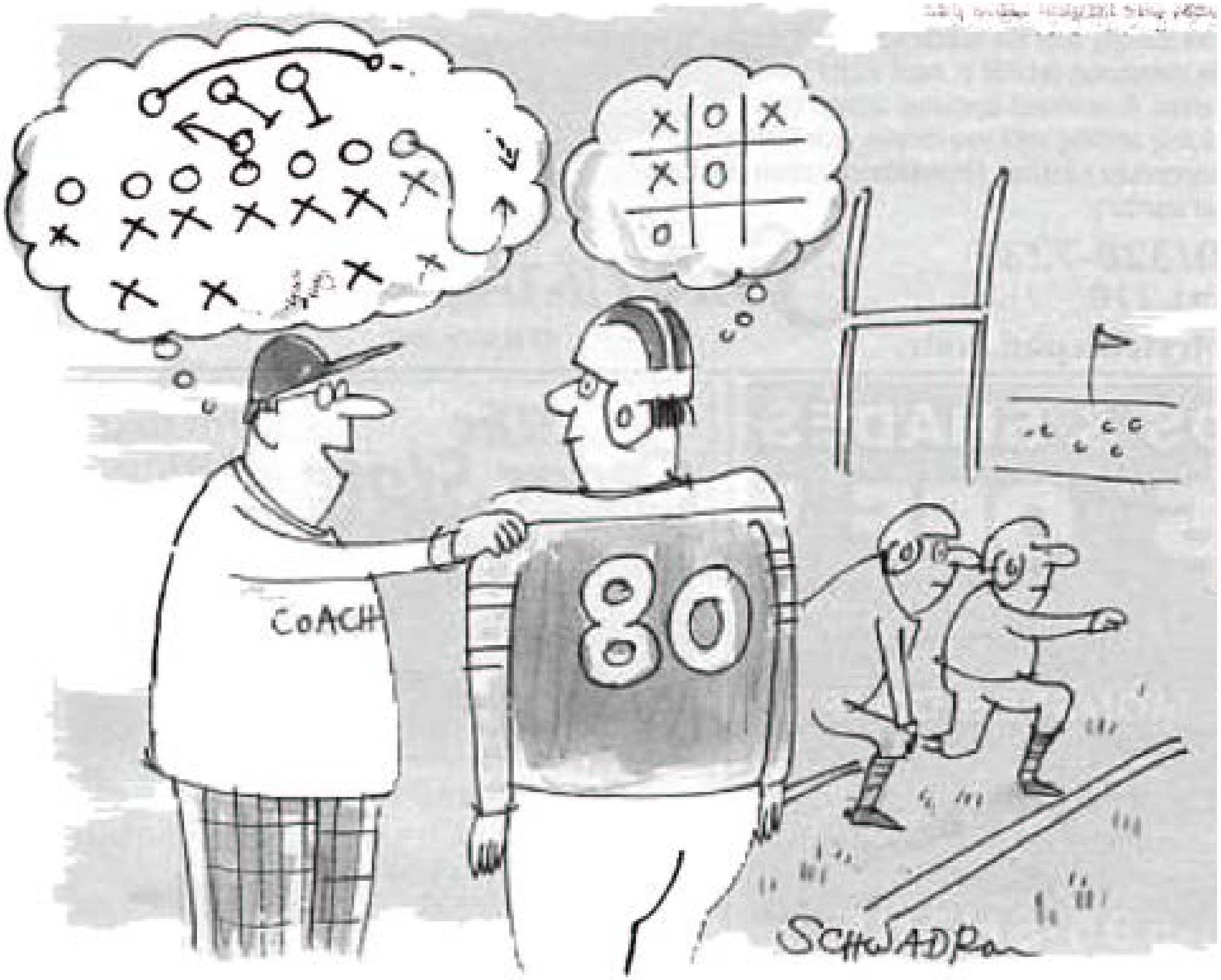
- The fun stuff
  - What is “clinical and translational” data management?
  - The changing landscape of clinical research
  - Learning health systems
  - National data networks
- The grunt work to do the fun stuff
  - Data harmonization
  - Data quality
  - My database can’t count

Harmonizing data into a common structure

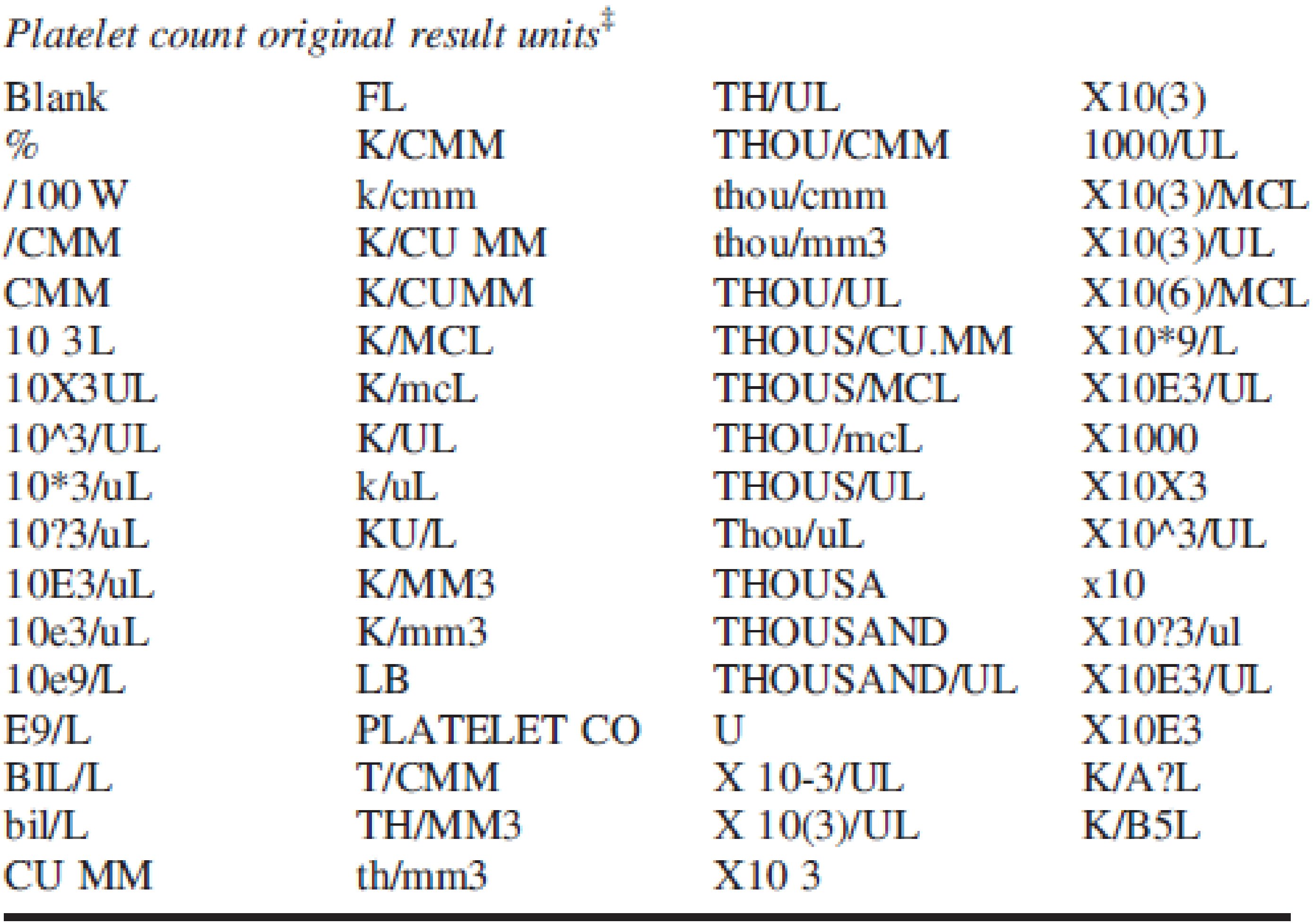


Common Query Interface

Terminology Harmonization – What are we talking about?



