Data Engineering in Healthcare:
progress and remaining challenges

Roberto A. Rocha, MD, PhD, FACMI
Managing Director
Semedy, Inc.
Assistant Professor of Medicine
Division of General Internal Medicine and Primary Care
Brigham and Women's Hospital, Harvard Medical School

1st International Conference on AI in Healthcare (ICAIH): March, 2018

CC BY NC SA

Disclosures

- Salary support from Semedy, Inc.
- Salary support from Wolters Kluwer Health (spouse)

Overview

- Motivation
- Background
- Interoperability
- Information models
- Conclusions

MOTIVATION

Data comparability and consistency

- "Progress on the road toward **integrating big data** — both high-volume genomic findings and heterogeneous clinical observations — into **practical clinical protocols** and **standard healthcare delivery** requires that providers, HIT vendors, federated knowledge resources, and patients can ultimately depend upon those data being **comparable** and **consistent**."
  - "Absent comparability, the data are more or less by definition **not able to support inferencing of any scalable** kind …"
  - "Without consistency, users of complex biomedical data will have to spend added **resources transforming the data into usable and predictable formats** …"

Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for big data in translational research. *Genet Med*. 2013 October ; 15(10): 802–809

Challenges for AI

- "There is a **great deal of interest** in the potential of using the vast data sets represented in electronic health records (EHRs), in combination with AI algorithms …"
  - "… AI can perform with great accuracy when the relationship between diagnostic data and the diagnosis is well defined, when the relationship between the data and the diagnosis suffers from **error, variability or difficulty in discrimination**, AI algorithms also perform less well."
- "**Extreme care is needed** in using EHRs as training sets for AI, where outputs may be **useless or misleading** if the training sets contain incorrect information or information with unexpected internal correlations."

JASON Report. Artificial Intelligence for Health and Health Care (JSR-17-Task-002). December 2017.

Relevance of data engineering

- "... two viable solutions to address heterogeneous data:
  - defining a "common representation" and transforming all data into that **common interlingua**, or
  - **adopting standards at the point of data generation** to obviate the costs and confusion that often emerge from data transformation."

- "... inferences will have hugely **more power and accuracy** if we aim big data methods at information that shares names and values"
  - "we do not want to waste analytic resources "discovering" that renal cancer behaves similarly to kidney cancer""

Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for big data in translational research. *Genet Med*. 2013 October ; 15(10): 802–809

Promising preliminary results

- "**Low-volume**, **structured** clinical data contain sufficient information to train classifiers to perform near physician-level."
  - 799 cases independently validated by more than 2 medical professionals (medical students and physicians)
- "**Collecting such data is possible** through human computation that is independently useful to clinicians."

W Scott, Lin I, Komarneni J, Nundy S. Machine Classifier Trained on Low-Volume, Structured Data Predicts Diagnoses Near Physician-Level: Chest Pain Case Study. *39th Annual Meeting of the Society for Medical Decision Making*. October 2017.

BACKGROUND

Data engineering?

