Methods for quality assessment in secondary use of EHR data

Nicole Weiskopf, PhD, **Chunhua Weng, PhD, FACMI**
Department of Biomedical Informatics
Columbia University

For Medical Informatics Initiative Germany

May 3rd, 2018, 8:30am-9:00am NYC Time

Outline

1. How shall we define EHR data quality issues?

2. How shall we measure EHR data quality, such as completeness and bias?

3. How shall we convert DQ assessment methods into actionable knowledge?

4. How can make DQ assessment more systematic?

EHR data reuse has many advantages

- Decreased cost and increased efficiency
  - Recruitment
  - Retention
  - Data collection
- Large volume of data
- Representative patient population
  - Patient-centered outcomes

EHR data are subject to quality problems

"With the advent of the information era in medicine, we are pouring out a torrent of medical record misinformation. Medical records, which have long been faulty, contain more distorted, deleted, and misleading information than ever before."

Burnum (1989) The misinformation era: the fall of the medical record.

EHR data are subject to quality problems

"…the quality and comprehensiveness of the clinical data were not up to research standards or the analytical methods used to overcome these limitations were inadequate…."

Weiner & Embi (2009) Toward reuse of clinical data for research and quality improvement: The end of the beginning?

What is data quality?

"Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. Data are fit for use if they are free of defects and possess desired features."

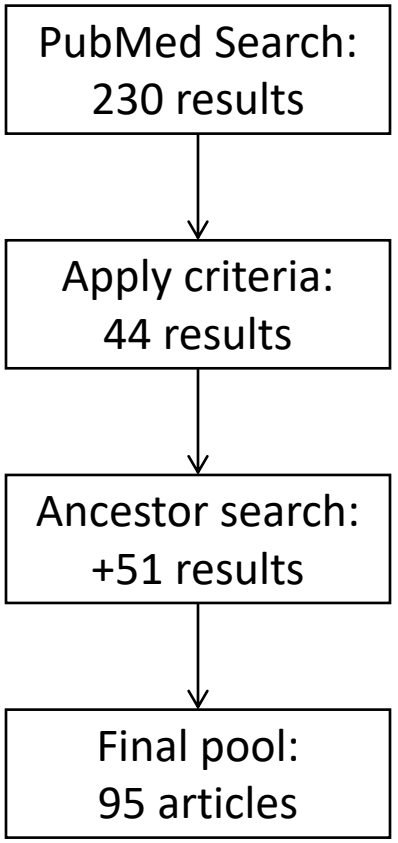
Redman, T (2001) Data quality: the field guide. Based on Juran's work.

Objective #1: to identify the most common dimensions of EHR data quality in the literature and map them these dimensions to methods of data quality assessment.

Literature review of EHR data quality assessment methods

Data collection:

- Literature review
- Inclusion criteria:
  - Original research using DQA methods
  - Data derived from EHR
  - Peer-reviewed
- Search: DQ terms & EHR terms
- Reviewed 230 articles
- Performed ancestor search
- Final pool: 95 relevant articles



Weiskopf & Weng (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.

Dimensions of Data Quality Derived from Literature

completeness	correctness	concordance	plausibility	currency
accessibility	accuracy	agreement	accuracy	recency
accuracy	corrections made	consistency	believability	timeliness
availability	misleading	reliability	trustworthiness	
missingness	PPV	variation	validity	
presence	quality			
quality	validity			
rate of recording				
sensitivity				
validity				

Dimensions of Data Quality Derived from Literature

completeness	correctness	concordance	plausibility	currency
accessibility	accuracy	agreement	accuracy	recency
accuracy	corrections made	consistency	believability	timeliness
availability	misleading	reliability	trustworthiness	
missingness	PPV	variation	validity	
presence	quality			
quality	validity			
rate of recording				
sensitivity				
validity				

Is a truth about a patient present in the EHR?

Dimensions of Data Quality Derived from Literature

completeness	correctness	concordance	plausibility	currency
accessibility	accuracy	agreement	accuracy	recency
accuracy	corrections made	consistency	believability	timeliness
availability	misleading	reliability	trustworthiness	
missingness	PPV	variation	validity	
presence	quality			
quality	validity			
rate of recording				
sensitivity				
validity				

Is an element that is present in the EHR true?

Dimensions of Data Quality Derived from Literature

completeness	correctness	concordance	plausibility	currency
accessibility	accuracy	agreement	accuracy	recency
accuracy	corrections made	consistency	believability	timeliness
availability	misleading	reliability	trustworthiness	
missingness	PPV	variation	validity	
presence	quality			
quality	validity			
rate of recording				
sensitivity				
validity				

Is there agreement between elements in the EHR, or between the EHR and another data source?

Dimensions of Data Quality Derived from Literature

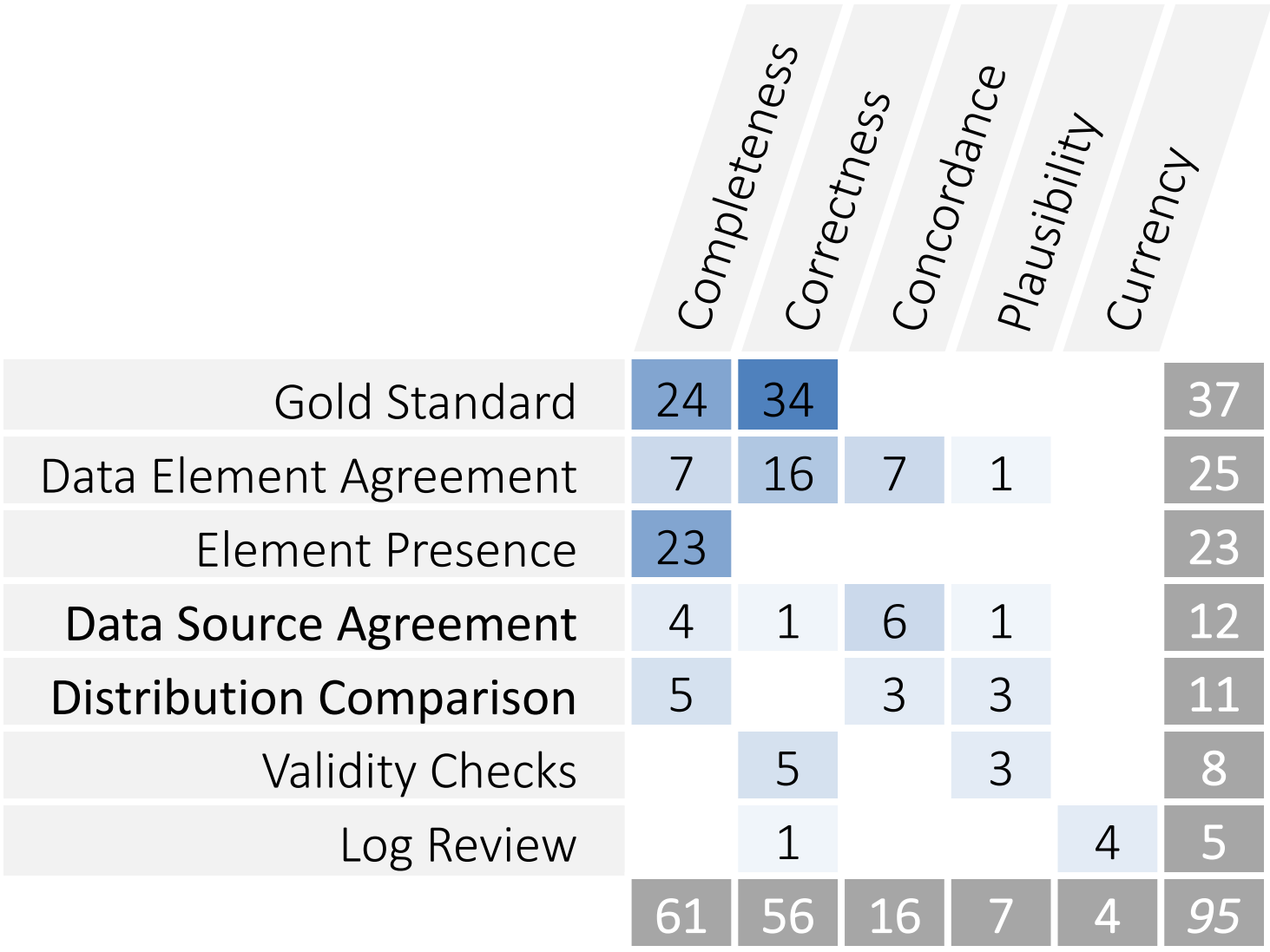
completeness	correctness	concordance	plausibility	currency
accessibility	accuracy	agreement	accuracy	recency
accuracy	corrections made	consistency	believability	timeliness
availability	misleading	reliability	trustworthiness	
missingness	PPV	variation	validity	
presence	quality			
quality	validity			
rate of recording				
sensitivity				
validity				

Does an element in the EHR makes sense in light of other knowledge about what that element is measuring?

Dimensions of Data Quality Derived from Literature

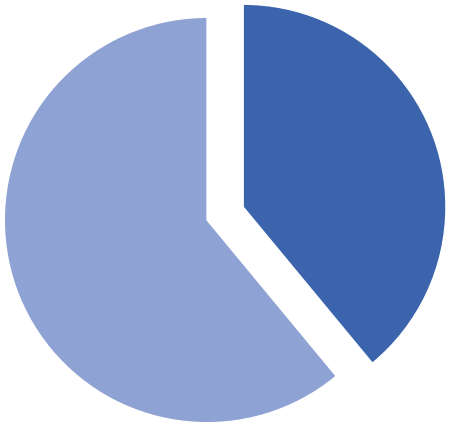
completeness	correctness	concordance	plausibility	currency
accessibility	accuracy	agreement	accuracy	recency
accuracy	corrections made	consistency	believability	timeliness
availability	misleading	reliability	trustworthiness	
missingness	PPV	variation	validity	
presence	quality			
quality	validity			
rate of recording				
sensitivity				
validity				

Is an element in the EHR a relevant representation of the patient state at a given point in time?