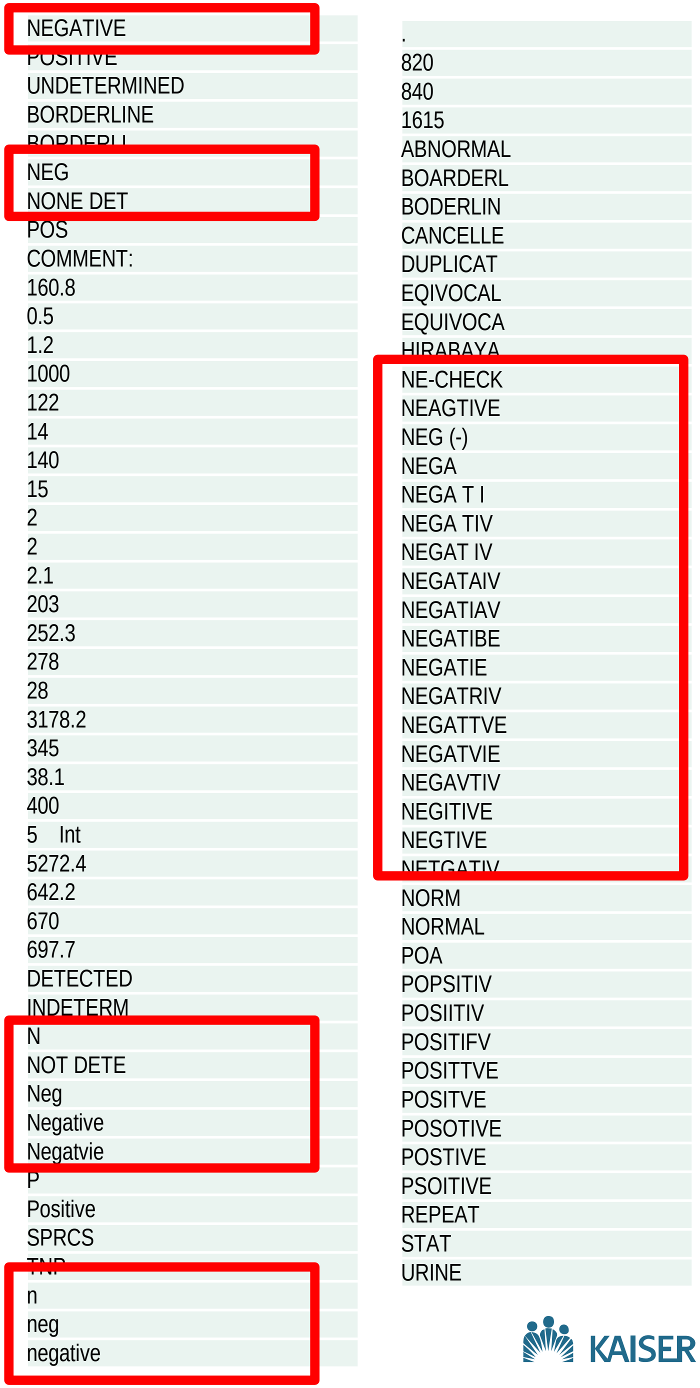
Examples of Variations in Platelet (Quantitative) Result Units in Source Data



Platelet count original result units[‡]

Blank
%
/100 W
/CMM
CMM
10 3L
10X3UL
10^3/UL
10*3/uL
10?3/uL
10E3/uL
10e3/uL
10e9/L
E9/L
BIL/L
bil/L
CU MM
FL
K/CMM
k/cmm
K/CU MM
K/CUMM
K/MCL
K/mcL
K/UL
k/uL
KU/L
K/MM3
K/mm3
LB
PLATELET CO
T/CMM
TH/MM3
th/mm3
TH/UL
THOU/CMM
thou/cmm
thou/mm3
THOU/UL
THOUS/CU.MM
THOUS/MCL
THOU/mcL
THOUS/UL
Thou/uL
THOUSA
THOUSAND
THOUSAND/UL
U
X 10-3/UL
X 10(3)/UL
X10 3
X10(3)
1000/UL
X10(3)/MCL
X10(3)/UL
X10(6)/MCL
X10*9/L
X10E3/UL
X1000
X10X3
X10^3/UL
x10
X10?3/ul
X10E3/UL
X10E3
K/A?L
K/B5L

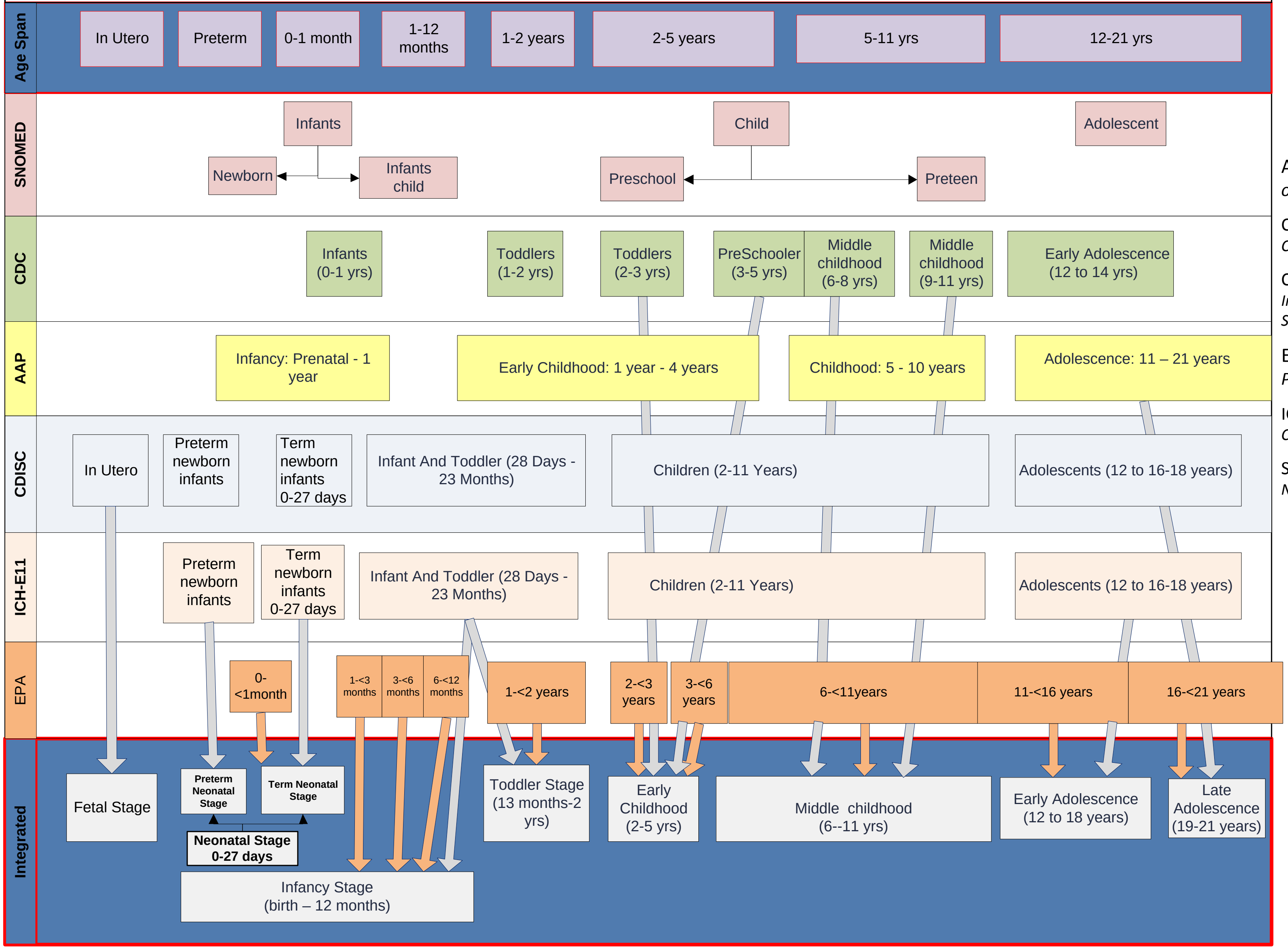
Examples of Variations in (Qualitative) Pregnancy Result Units in Source Data (aka, how many ways can you spell negative?)