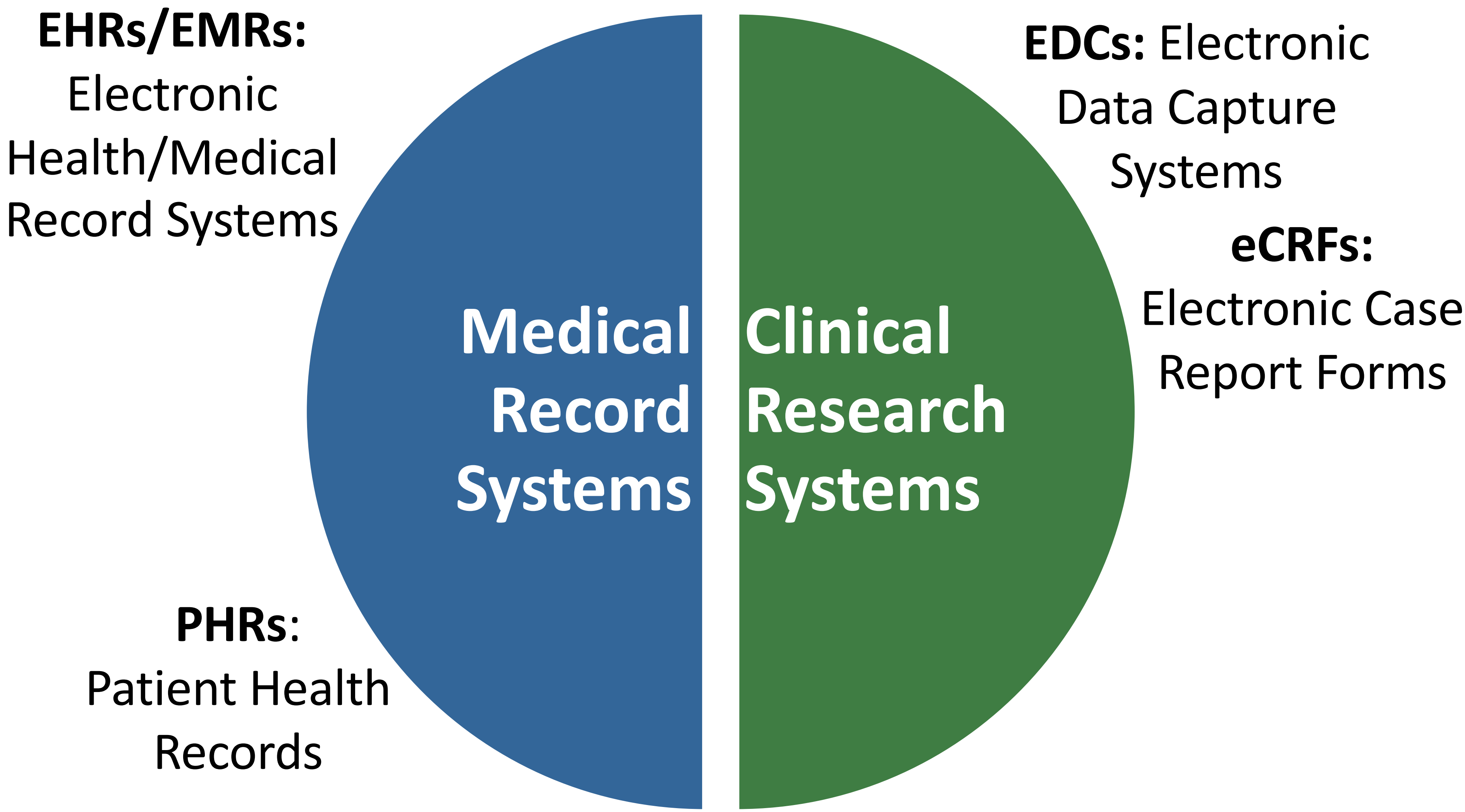
- “The design, implementation, modeling, theory and application of database systems and their technology.” (IEEE TCDE)

- Purposeful design and implementation of **models** and related **artifacts** to ensure **consistent**, **extensible**, and **interoperable** data representation
  - Akin to *knowledge* and *terminology* engineering

Clinical data

- Highly complex
  - Diverse types of data
    - structured and unstructured, images, sounds, etc.
  - Dynamic (changing) nature
    - flexible and extensible underlying models
- Large quantities
  - High performance database environment
    - response time is the critical factor
- Confidentiality and privacy (security)

Clinical systems (data) dichotomy



EHR systems

- Electronic Health Record (EHR)
  - "electronic version of a patients medical history, … maintained by the provider over time, … clinical data relevant to that persons care … including **demographics**, **progress notes**, **problems**, **medications**, **vital signs**, **past medical history**, **immunizations**, **laboratory** … **radiology**" (CMS definition)

- US market
  - Commercial EHRs dominate; small number of vendors
  - Designed for **data storage** & **communication**: human users
  - High tolerance for **incomplete**, **incorrect**, **ambiguous data**
  - **Limited** capability for computer-assisted decision making
  - Recent emphasis on **data sharing** (government incentives)