Completeness
Correctness
Concordance
Plausibility
Currency
Gold Standard
24
34
37
Data Element Agreement
7
16
7
1
25
Element Presence
23
23
Data Source Agreement
4
1
6
1
12
Distribution Comparison
5
3
3
11
Validity Checks
5
3
8
Log Review
1
4
5
61
56
16
7
4
95

Gold Standard

- A dataset drawn from another source or multiple sources, with or without information from the EHR, is used as a gold standard.
- Used for: correctness and completeness
- 39% of articles used a gold standard



Gold Standard

- Paper records (Ayoub 2007, Barrie 1992, Dambro & Weiss 1988, Hohnloser 1994, Mikkelsen 2005, Nazareth 1993, Pearson 1996, Ricketts 1993, Roukema 2006, Wallace 2002)
- Triangulated data (Aronsky & Haug 2000, Margulis 2009, van Weel 1995, Wagner & Hogan 1996)
- Information from patients (Dawson 1997, Kaboli 2004, Olala 2011, Ndira 2008, Porter 1999, Powell 2006, Pyper 2004, Staroselsky 2006, Staroselsky 2008, van Weel 1995, Weingart 2007, Whitelaw 1996)
- Information from physicians (Lewis 2004, Lo Re 2009, van Staa 2000)
- Patient encounters (Bentsen 1976, Logan 2001, Meara 1999, Smith 2005)
- Trained standard patients (Berner 2005, Peabody 2004)
- Alternate data sources (Madsen 1998, Staes 2006)
- Automatically recorded data (Vawdrey 2007)

Data Element Agreement

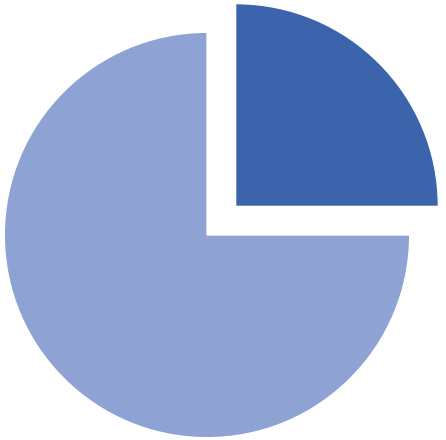
- Two or more elements within an EHR are compared to see if they report the same or compatible information.
- Used to assess completeness, correctness, concordance, and plausibility
- 26% of articles used agreement between data elements

Data Element Agreement

- Compared structured and unstructured data (Botsis 2010, Goulet 2007, Hogan & Wagner 1996, Hohnloser 1996)
- Compared related EHR elements (Basden 1980, Benson 2001, de Lusignan 2010, Jelovsek 1978, Horsfield 2002, Owen 2004)
  - Specifically diagnoses and related elements (de Burgos-Lunar, de Lusignan 2004, de Lusignan 2005, de Lusignan 2010, Faulconer 2004, Hassey 2001, Linder 2009, Pringle 1995, Stein 2000, Tang 2007, Terry 2009, Thiru 1999)
  - Data Quality Probes (Brown & Warmington 2002 & 2003)
- Identified copy-and-pasted data (Hammond 2003, Weir 2003)

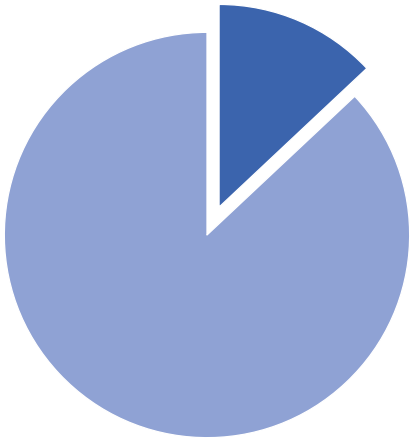
Element Presence

- A determination is made as to whether or not desired or expected data elements are present.
  - Desired elements (Agnew-Blais 2009, Asche 2008, Botsis 2010, de Lusignan 2004, Jensen 2009, Linder 2009, Lo Re 2009, Williams 2003)
  - Expected elements (Agnew-Blais 2009, Einbinder 1995, Forster 2008, Goulet 2007, Hahn 2011, Jelovsek 1978, Jones 1986, Ndira 2008, Olola 2011, Pearson 1996, Porcheret 2004, Pringle 1995, Scobie 1995, Soto 2002, Tang 1999, Thiru 1999)
- Used to assess completeness
- 24% of articles used element presence



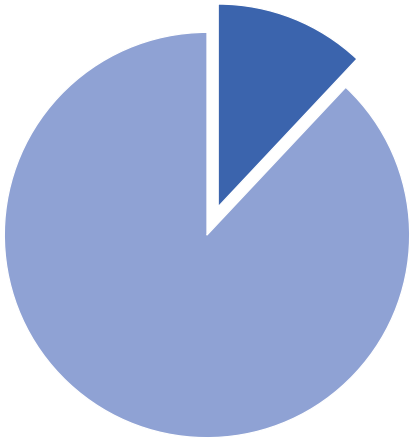
Data Source Agreement

- Elements from the EHR are compared to data from other sources to determine if they are in agreement
  - Billing data (Roos 1989)
  - Shared data warehouse (Noel 2010)
  - Paper records (Jick 1991, Jick 1992, Mikkelson 2001, Neal 1996, Scobie 1995, Stausberg 2003)
  - Order system (Scobie 1995)
  - Survey data (Conroy 2005)
- Used to assess concordance
- 13% of articles used agreement between data sources



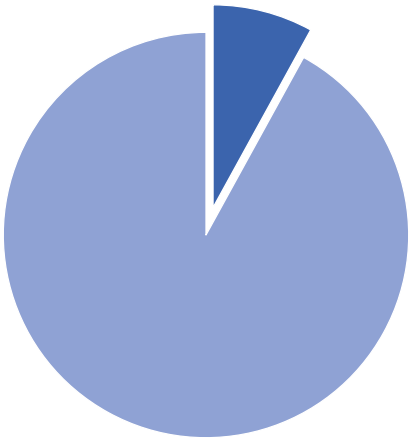
Distribution Comparison

- Distributions or summary statistics of aggregated data from the EHR are compared to the expected distributions for the clinical concepts of interest.
  - Between practices (de Lusignan 2003, de Lusignan 2005, Haynes 2011, Pringle 1995)
  - With national rates (Haynes 2009, Iyen-Omofoman 2011, Johnson 1991, Kaye 2000, Lewis 2004)
- Used to assess completeness, correctness, concordance, and plausibility
- 12% of articles used comparisons of data distributions



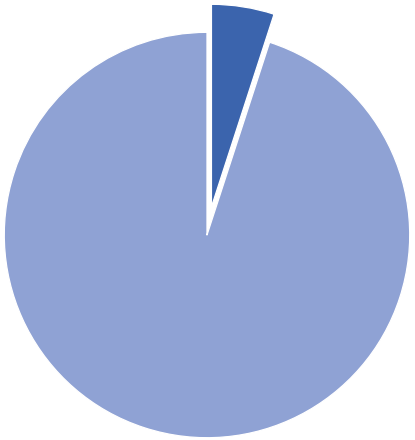
Validity Checks

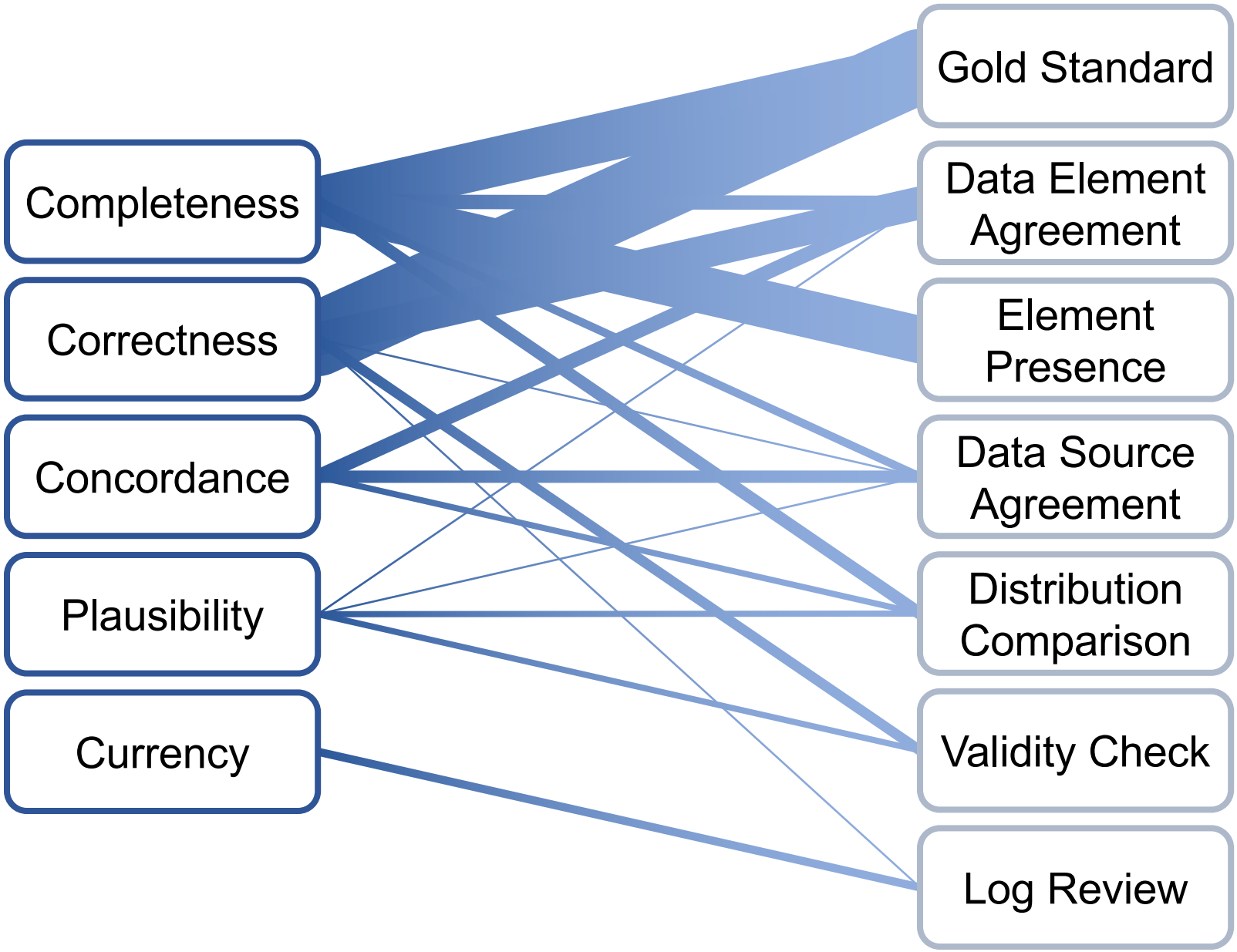
- Data in the EHR are assessed using various techniques that determine if values “make sense.”
  - Changes in sequential data (Haerian 2009, Noel 2010)
  - End-digit preference (Alsanjari 2011, de Lusignan 2004)
  - Range checks (Basden 1980, Noel 2010, Staes 2006)
  - Zero value checks (Benson 2001)
  - Diagnoses appropriateness (Pearson 1996)
- Used to assess plausibility and correctness
- 8% of articles used validity checks