NEGATIVE
POSITIVE
UNDETERMINED
BORDERLINE
BORDERLI
NEG
NONE DET
POS
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
N
NOT DETE
Neg
Negative
Negatvie
P
Positive
SPRCS
TNP
n
neg
negative

.
820
840
1615
ABNORMAL
BOARDERL
BODERLIN
CANCELLE
DUPLICAT
EQIVOCAL
EQUIVOCA
HIRABAYA
NE-CHECK
NEAGTIVE
NEG (-)
NEGA
NEGA T I
NEGA TIV
NEGAT IV
NEGATAIV
NEGATIAV
NEGATIBE
NEGATIE
NEGATRIV
NEGATTVE
NEGATVIE
NEGAVTIV
NEGITIVE
NEGTIVE
NETGATIV
NORM
NORMAL
POA
POPSITIV
POSIITIV
POSITIFV
POSITTVE
POSITVE
POSOTIVE
POSTIVE
PSOITIVE
REPEAT
STAT
URINE

Integrated child-life stages for NICHD Pediatric Terminology as mapped to existing medical terminologies



Age Span

In Utero

Preterm

0-1 month

1-12 months

1-2 years

2-5 years

5-11 yrs

12-21 yrs

SNOMED

Infants

Newborn

Infants child

Child

Preschool

Preteen

Adolescent

CDC

Infants (0-1 yrs)

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)

AAP

Infancy: Prenatal - 1 year

Early Childhood: 1 year - 4 years

Childhood: 5 - 10 years

Adolescence: 11 – 21 years

CDISC

In Utero

Preterm newborn infants

Term newborn infants 0-27 days

Infant And Toddler (28 Days - 23 Months)

Children (2-11 Years)

Adolescents (12 to 16-18 years)

ICH-E11

Preterm newborn infants

Term newborn infants 0-27 days

Infant And Toddler (28 Days - 23 Months)

Children (2-11 Years)

Adolescents (12 to 16-18 years)

EPA

0-<1month

1-<3 months

3-<6 months

6-<12 months

1-<2 years

2-<3 years

3-<6 years

6-<11years

11-<16 years

16-<21 years

Integrated

Fetal Stage

Preterm Neonatal Stage

Term Neonatal Stage

Neonatal Stage 0-27 days

Infancy Stage (birth – 12 months)

Toddler Stage (13 months-2 yrs)

Early Childhood (2-5 yrs)

Middle childhood (6--11 yrs)

Early Adolescence (12 to 18 years)

Late Adolescence (19-21 years)

AAP: *American Academy of Pediatrics*

CDC: *Centers for Disease Control and Prevention*

CDISC: *Clinical Data Interchange Standards Consortium*

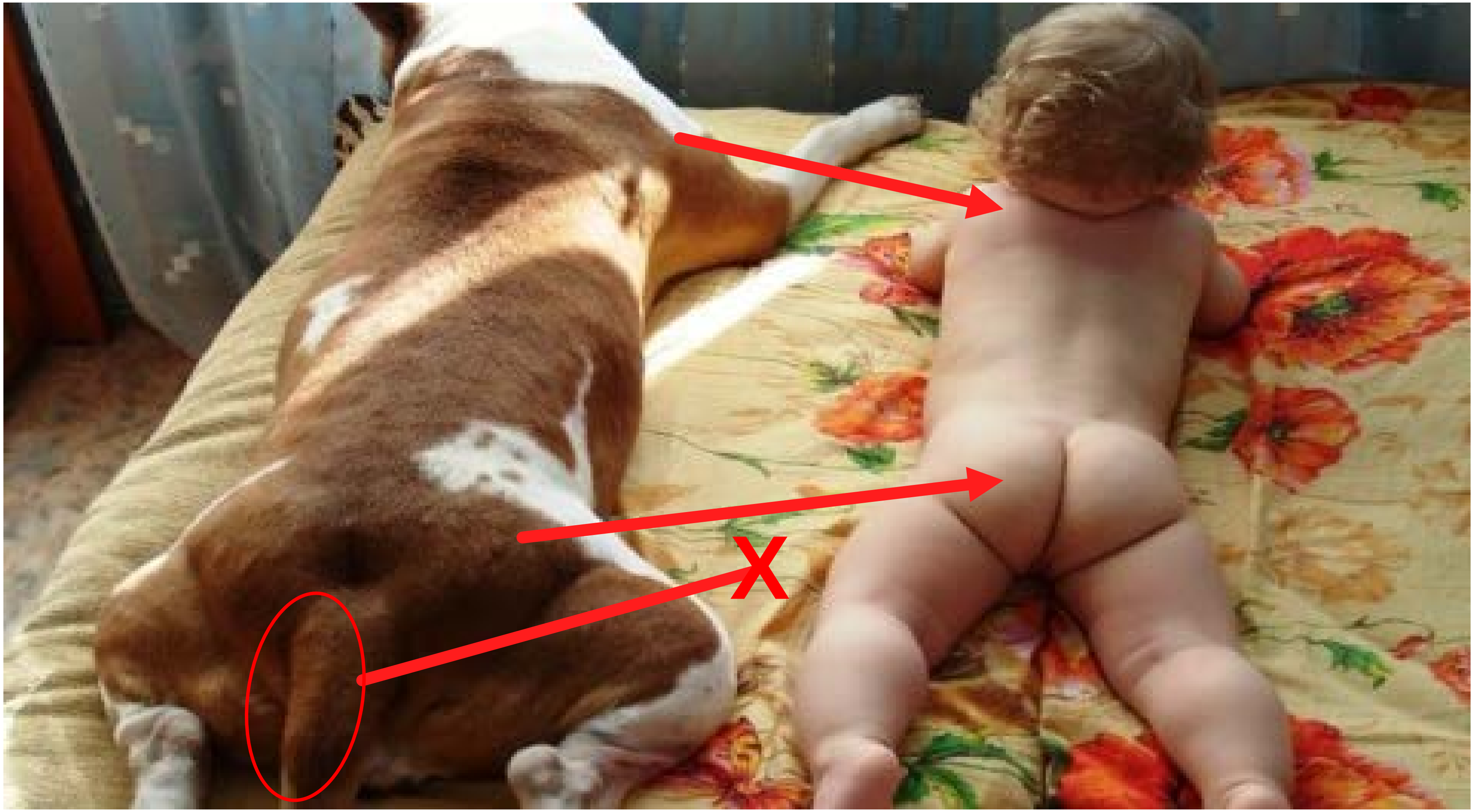
EPA: *Environmental Protection Agency*

ICH-E11: *International Conference on Harmonisation*

SNOMED: *Systematized Nomenclature of Medicine*

From: Steven Hirschfeld MD, PhD

Aligning Terminologies



SNOMED CT: A “mandated” clinical standard

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

Attention deficit hyperactivity disorder (disorder)

- loss of scalp hair
- scalp hair loss

Diffuse loss of scalp hair (finding)

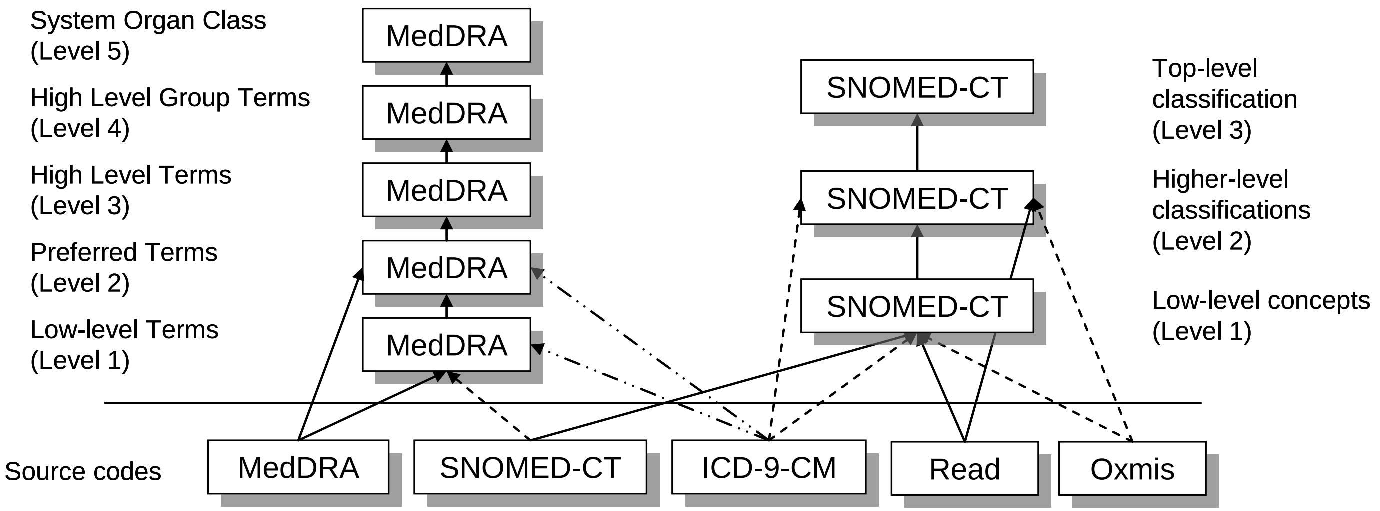
- Epistaxis
- nosebleeds

Epistaxis (disorder)

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)

MedDRA

MedDRA

MedDRA

MedDRA

MedDRA

SNOMED-CT

SNOMED-CT

SNOMED-CT

Top-level classification (Level 3)

Higher-level classifications (Level 2)