Types of clinical data

- Unstructured
  - "Mr. Jones has **Appendicitis**"
  - No structure or codes
- Structured
  - *Diagnosis*: "**Appendicitis**"
  - Question is defined (coded) but answer (value) is free-text
- Structured & Coded
  - *Diagnosis*: **K35** (**Appendicitis**)
  - Question & answer are defined (coded)

Unstructured (narrative) data

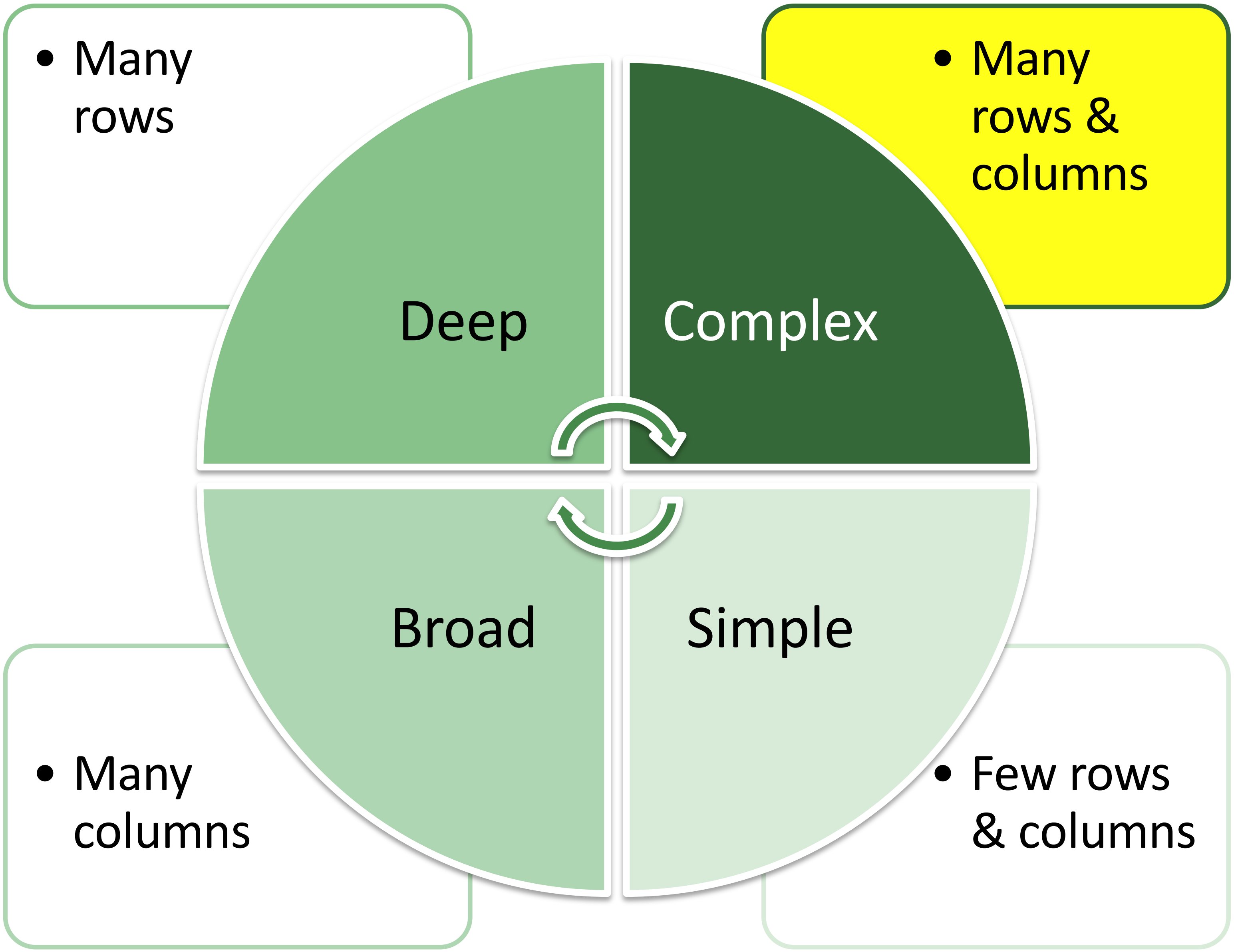
- Significant portion of the medical record is available as narrative data (**70-95%**)
    - medical history, physical exam, progress notes, discharge summary, radiology reports, operative notes
  - *advantages*: widespread, comprehensive, convenient, **expressive**, **natural**
  - *disadvantages*: ambiguous, complex, different styles, redundant, embedded errors, loose structure