Log Review

- Information on data entry practices (e.g. dates, times, edits) is examined.
- Used to assess
  - Currency (Falconer 2004, Ndira 2008, Williams 2003, Vawdrey 2007)
  - Correctness (Benson 2001)
- 5% of articles used log review




Completeness
Correctness
Concordance
Plausibility
Currency
Gold Standard
Data Element Agreement
Element Presence
Data Source Agreement
Distribution Comparison
Validity Check
Log Review

Discussion: Terminology and Dimensions

- Inconsistent terminology
  - Overlap
- Agreement with previous frameworks of DQ
  - Wang & Strong (1996): match 4 of 15 dimensions
  - IOM (1997): match 2 of 3 relevant dimensions
- Fundamental vs. “proxy” dimensions
  - Concordance and plausibility
  - May be useful for when it’s not possible to directly assess fundamental dimensionss

Discussion: DQ Assessment Methods

- Reliance on gold standards
  - GS often not available; de-identified databases
  - Not truly “gold”
- Most assessments relied upon ad hoc methods
  - Limited discussion of how to generalize methods
- Incorporation of “fitness for use”
  - Study and data needs should be made explicit prior to assessing DQ

Objective 2

Defining and measuring completeness of electronic health records for secondary use

Nicole G Weiskopf, George M Hripcsak, Alex Rusanov, Sushmita Swaminathan, Chunhua Weng

Completeness of EHR data is variable

- Hogan and Wagner (1997); 20 articles
  - Completeness: 1.1%-100%
- Thiru et al. (2003); 52 articles:
  - Sensitivity: 0.26 – 1.00
- Chan et al. (2010); 35 articles:
  - Completeness of BP: 0.1% – 51%

Why is there so much inconsistency?

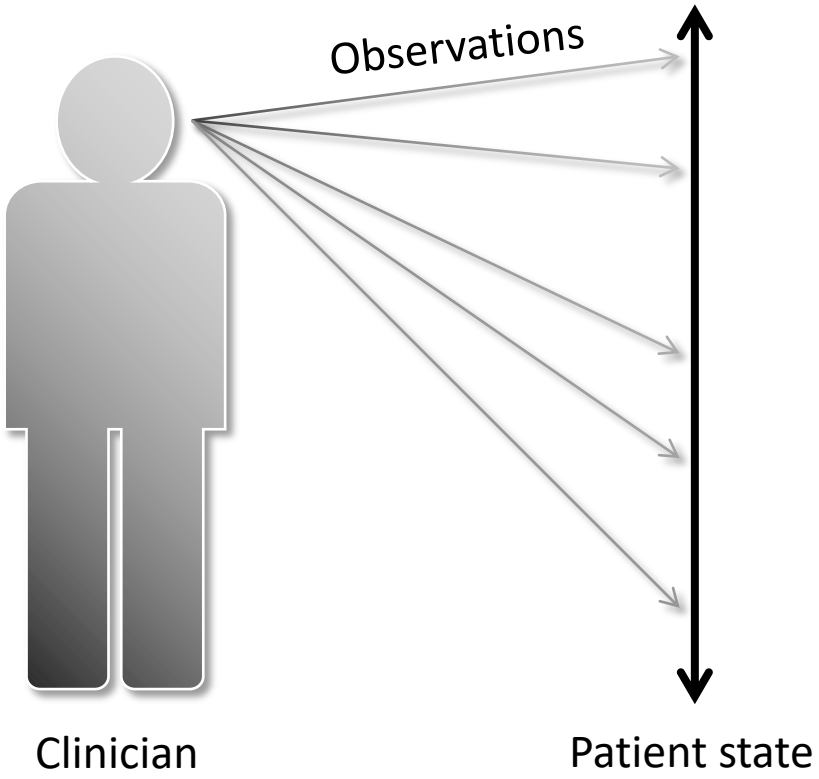
Hogan & Wagner (1997) Accuracy of data in computer-based patient records.
Thiru et al. (2003) Systematic review of scope and quality of electronic patient record data in primary care.
Chan et al. (2010) Electronic health records and the reliability and validity of quality measures: a review of the literature.

Hypothesis:

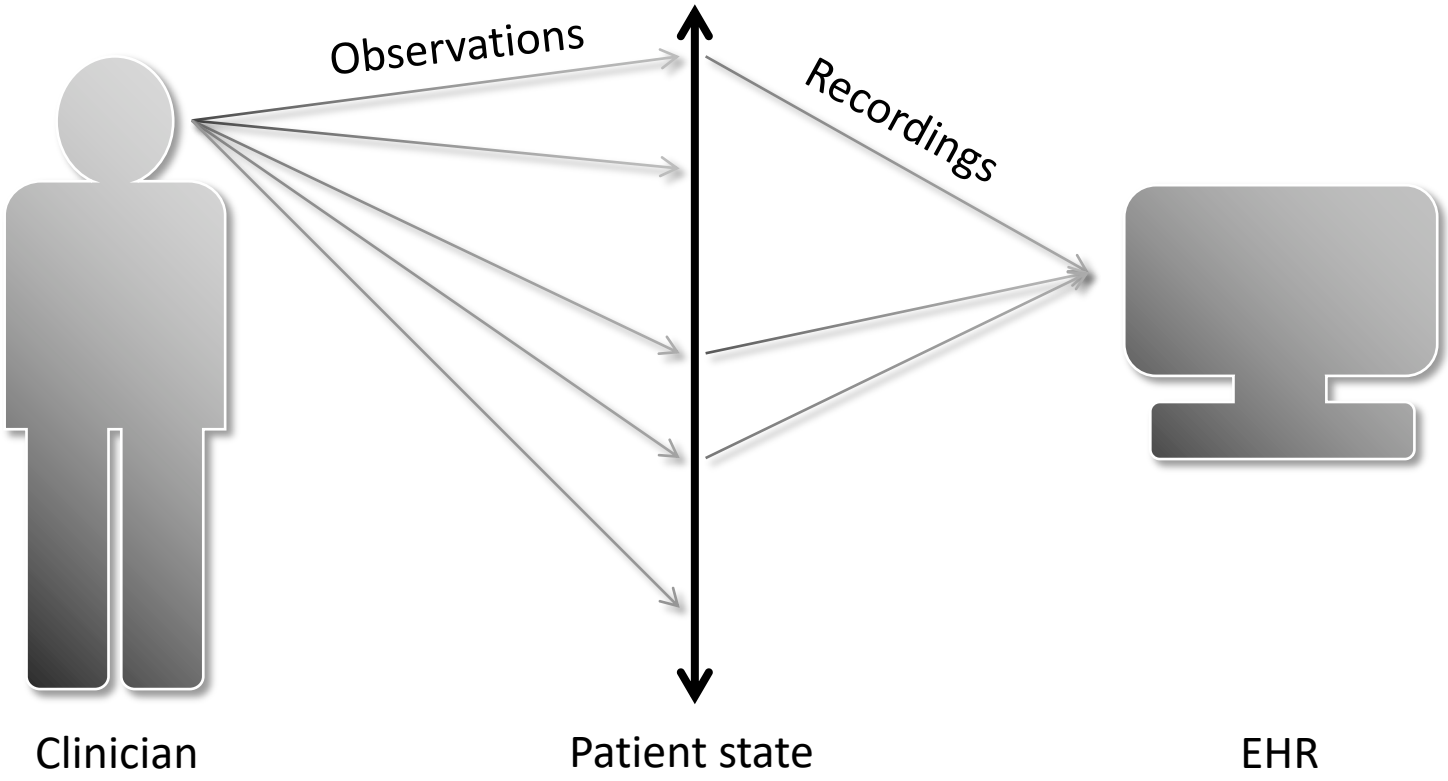
EHR data completeness is task-dependent

Weiskopf, Hripcsak, Swaminathan, Weng (in press) Defining and measuring completeness of electronic health records for secondary use.

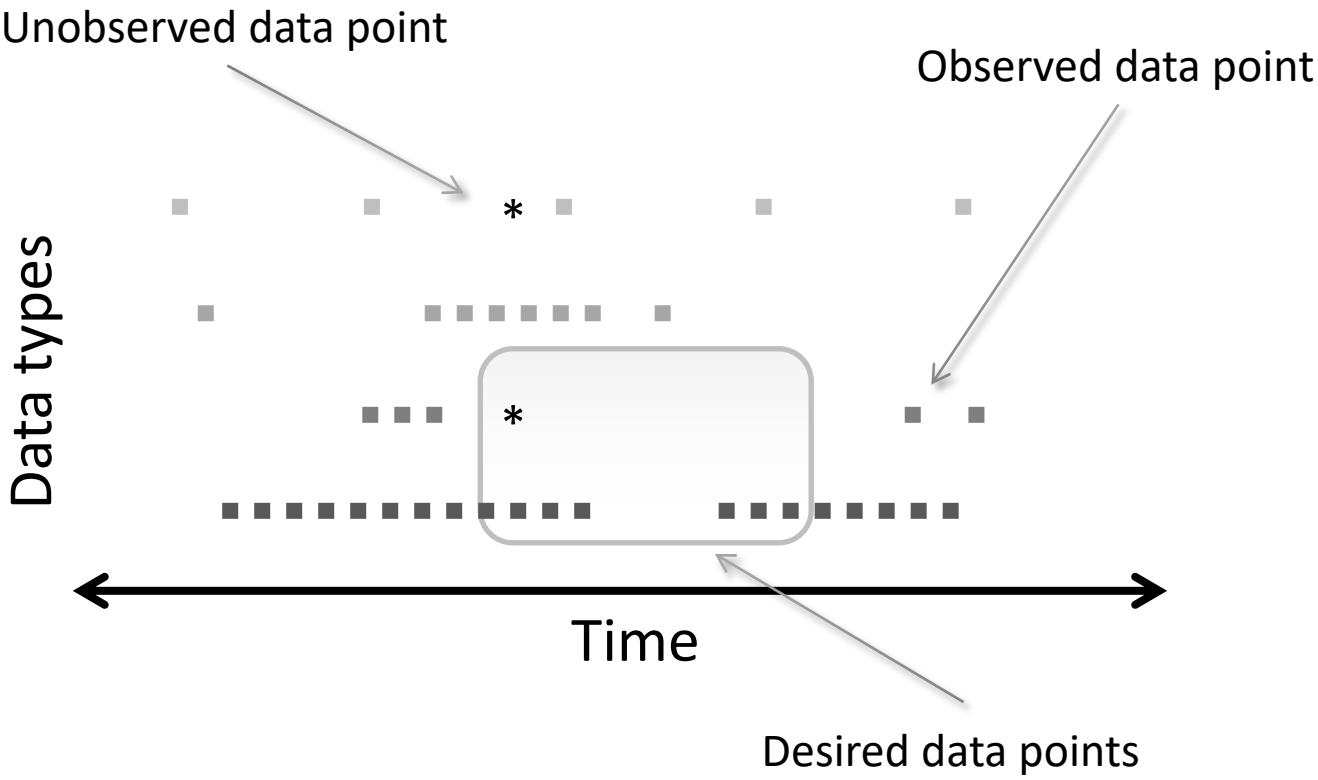
A potential data point may be observed or unobserved…