Low-level concepts (Level 1)

Source codes

MedDRA

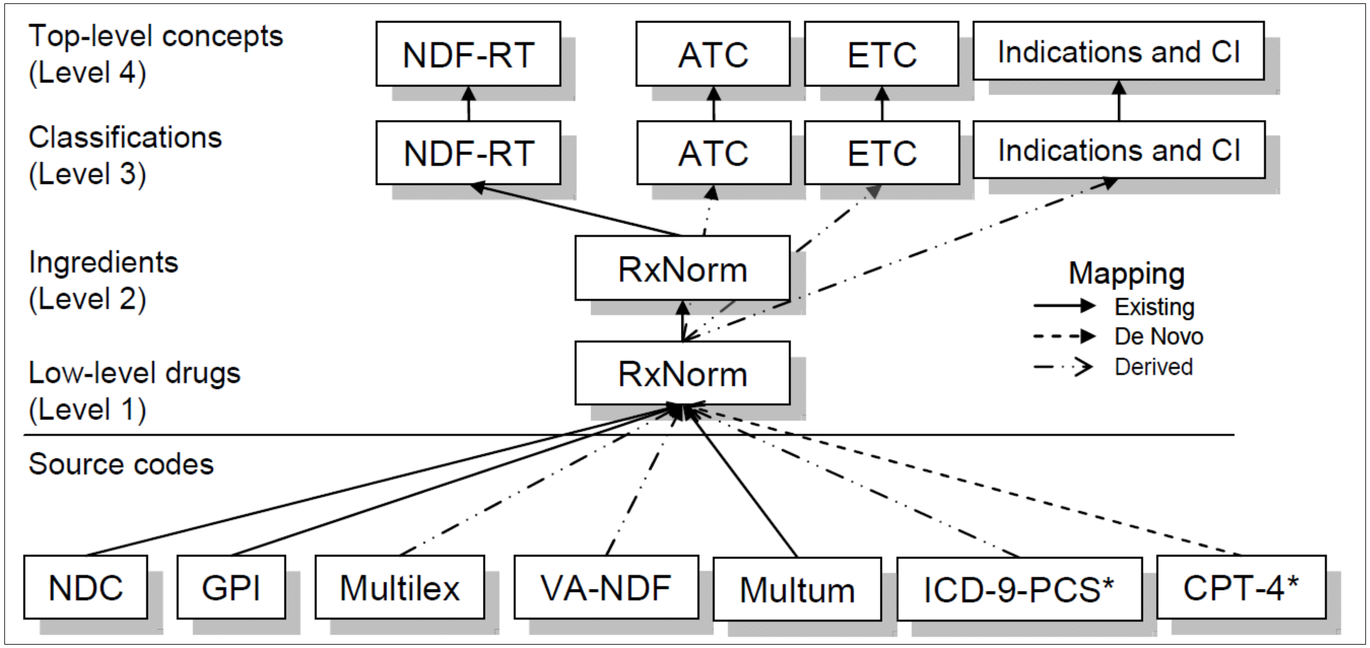
SNOMED-CT

ICD-9-CM

Read

Oxmis

Standardizing drugs:



Top-level concepts (Level 4)

Classifications (Level 3)

Ingredients (Level 2)

Low-level drugs (Level 1)

Source codes

NDF-RT

ATC

ETC

Indications and CI

NDF-RT

ATC

ETC

Indications and CI

RxNorm

RxNorm

Mapping

Existing

De Novo

Derived

NDC

GPI

Multilex

VA-NDF

Multum

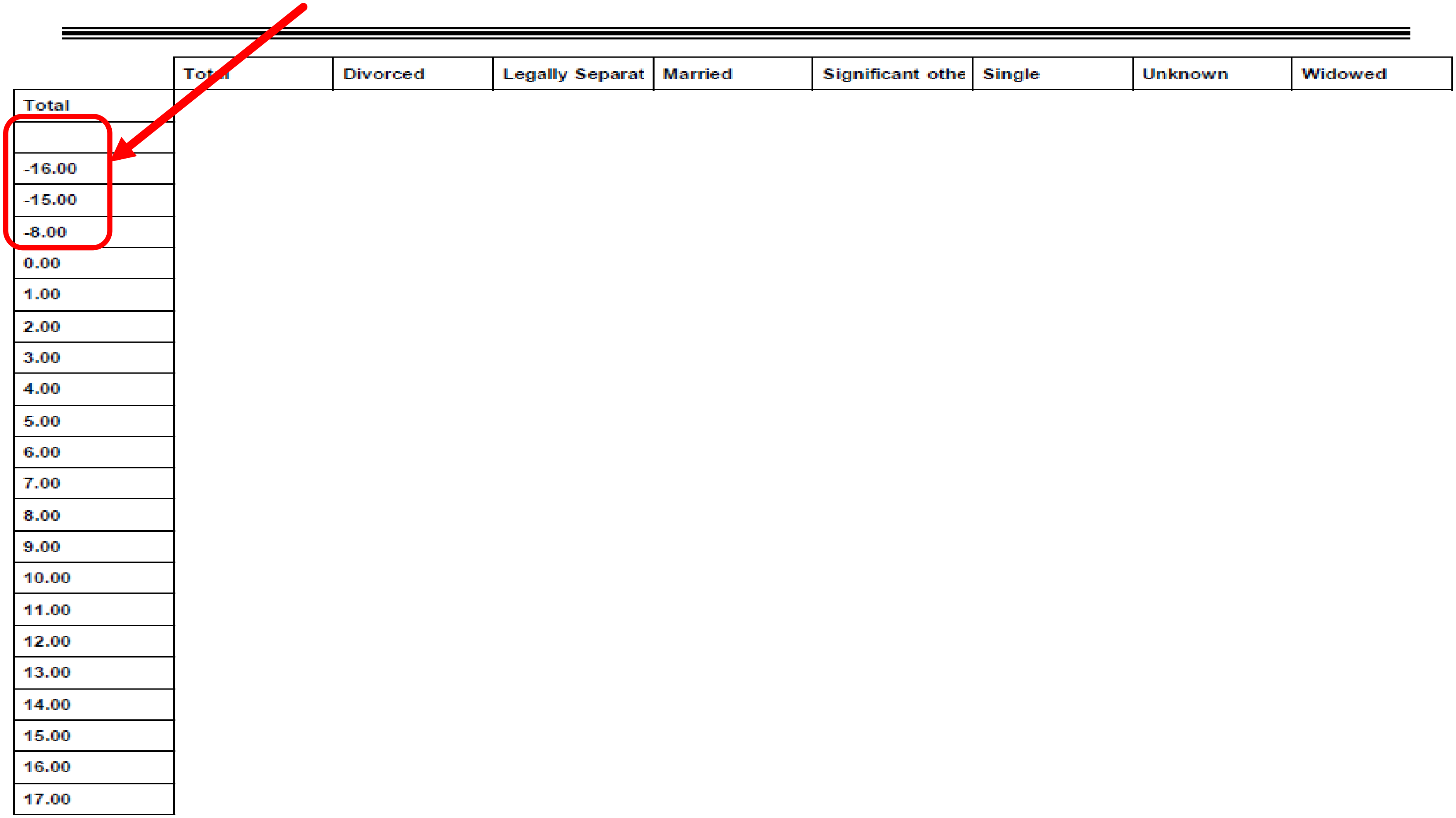
ICD-9-PCS*

CPT-4*

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



	Total	Divorced	Legally Separat	Married	Significant othe	Single	Unknown	Widowed
Total								
-16.00								
-15.00								
-8.00								
0.00								
1.00								
2.00								
3.00								
4.00								
5.00								
6.00								
7.00								
8.00								
9.00								
10.00								
11.00								
12.00								
13.00								
14.00								
15.00								
16.00								
17.00								