Clinical phenotyping data

- "Intrinsically **complex**, fraught with **heterogeneity**, and amply having the potential for enormous **depth** (*many records*)."
- "A single patient may have **many thousands of unique attributes**, each of which may have arbitrarily repeated measures."

Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for big data in translational research. *Genet Med*. 2013 October ; 15(10): 802–809

Data complexity



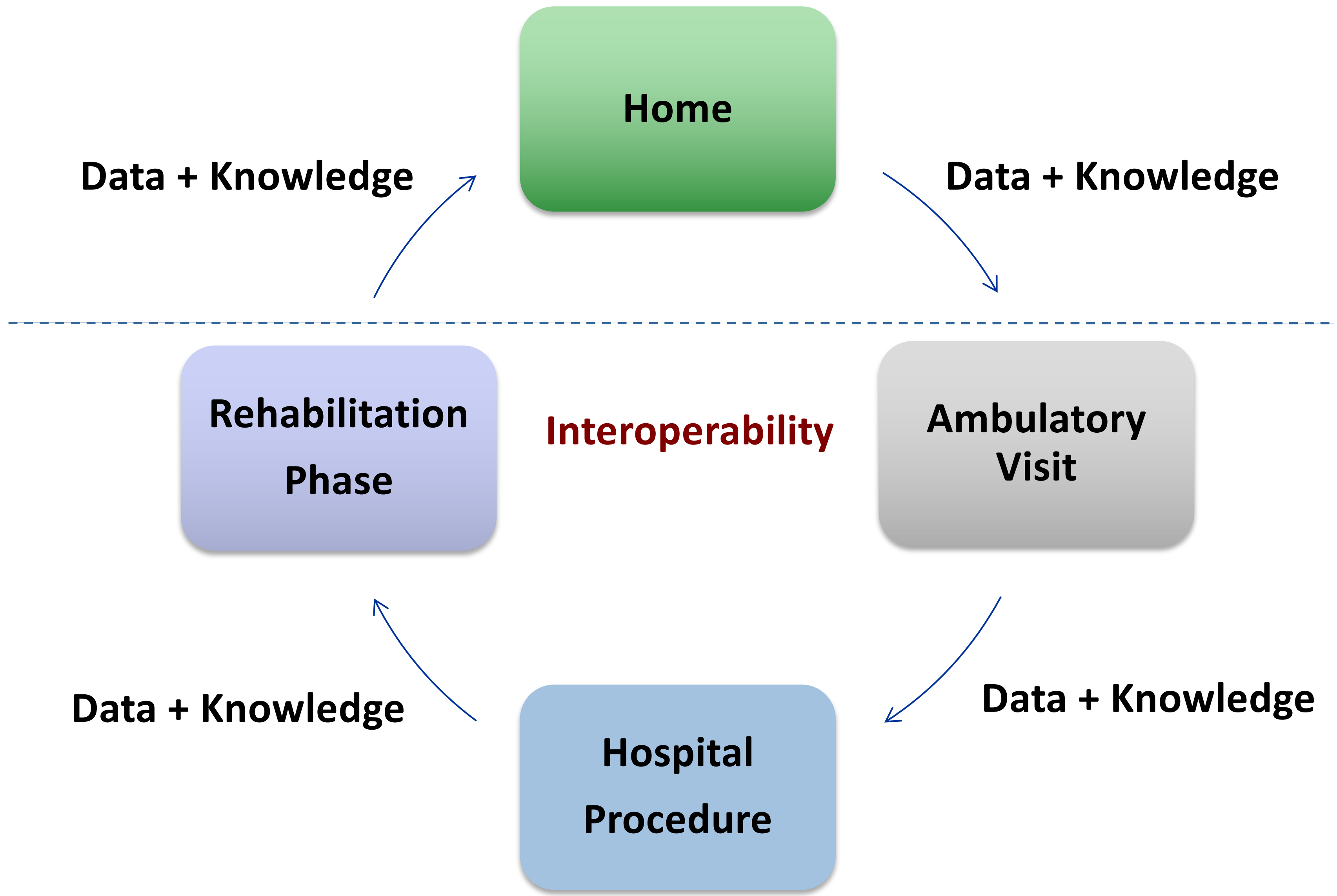
Why Data Engineering?