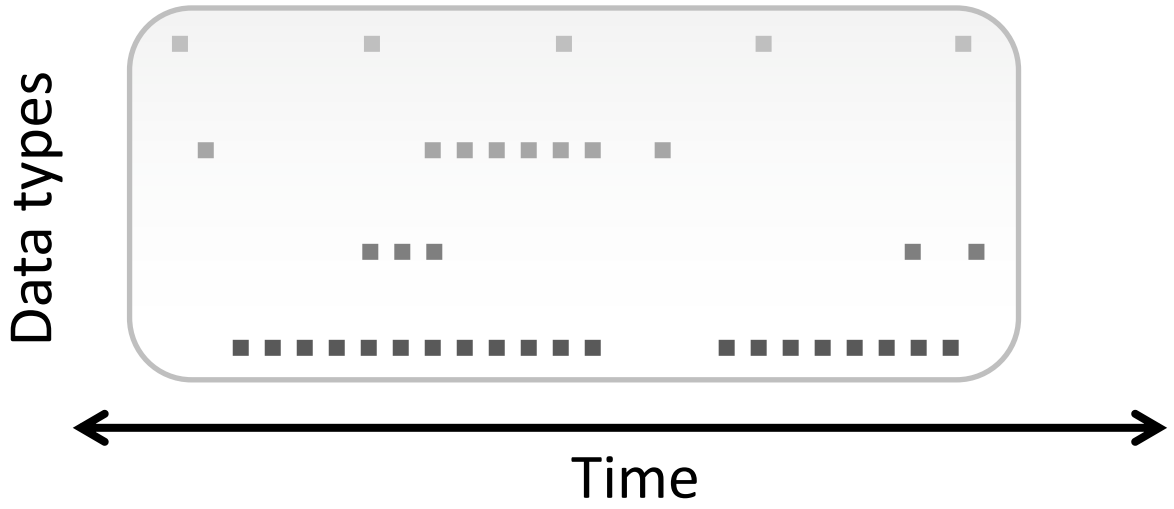
…and recorded or unrecorded



Patient Record Representation

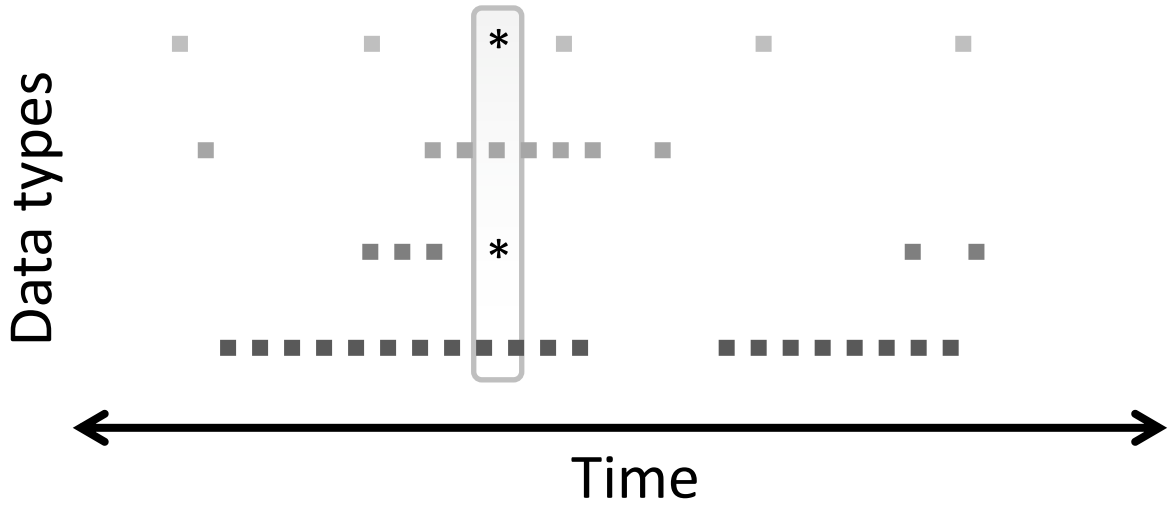


A record is complete when it contains all observations



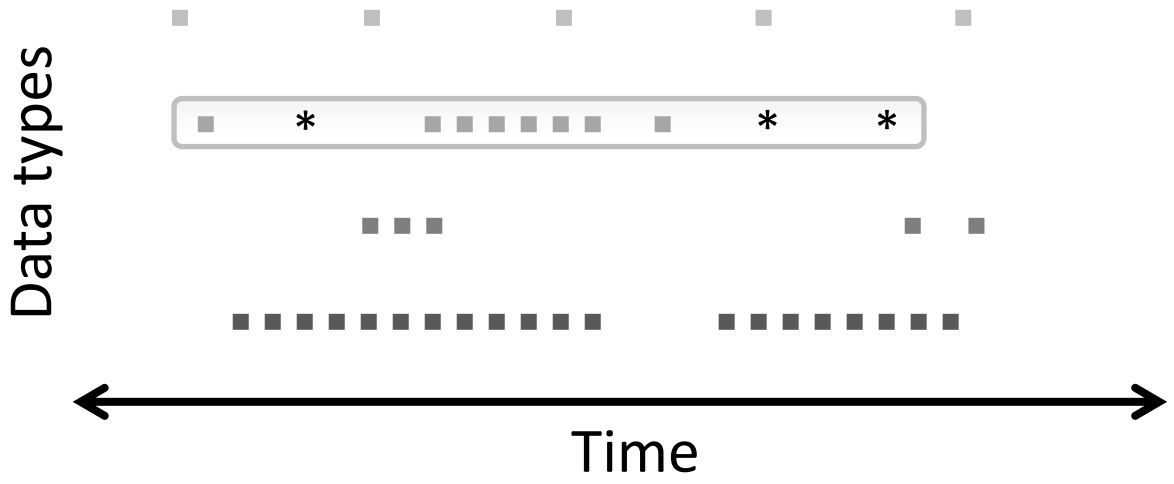
Documentation Completeness

A record is complete when it contains all desired or expected types of data



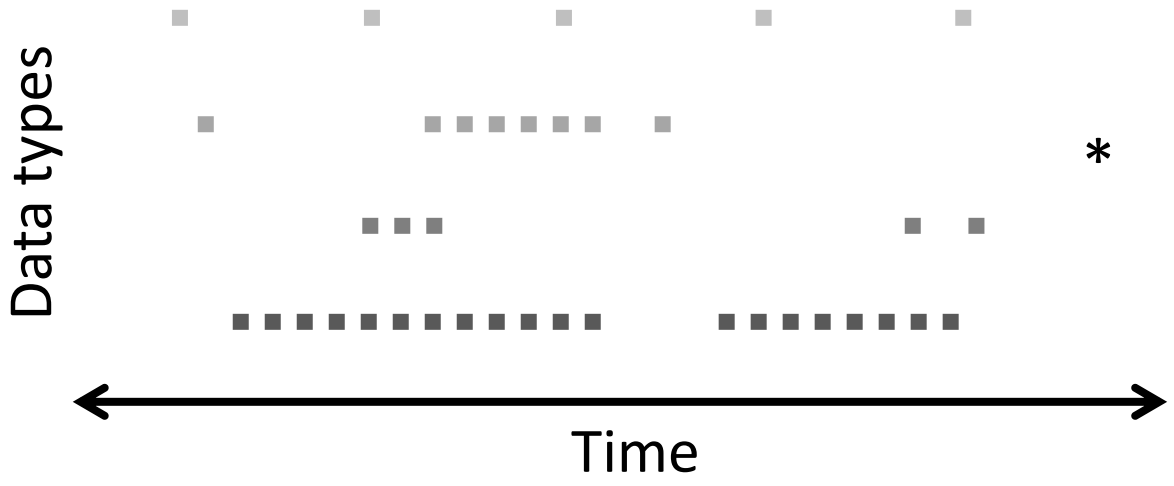
Breadth Completeness

A record is complete when it contains a specified frequency of data points over time



Density Completeness

A record is complete when it contains sufficient information to predict a clinical phenomenon of interest