CCTSI

Colorado Clinical & Translational Sciences Institute

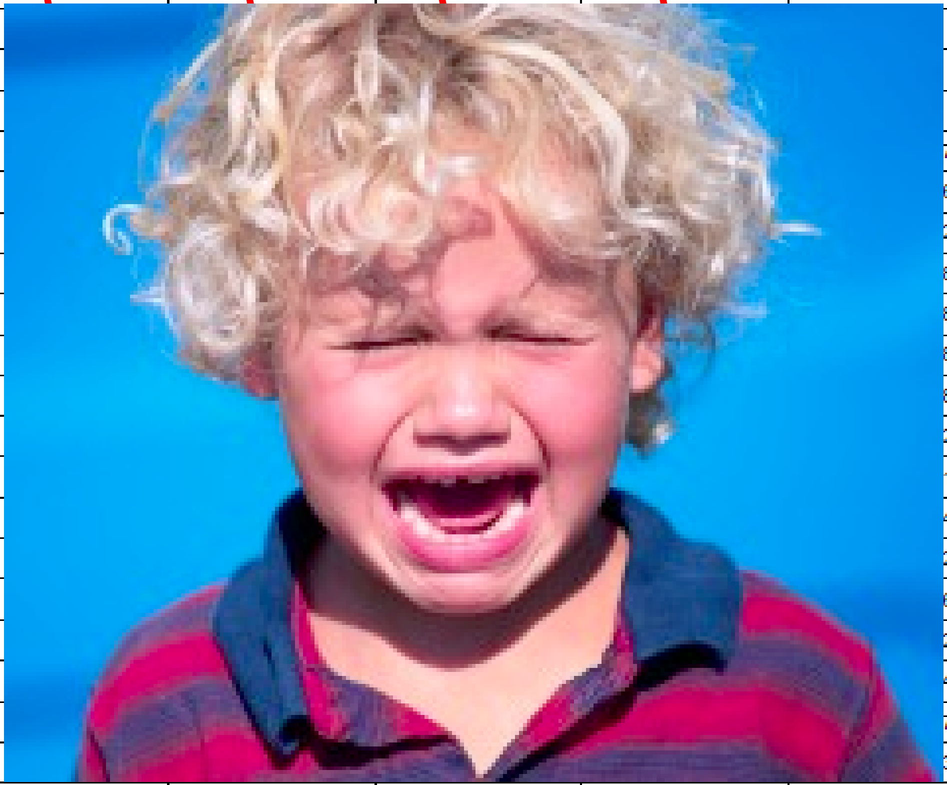
University of Colorado Denver

A trivial example: Martial Status by Age

It’s tough being 6 years old in Denver…….

Would these results be worrisome?

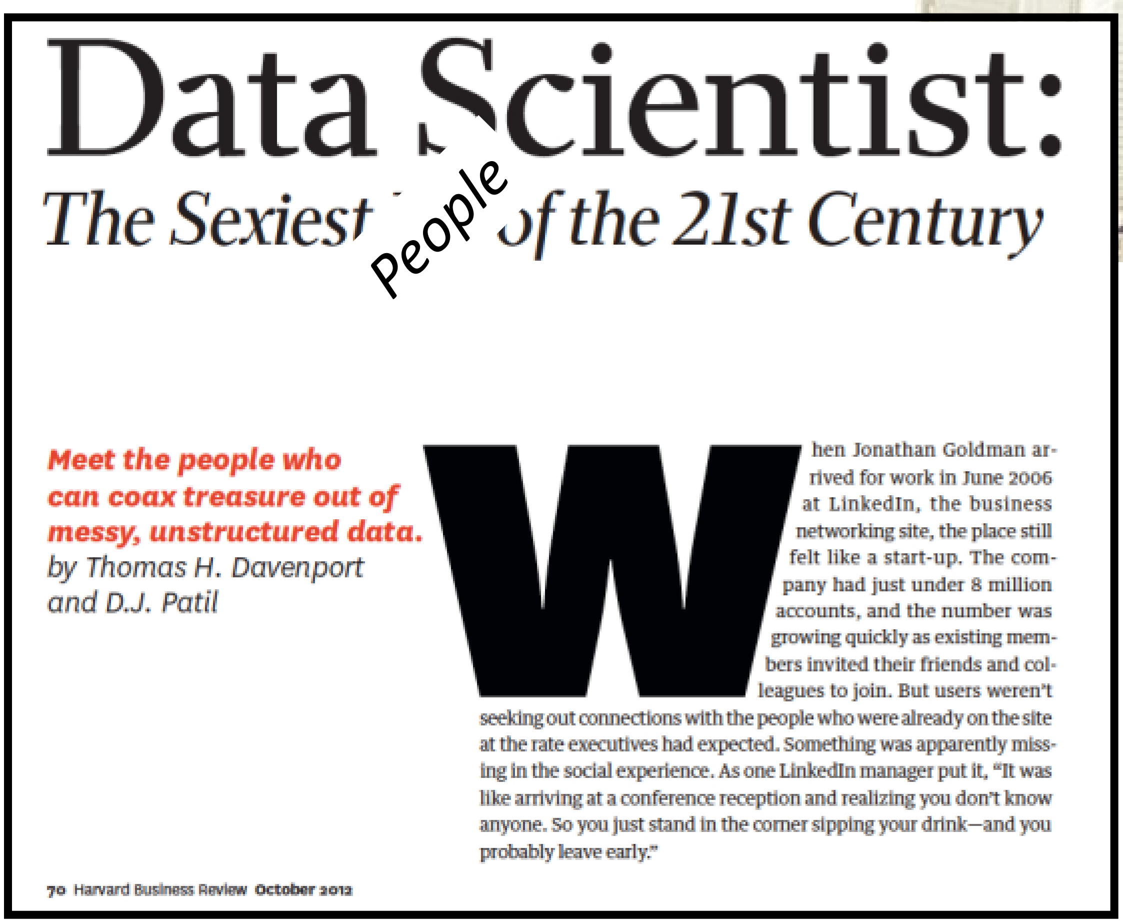
	Total	Divorced	Legally Separat	Married	Significant othe	Single	Unknown	Widowed
Total	423,508	33	3	1,606	81	420,944	830	11
	70	0	0	0	0	70	0	0
-16.00	2					2	0	0
-15.00	1					1	0	0
-8.00	1					1	0	0
0.00	768					768	0	0
1.00	13,660					652	5	0
2.00	21,350					290	25	0
3.00	24,960					885	31	0
4.00	27,861					806	32	0
5.00	29,933					889	24	0
6.00	30,932					810	40	0
7.00	27,381					268	46	0
8.00	24,198					124	31	0
9.00	22,522					448	35	0
10.00	20,283					231	22	0
11.00	18,705					659	16	0
12.00	17,340					296	19	0
13.00	16,510					470	17	0
14.00	15,792					761	15	0
15.00	15,354					302	21	0
16.00	15,474	2	0	19	1	15,439	13	0
17.00	15,208	1	0	9	0	15,181	17	0



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 People of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

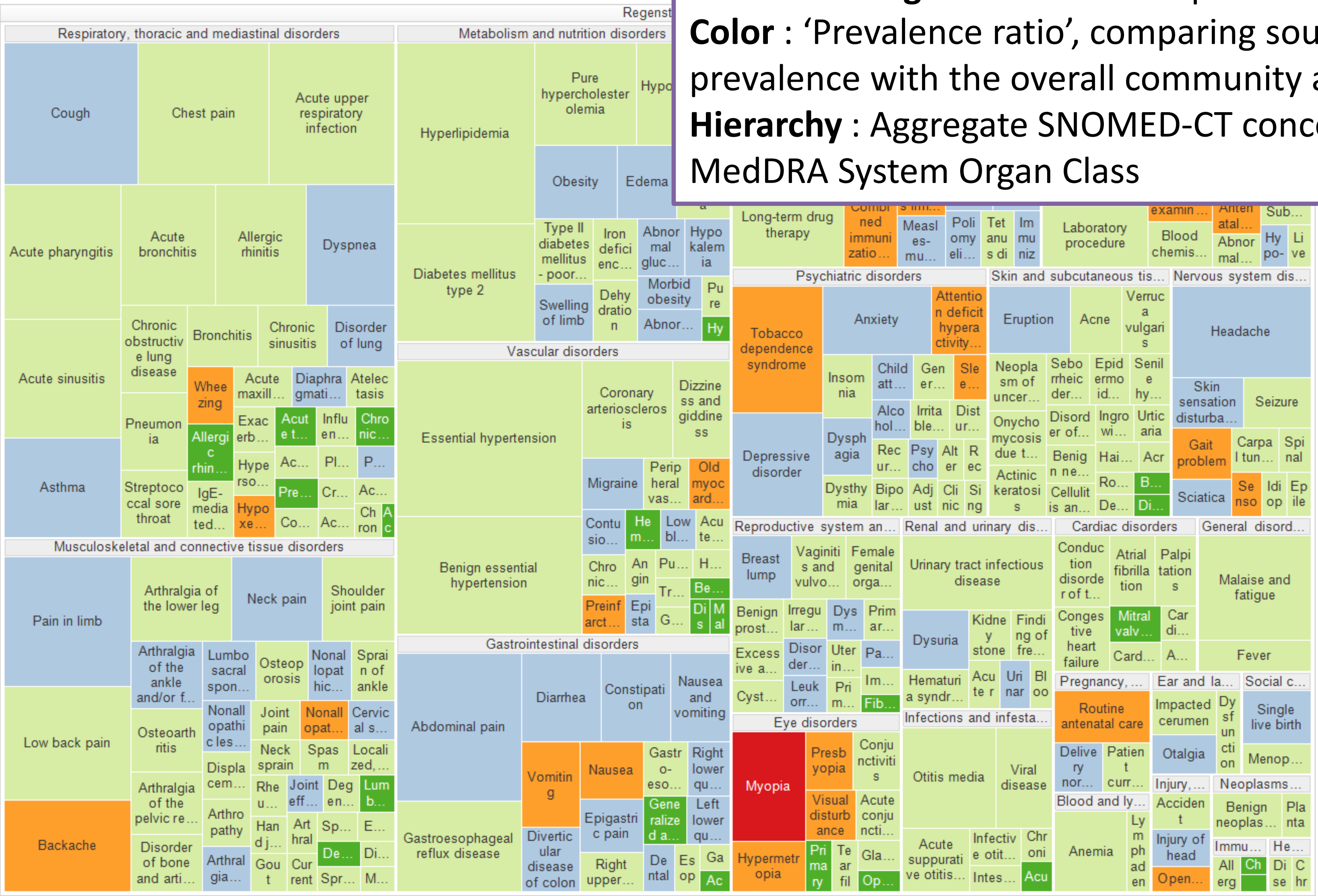
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren’t seeking 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.”