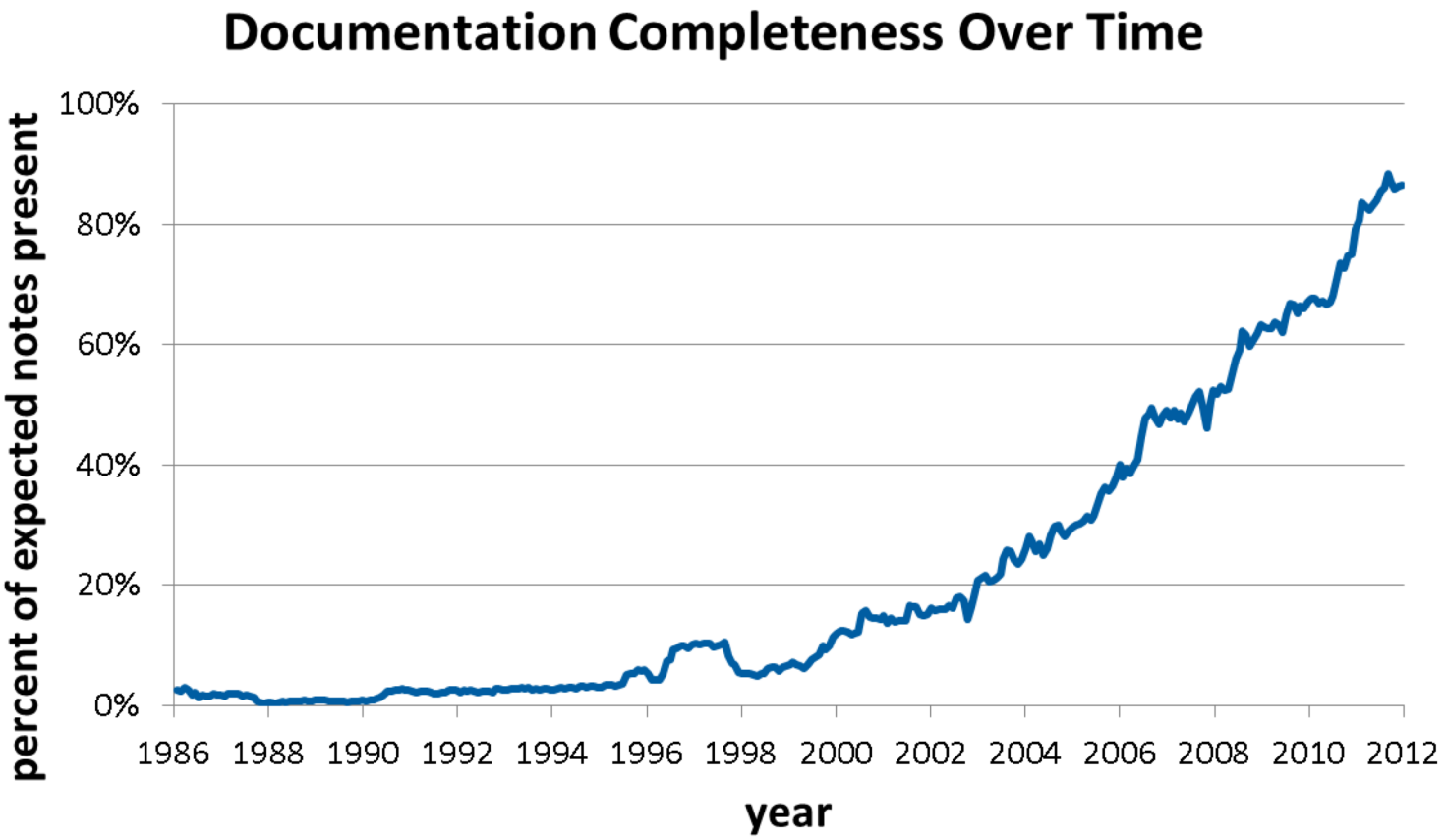


Predictive Completeness

Data

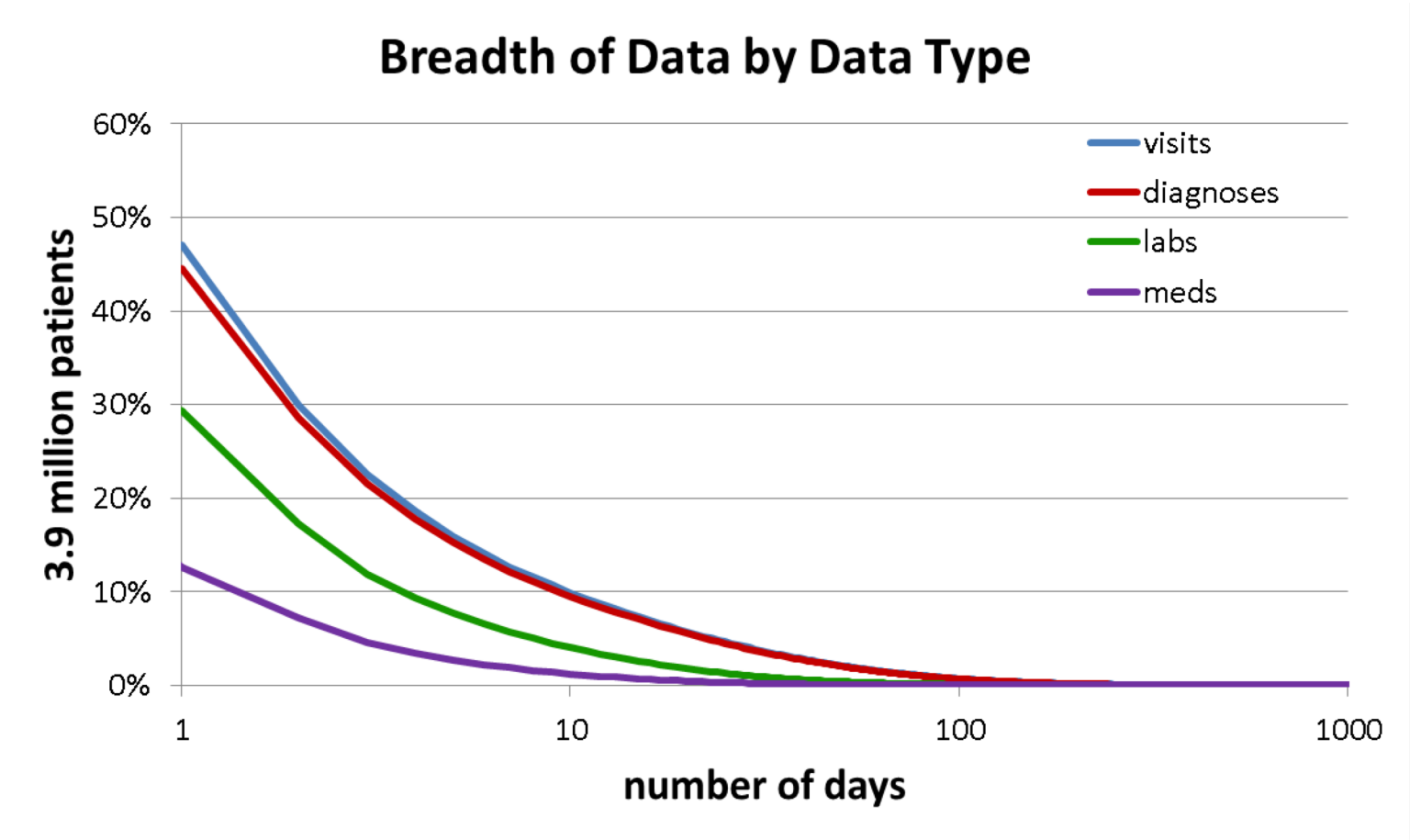
- NewYork-Presbyterian Hospital
  - Milstein Hospital, Allen Hospital, Morgan Stanley Children's Hospital
- 300,000 unique patients per year
- 3.9 million unique patient records in electronic clinical data warehouse
- Population
  - 56% female
  - average age of 51 years
  - 32% Hispanic, 10% Asian, 19% Black, and 39% White

Completeness has increased with improved adoption of HIT

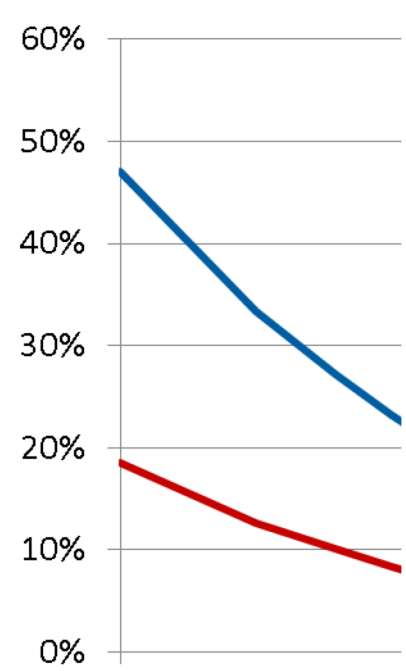


Data: narrative notes

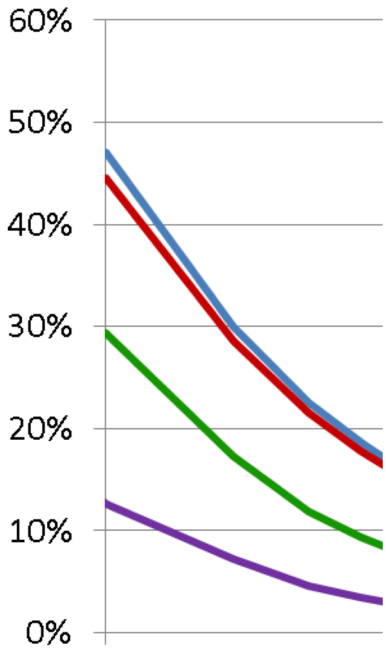
Many visits are not accompanied by common data types



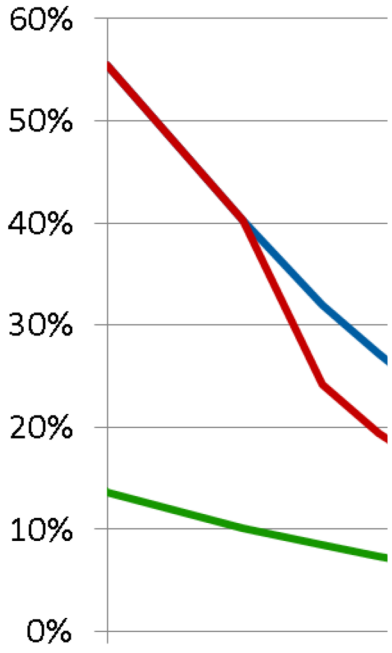
By any definition, approximately half or more of all patient records in our EHR are incomplete



Documentation

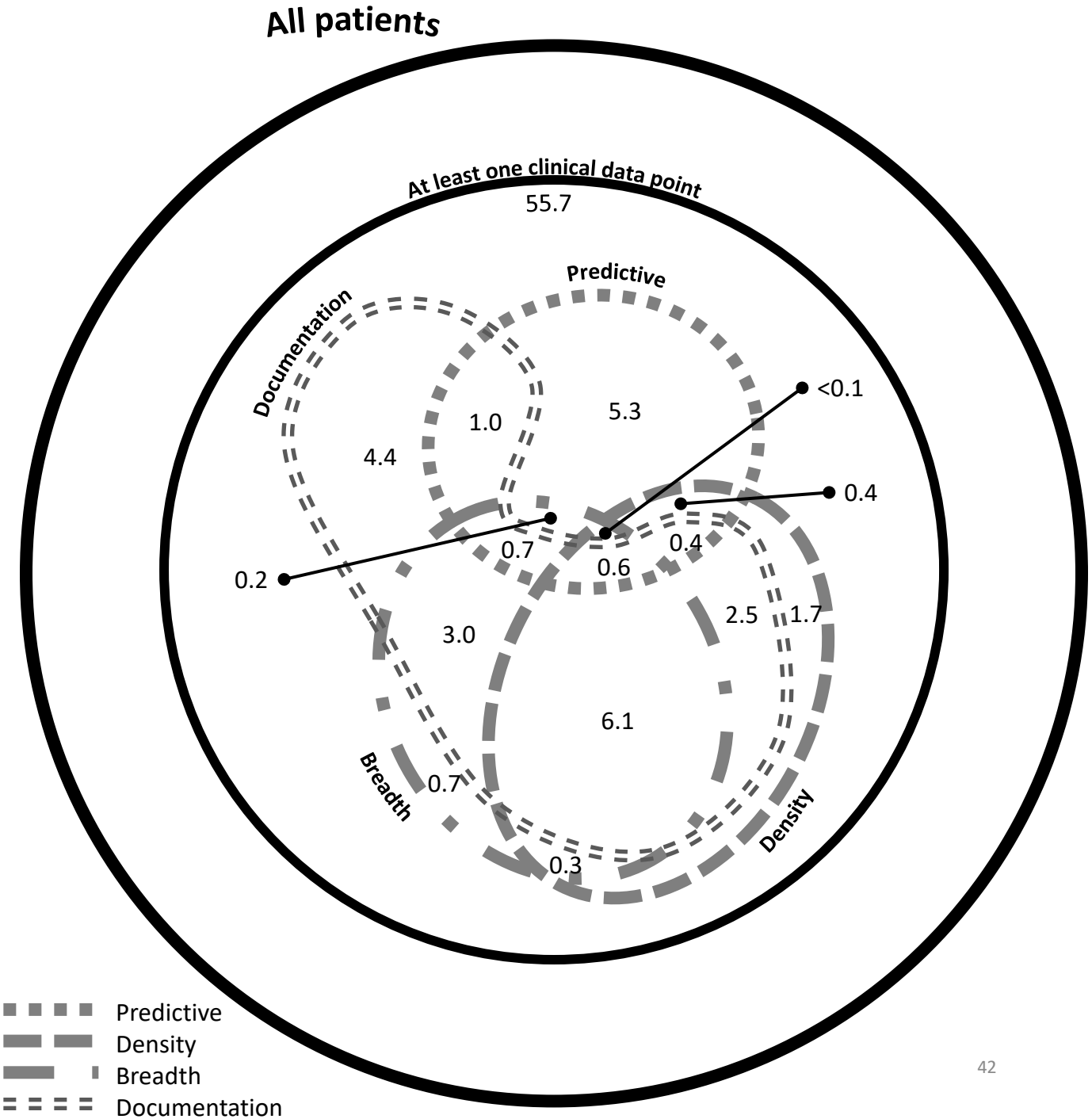


Breadth



Density

Overall, **55.7%** of patients in the CDW have at least one point of clinical data, and **26.9%** meet the criteria for at least one definition of completeness. Patients with documentation complete records—meaning they had at least one visit with an associated note—accounted for **18.5%** of all patients. In terms of density, only **11.8%** have a complete record when completeness is defined as at least 15 laboratory results or medication orders adjusted for temporal variance. When completeness is defined as a breadth of five data types of interest (date of birth, sex, medication order, laboratory test, and diagnosis), **11.4%** of patients have complete records. Finally, the presence or absence of a gap of 180 days or more could be correctly predicted for **8.4%** of patients. Only **0.6%** of patient records could be considered complete according to the implementations of all four definitions.



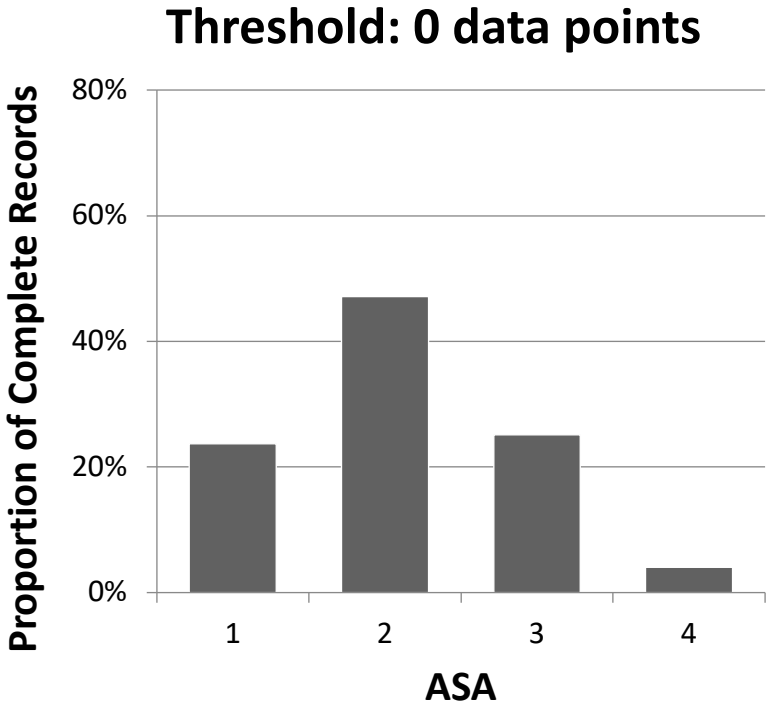
Hypothesis: sick patients are likely to have more complete records

Weiskopf, Rusanov, Weng. Under review.

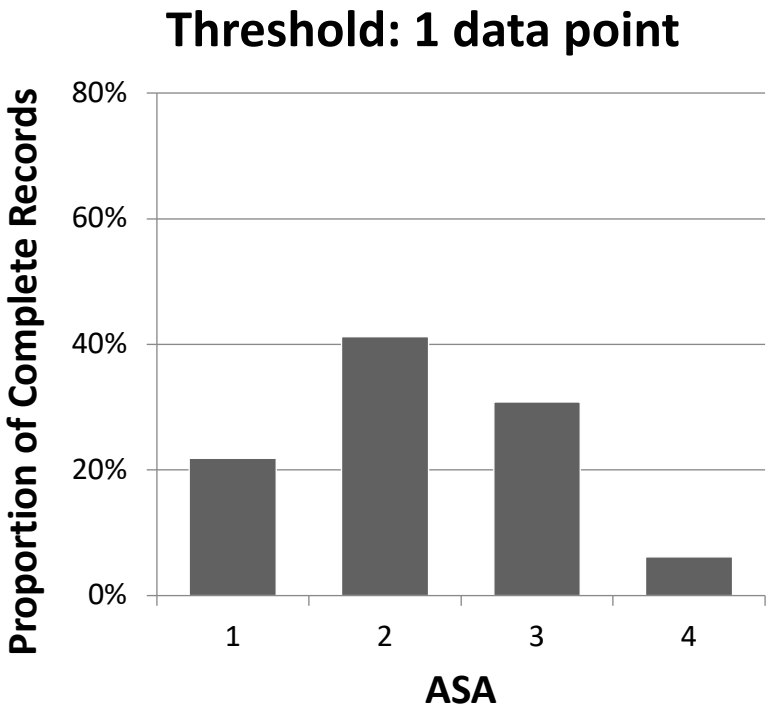
Methods

- Population
  - 5,000 patients who have received anesthesia services
- Data
  - American Society of Anesthesiology (ASA) Physical Status Classification (1 to 6 scale; limited to 1 to 4)
  - Density of laboratory results (no time correction)
  - Density of medication orders (no time correction)

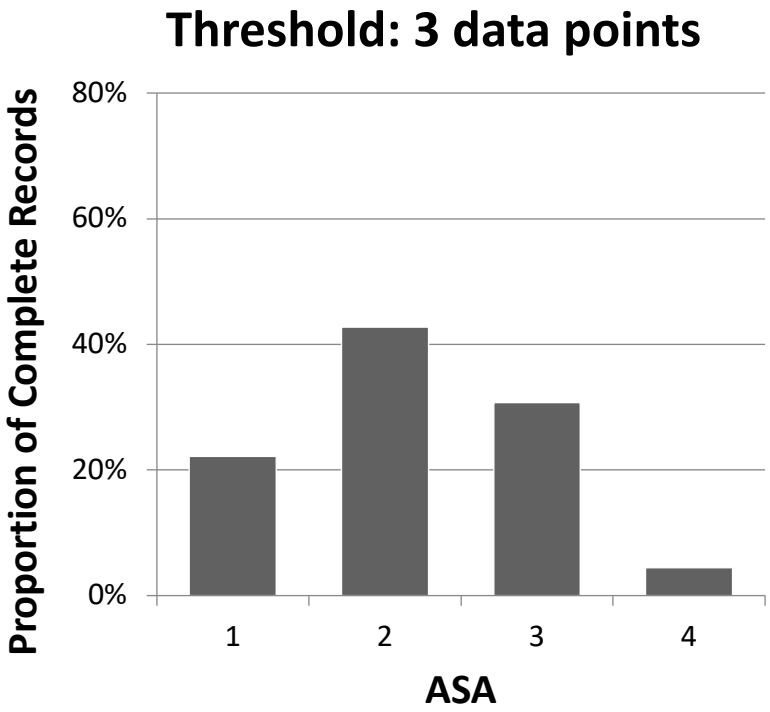
Distribution of ASA scores changes with threshold for completeness



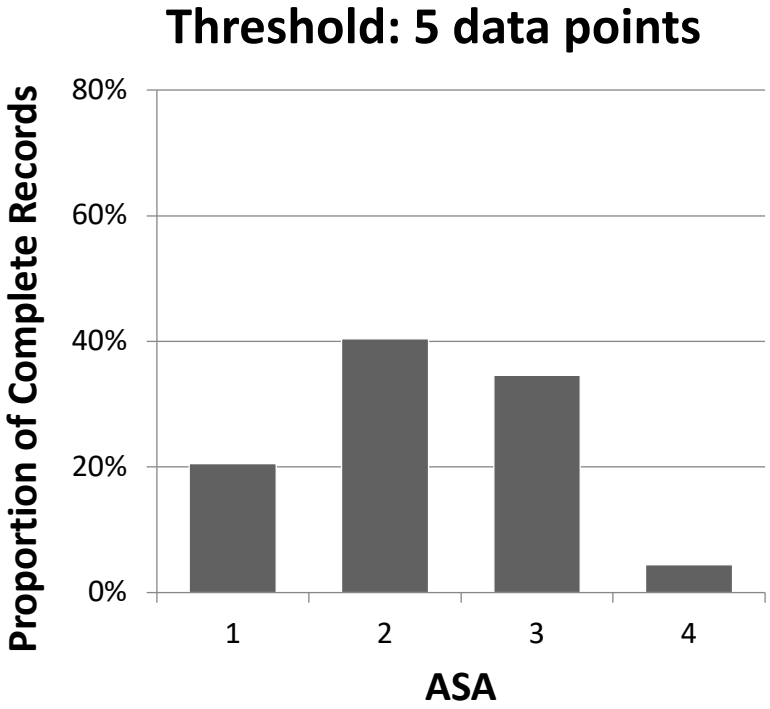
Distribution of ASA scores changes with threshold for completeness