70 Harvard Business Review **October 2012**

OBSERVATIONAL
MEDICAL
OUTCOMES
PARTNERSHIP

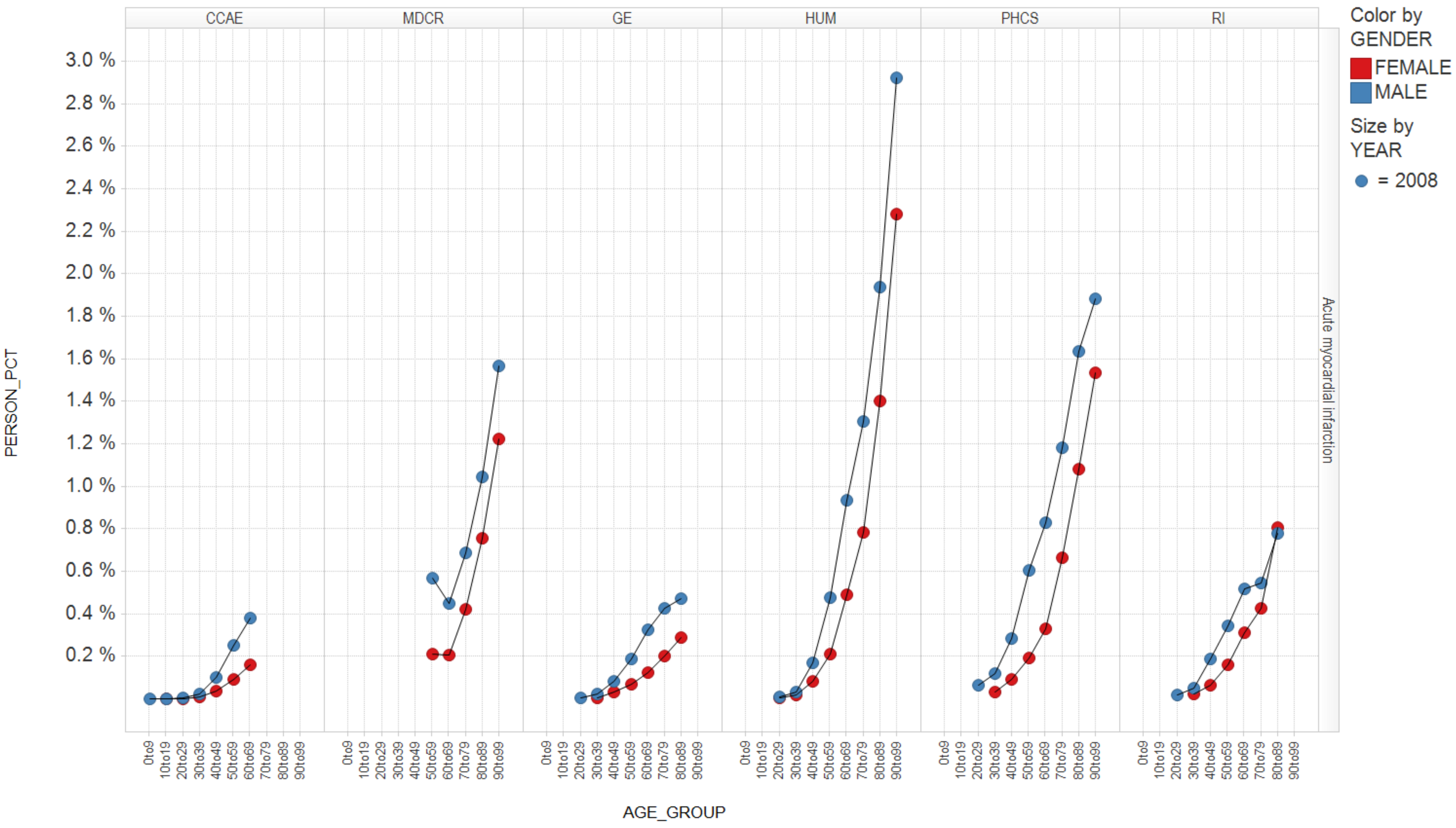
Exploring prevalence of all diseases

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



OBSERVATIONAL
MEDICAL
OUTCOMES
PARTNERSHIP

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”

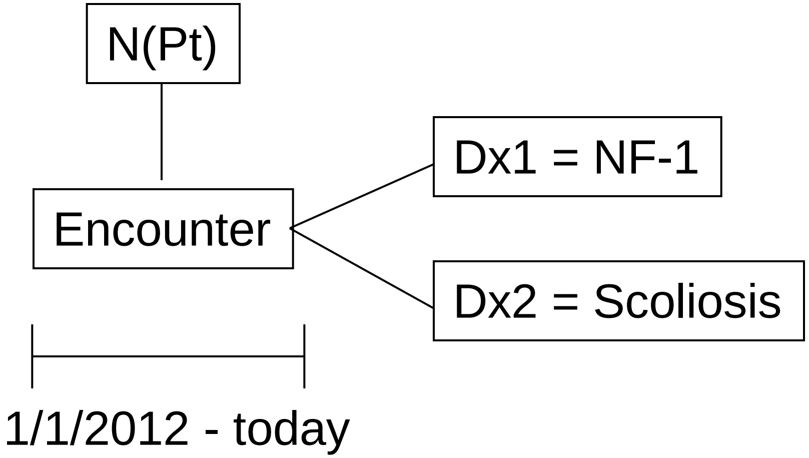
The Tale of A Simple Data Query

- First query result: N = 15

Clinical investigator did not believe this result even though we used her definitions.