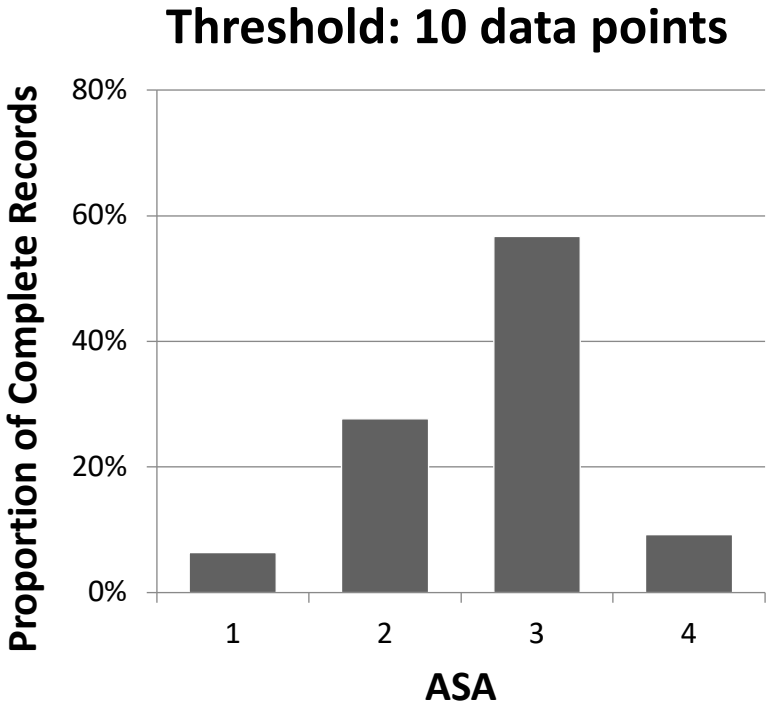
Distribution of ASA scores changes with threshold for completeness



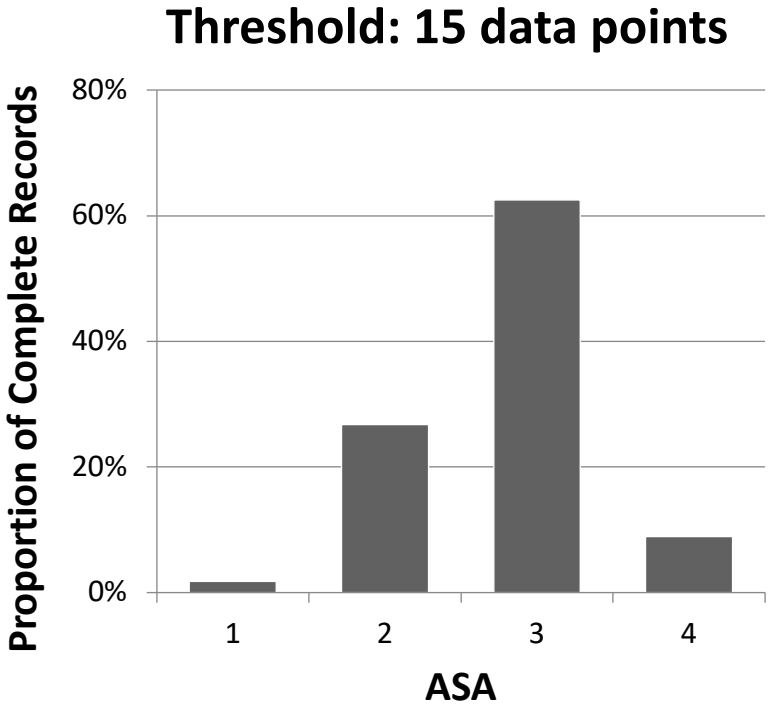
Distribution of ASA scores changes with threshold for completeness



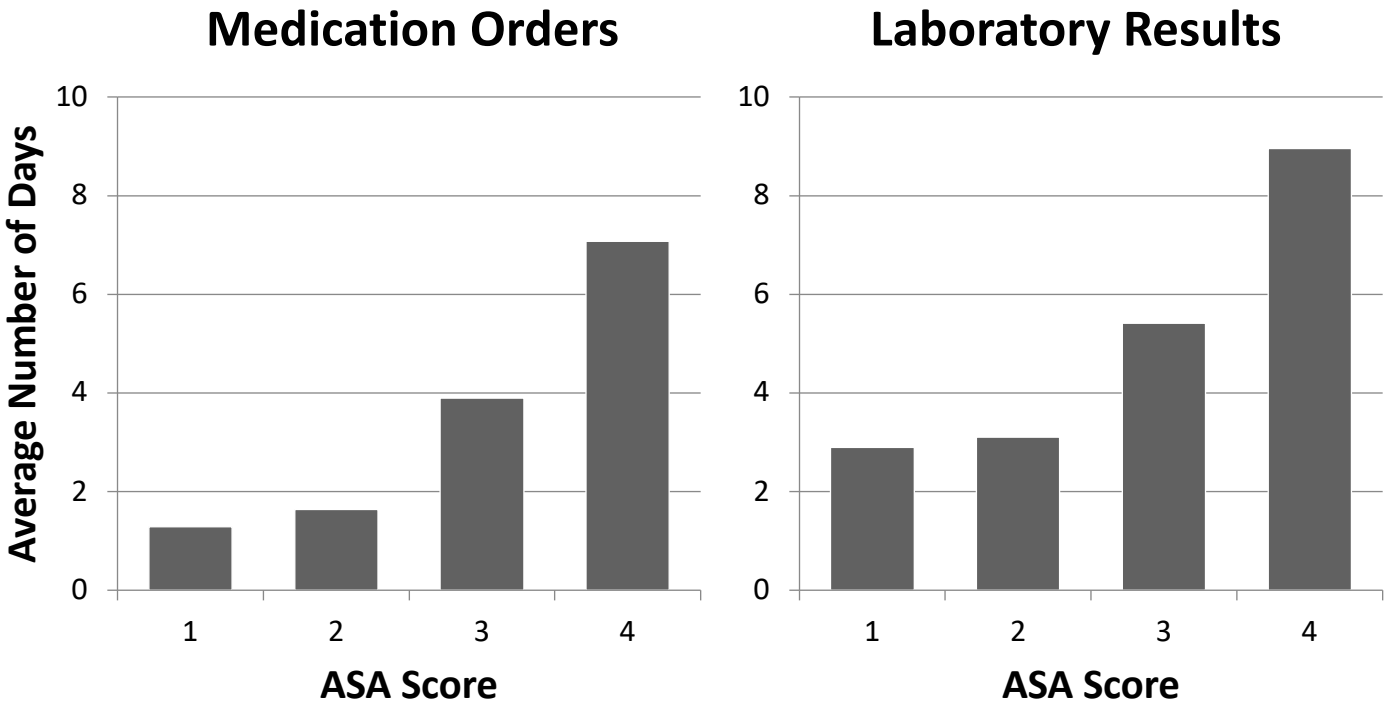
Distribution of ASA scores changes with threshold for completeness



Distribution of ASA scores changes with threshold for completeness



Counts of available data points differ across ASA scores



Take Home:

- Be mindful of the potential limitations of a dataset prior to committing to its use
- Be explicit in your choice of completeness definition before assessing quality and suitability of a dataset
- Be transparent about data quality findings when reporting results

Take Home:

- Be mindful of the potential limitations of a dataset prior to committing to its use
- Be explicit in your choice of completeness definition before assessing quality and suitability of a dataset
- Be transparent about data quality findings when reporting results

Take Home:

- Be mindful of the potential limitations of a dataset prior to committing to its use
- Be explicit in your choice of completeness definition before assessing quality and suitability of a dataset
- Be transparent about data quality findings when reporting results

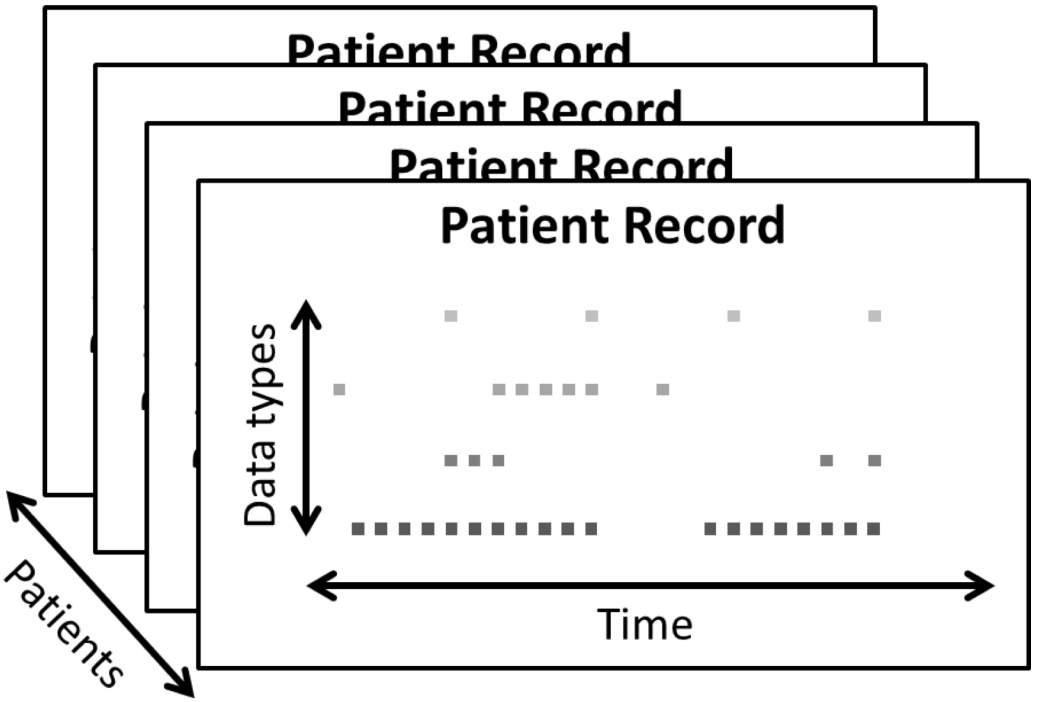
Take Home:

- Be mindful of the potential limitations of a dataset prior to committing to its use
- Be explicit in your choice of completeness definition before assessing quality and suitability of a dataset
- Be transparent about data quality findings when reporting results

EHR Data Quality Assessment Framework Based on Clinical Researcher Needs

Dimensions of EHR Data

"Completeness: breadth, depth, and scope of information contained in the data."



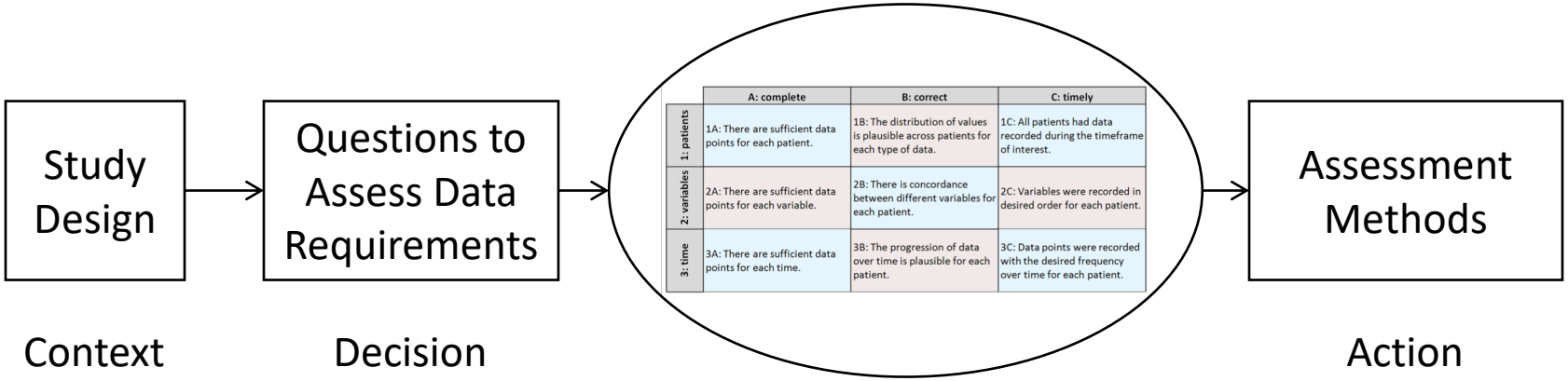
Wang & Strong (1996) Beyond accuracy: what data quality means to data consumers.

Weiskopf, Weng, Swaminathan, Hripcasak. (2013) Defining and measuring completeness of electronic health records for secondary use.

Data Quality Assessment Framework

	A: complete	B: correct	C: timely
1: patients	1A: There are sufficient data points for each patient.	1B: The distribution of values is plausible across patients for each type of data.	1C: All patients had data recorded during the timeframe of interest.
2: variables	2A: There are sufficient data points for each variable.	2B: There is concordance between different variables for each patient.	2C: Variables were recorded in desired order for each patient.
3: time	3A: There are sufficient data points for each time.	3B: The progression of data over time is plausible for each patient.	3C: Data points were recorded with the desired frequency over time for each patient.

Using DQA Framework

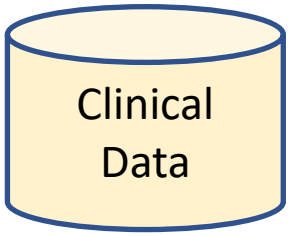


Extension of the Harmonized Data Quality Framework: Early Thoughts

Jimmy Rogers, MS; Chunhua Weng, PhD

Department of Biomedical Informatics, Columbia University

Current Practice: IT Centered DQ



"Do I apply all DQ rules? Or are only some of the DQ rules applicable?"

"Without domain knowledge and clinical or research tasks, how would I know what rules are relevant to measure the fitness of use of clinical data?"

Potential Stakeholder DQ Needs

- *"I want to apply the same set of DQ rules that Michael applied to his database to my CDW and compare the quality of our CDWs"*
- *"I want to get a set of DQ rules for measuring the completeness of problem lists" (or replace problem lists with ICD-9 diagnosis codes)*
- *"find me all the DQ rules for diabetes"*
- *"find me all the DQ rules for breast cancer"*
- *"For this project, relevant DQ rules include completeness and correctness of timestamps; can I review and select from all related rules?"*
- *"DQ rules for elevated troponin may vary among different contexts; In clinical setting, elevated troponin has a threshold of 0.01 ng/ml, while in this research study, abnormal troponin should be > 0.04 ng/ml"*
- *"find me all the DQ rules identifying data errors, questionable data, potentially questionable data, respectively"*

The Harmonized Framework for Data Quality (Kahn MG et al, 2016)

- **Conformance**
  - Value → comply with pre-specified formatting constraints
  - Relational → comply with primary/foreign key relationships
  - Calculation → computational accuracy and feasibility
- **Completeness**
  - Atemporal → value present at a single point
  - Temporal → value present at multiple points across time
- **Plausibility**
  - Uniqueness → presence of duplicate values
  - Atemporal → values and distributions of values are feasible at a single point
  - Temporal → values and distributions of values are feasible across time

A Use Case for The Framework: DQ Rule Categorization (Callahan et al, 2017)

Table 2. DQ Check Coverage by DQA Category by Organization

DQ HARMONIZATION TERMINOLOGY CATEGORIES		ORGANIZATIONS						
		CESR N (%)	MURDOCK N (%)	OHDSI N (%)	PEDSnet N (%)	PHIS N (%)	SENTINEL N (%)	TOTAL N (%)
Conformance	Value	1,434 (41.76)	43 (1.34)	0 (0.00)	3 (0.34)	65.5 (3.57)	421 (28.31)	19,66.5 (17.84)
	Relational	786 (22.89)	36 (1.12)	25 (14.53)	13 (1.49)	114 (6.21)	42 (2.82)	1,016 (9.22)
	Calculation	50 (1.46)	0 (0.00)	5 (2.91)	0 (0.00)	10 (0.54)	1 (0.07)	66 (0.60)
Completeness	Atemporal	754 (21.96)	9 (0.28)	3 (1.74)	367.5 (42.00)	186.5 (10.16)	111 (7.46)	1,431 (12.98)
	Temporal	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	22 (1.20)	0 (0.00)	22 (0.20)
Plausibility	Uniqueness	1 (0.03)	0 (0.00)	0 (0.00)	0 (0.00)	29 (1.58)	18 (1.21)	48 (0.44)
	Atemporal	207 (6.03)	3,031 (94.13)	87 (50.58)	315 (36.00)	1,300 (70.84)	527 (35.44)	5,467 (49.60)
	Temporal	202 (5.88)	101 (3.14)	52 (30.23)	176.5 (20.17)	108 (5.89)	367 (24.68)	1,006.5 (9.13)
Provided DQ Checks		**3,434**	**3,220**	**172**	**875**	**1,835**	**1,487**	**11,023**

Towards A System for Indexing, Querying, and Retrieval of DQ Rules

- The harmonized DQ framework (Kahn et al.): establishes a good foundation with high-level concepts and can be extended
- The DQ ontology (Johnson et al.): focus on concept definition rather than support of real-world DQ application and utilization
- A Data Quality Assessment Guideline for EHR data reuse (Weiskopf et al.): focus on decision support and rule selection for various tasks, not on rule management

All expert-driven, not data driven

Can we enrich DQ rules with

adoption status

Reference standard

creator

disease domain

source

lab name

Measurement thresholds

Error type

Data source

Error cause

…Can add more metadata tags

Our Goal

- A knowledge base of human reviewable and machine executable DQ rules for
  - DQ Knowledge management and sharing
  - Stakeholders to perform knowledge-based DQ
- Supports the following tasks around DQ rules
  - Indexing: how can we organize rules in an understandable way?
  - Retrieval: which DQ rules suits my task?
  - Auditing: how are rules related? Is there redundancy? When will be a new rule be necessary?
  - Machine learning: generate rules systematically and automatically?
- Be
  - Stakeholder friendly
  - Fine grained
  - extensible

References

1. Callahan, T. J. et al. **A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks.** EGEMs Gener. Evid. Methods Improve Patient Outcomes 5, (2017).
2. Kahn MG, Callahan TJ, Barnard JG, Bauck A, Brown JS, Estiri H, Görg C, Holve E, Johnson SG, Liaw ST, Lopez MH, Meeker D, Ong TC, Ryan PB, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling LM, **A Harmonized Data Quality Assessment Framework for the Secondary Use of Electronic Health Record Data**, eGEMS, 2016, in press.
3. Weiskopf NG, Bakken S, Hripcsak G, **Weng C**, **A Data Quality Assessment Guideline for Electronic Health Record Data Reuse,** *eGEMS*, accepted
4. Khare R, Utidjian L, Ruth BJ, Kahn MG, Burrows E, Marsolo K, Patibandla N, Razzaghi H, Colvin R, Ranade D, Kitzmiller M, Eckrich D, Bailey LC. **A longitudinal analysis of data quality in a large pediatric data research network.** J Am Med Inform Assoc, 2017, April 8.
5. Shang N, Weng C, Hripcsak G, **A Conceptual Framework for Evaluating Data Suitability for Observational Studies,** JAMIA, in press.
6. Huser, Vojtech; DeFalco, Frank J.; Schuemie, Martijn; Ryan, Patrick B.; Shang, Ning; Velez, Mark; Park, Rae Woong; Boyce, Richard D.; Duke, Jon; Khare, Ritu; Utidjian, Levon; and Bailey, Charles (2016) **Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets**, eGEMs: Vol. 4: Iss. 1, Article 24. DOI: http://dx.doi.org/10.13063/2327-9214.1239
7. Weiskopf NG, Hripcsak G, Sushmita S, Weng C, **Defining and measuring completeness for electronic health records for secondary use.** *J Biomed Inform,* 2013, Jun 29. doi:pii: S1532-0464(13)00085-3.
8. Weiskopf NG, Weng C, **Methods and Dimensions of EHR Data Quality Assessment: Enabling Reuse for Clinical Research**, *J Am Med Inform Assoc.* 2013 Jan 1;20(1):144-51.
9. Rusanov A, Weiskopf NG, Wang S, Weng C, **Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research,** *BMC Medical Informatics and Decision Making*, 2014.
10. Weiskopf NG, Bakken S, Weng C, **Are EHR Data Suitable for Secondary Use? Researcher Views.** *2014 AMIA Joint Summits on Translational Science*.
11. Stephens, KA. "**Data QUEST: Data Quality Testing – DQe Tools**." PCORI Data Quality Meeting, 28 February 2017. Presentation.
12. Lyman, KA. "**REACHnet Data Quality Approach.**" PCORI Data Quality Meeting, 11 May 2017. Presentation.