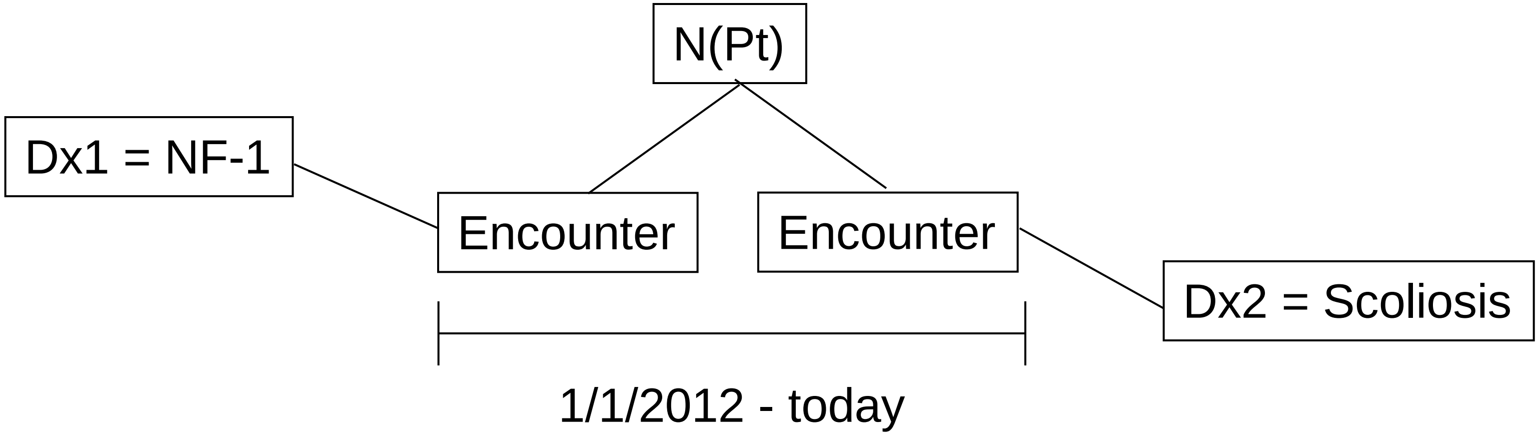
The Tale of A Simple Data Query

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



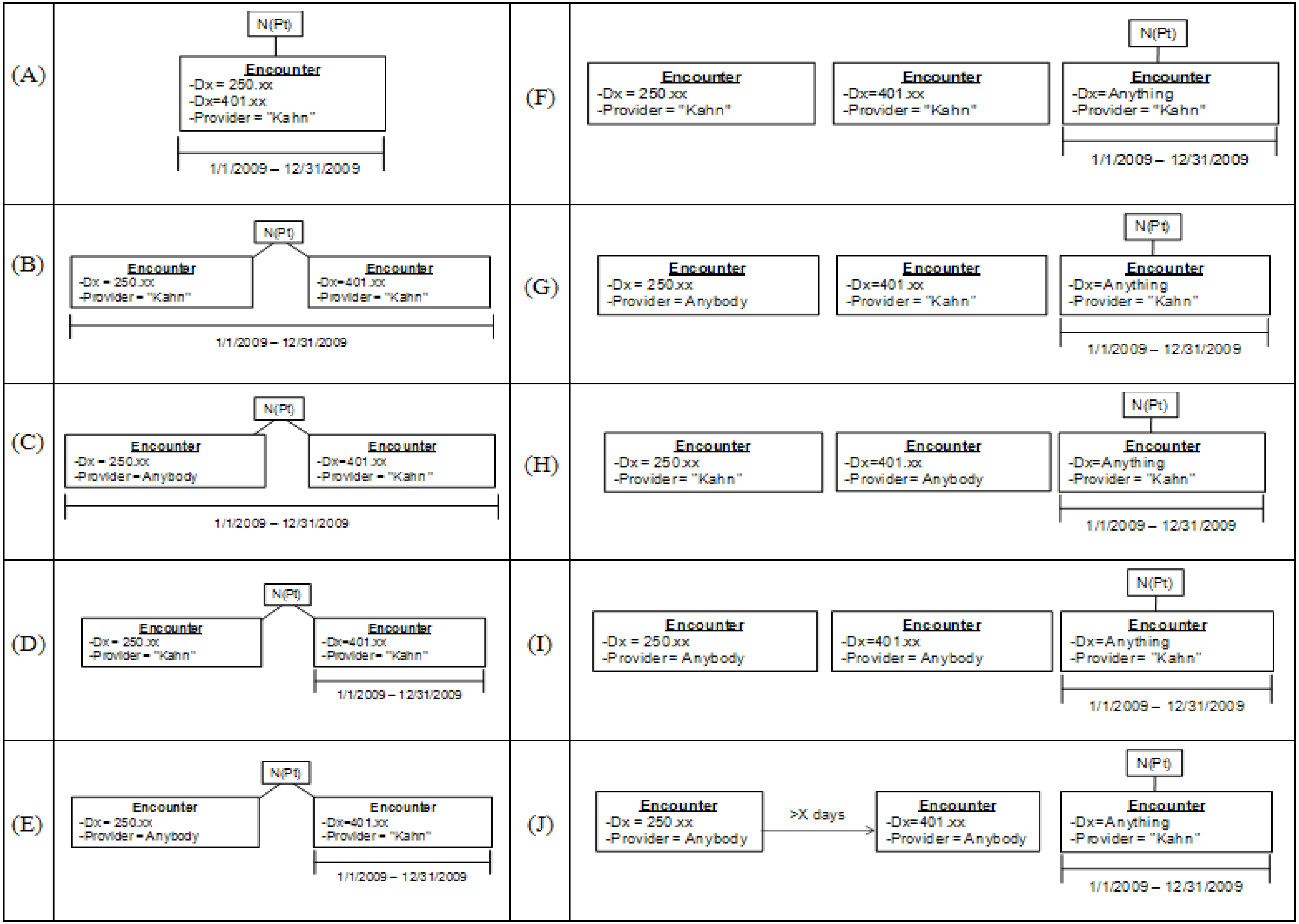
The Tale of A Simple Data Query

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



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.



(A)

N(Pt)

Encounter
-Dx = 250.xx
-Dx=401.xx
-Provider = "Kahn"

1/1/2009 – 12/31/2009

(B)

N(Pt)

Encounter
-Dx = 250.xx
-Provider = "Kahn"

Encounter
-Dx=401.xx
-Provider = "Kahn"

1/1/2009 – 12/31/2009

(C)

N(Pt)

Encounter
-Dx = 250.xx
-Provider = Anybody

Encounter
-Dx=401.xx
-Provider = "Kahn"

1/1/2009 – 12/31/2009

(D)

N(Pt)

Encounter
-Dx = 250.xx
-Provider = "Kahn"

Encounter
-Dx=401.xx
-Provider= "Kahn"

1/1/2009 – 12/31/2009

(E)

N(Pt)

Encounter
-Dx = 250.xx
-Provider = Anybody

Encounter
-Dx=401.xx
-Provider = "Kahn"

1/1/2009 – 12/31/2009

(F)

Encounter
-Dx = 250.xx
-Provider = "Kahn"

Encounter
-Dx=401.xx
-Provider = "Kahn"

N(Pt)

Encounter
-Dx= Anything
-Provider = "Kahn"

1/1/2009 – 12/31/2009

(G)

Encounter
-Dx = 250.xx
-Provider = Anybody

Encounter
-Dx=401.xx
-Provider = "Kahn"

N(Pt)

Encounter
-Dx=Anything
-Provider = "Kahn"

1/1/2009 – 12/31/2009

(H)

Encounter
-Dx = 250.xx
-Provider= "Kahn"

Encounter
-Dx=401.xx
-Provider= Anybody

N(Pt)

Encounter
-Dx=Anything
-Provider = "Kahn"

1/1/2009 – 12/31/2009

(I)

Encounter
-Dx = 250.xx
-Provider= Anybody

Encounter
-Dx=401.xx
-Provider = Anybody

N(Pt)

Encounter
-Dx=Anything
-Provider= "Kahn"

1/1/2009 – 12/31/2009

(J)

Encounter
-Dx = 250.xx
-Provider = Anybody

>X days

Encounter
-Dx= 401.xx
-Provider = Anybody

N(Pt)

Encounter
-Dx=Anything
-Provider = "Kahn"

1/1/2009 – 12/31/2009

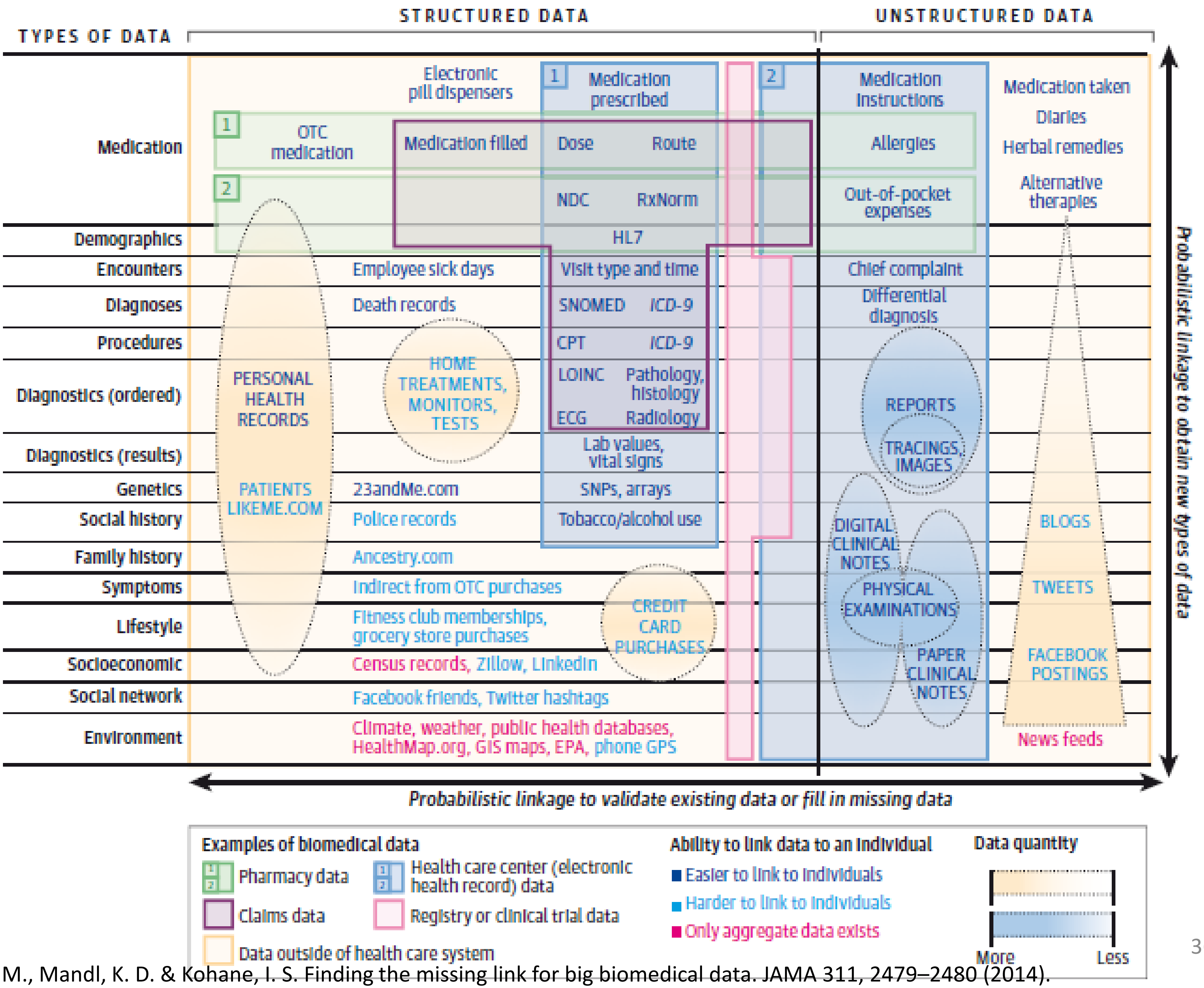
Guide to the Presentation

- The fun stuff
  - What is “clinical … management?
  - Th… rch
  - …
  - …
- The
  - D
  - Da
  - My database can’t count

It is a wonderful time in this field!

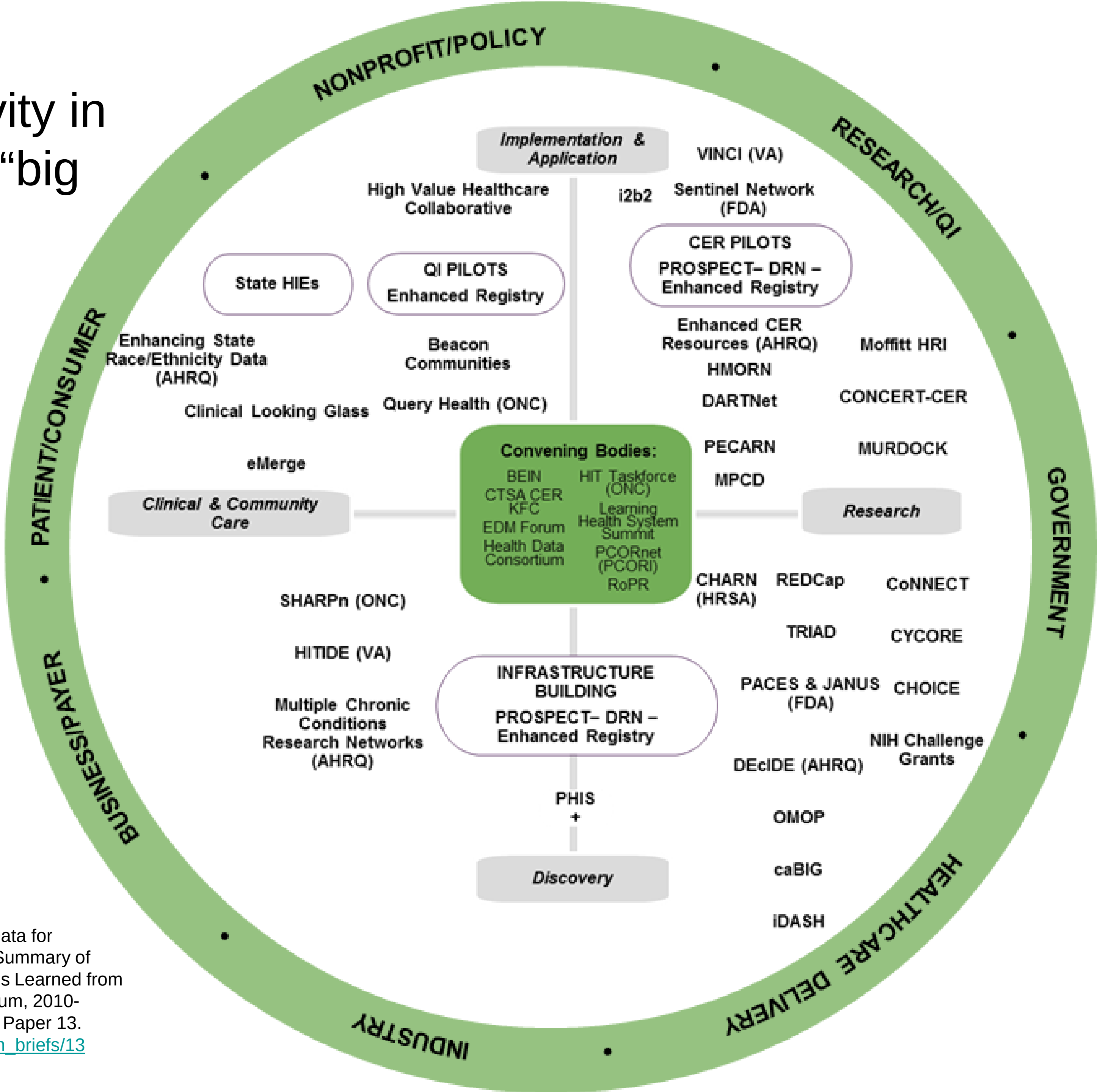
The fun stuff >>> The grunt work
(And even the grunt work ain’t bad!)

Figure. The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health Care



Weber, G. M., Mandl, K. D. & Kohane, I. S. Finding the missing link for big biomedical data. JAMA 311, 2479–2480 (2014).

Explosive activity in “big data” and “big data analytics” in healthcare



NONPROFIT/POLICY

RESEARCH/QI

GOVERNMENT

HEALTHCARE DELIVERY

INDUSTRY

BUSINESS/PAYER

PATIENT/CONSUMER

Implementation & Application

Clinical & Community Care

Research

Discovery

Convening Bodies:

BEIN
CTSA CER KFC
EDM Forum
Health Data Consortium

HIT Taskforce (ONC)
Learning Health System Summit
PCORnet (PCORI)
RoPR

State HIEs

QI PILOTS
Enhanced Registry

CER PILOTS
PROSPECT– DRN – Enhanced Registry

INFRASTRUCTURE BUILDING
PROSPECT– DRN – Enhanced Registry

High Value Healthcare Collaborative

VINCI (VA)

i2b2

Sentinel Network (FDA)

Enhancing State Race/Ethnicity Data (AHRQ)

Beacon Communities

Clinical Looking Glass

Query Health (ONC)

eMerge

Enhanced CER Resources (AHRQ)

Moffitt HRI

HMORN

DARTNet

CONCERT-CER

PECARN

MURDOCK

MPCD

CHARN (HRSA)

REDCap

CoNNECT

SHARPn (ONC)

TRIAD

CYCORE

HITIDE (VA)

PACES & JANUS (FDA)

CHOICE

Multiple Chronic Conditions Research Networks (AHRQ)

NIH Challenge Grants

DEcIDE (AHRQ)

PHIS +

OMOP

caBIG

iDASH

EDM Forum, "Using Electronic Health Data for Research and Quality Improvement: A Summary of Innovations, Achievements, and Lessons Learned from the Electronic Data Methods (EDM) Forum, 2010-2013" (2014). Issue Briefs and Reports. Paper 13. http://repository.academyhealth.org/edm_briefs/13

CCTSI Colorado Clinical & Translational Sciences Institute
University of Colorado Denver

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

Funding was provided by a contract from AcademyHealth. Additional support was provided by AHRQ 1R01HS019912-01 (Scalable PArtnering Network for CER: Across Lifespan, Conditions, and Settings), AHRQ 1R01HS019908 (Scalable Architecture for Federated Translational Inquiries Network), and NIH/NCRR Colorado CTSI Grant Number UL1 RR025780 (Colorado Clinical and Translational Sciences Institute).