- Opportunity
  - Data defined with standard **reference models** – **consistency**, **completeness**, and **interoperability**
  - **Sustainable** process – new domains, single electronic record for **all settings** and **disciplines**

- Challenges
  - Data definitions **not shared** across EHR modules or settings – similar data stored and encoded **differently**
  - Large libraries of definitions – promote **inconsistency**, distinctions **not documented**, overlapping domains/topics
  - Manual verification – constantly **evolving** data collection tools (e.g. forms, flowsheets, templates, macros, etc.)

INTEROPERABILITY

Data & Knowledge interoperability



Clinical data representation: design/modeling

Define “**data points**” (**elements** and **values**) using available standards

Define logical **models** that combine **data points** and provide meaningful clinical **information**

Obtain detailed **provenance** to understand **how** and **when** data was created, and also **who** provided the data

Represent **semantic** and **temporal linkages** between data instances

Clinical data: standards

- **Data elements** and **data values**
  - *Available*: reference terminologies and ontologies
- **Information models**
  - *Work in progress*: isolated efforts and collections
  - *Available*: clinical documents (multiple types)
- **Provenance models**
  - *Work in progress*: competing models
- **Semantic** and **temporal linkages**
  - *Preliminary work*

Additional work is needed → opportunities

Development of standards

- SDO: Standards Developing Organization
- Many national and international organizations
  - Interdependencies and overlapping efforts
- Several with specific focus on Healthcare
  - Examples: HL7, IHE, DICOM, CDISC, …

LOINC

- **Logical Observation Identifiers Names and Codes**
- Organization: LOINC Committee
- Purpose: identification of laboratory and clinical observations (HL7 messages)
- Content: laboratory tests, clinical measurements, documents, etc.
- Information:
  - http://loinc.org

LOINC® from Regenstrief

systolic blood pressure 搜索

LOINC	详称	成分	属性	时间	体系	精度
24372-5	Peak systolic blood pressure --during right ventricular septal defect maximum velocity measurement	血管内心脏收缩期高峰^在右心室间隔缺损最大速度测量过程中	压力或压强	时间点	动脉系统.XXX	定量型
75997-7	Systolic blood pressure by Continuous non-invasive monitoring	血管内心脏收缩期	压力或压强	时间点	动脉系统	定量型
11378-7	Systolic blood pressure at First encounter	血管内心脏收缩期	压力或压强	就医过程持续时间^第一	动脉系统	定量型
20185-5	End systolic blood pressure by US	血管内心脏收缩期末期.XXX	压力或压强	时间点	循环系统.XXX	定量型
20186-3	Peak systolic blood pressure by US	血管内心脏收缩期高峰.XXX	压力或压强	时间点	循环系统.XXX	定量型
24370-9	Peak systolic blood pressure --during mitral regurgitation maximum velocity measurement	血管内心脏收缩期高峰^在二尖瓣反流最大速度测量过程中	压力或压强	时间点	动脉系统.XXX	定量型
24371-7	Peak systolic blood pressure --during aorta stenosis maximum velocity measurement	血管内心脏收缩期高峰^在主动脉狭窄最大速度测量过程中	压力或压强	时间点	动脉系统.XXX	定量型
45372-0	Blood pressure systolic and diastolic--post phlebotomy	收缩期与舒张期血压^在静脉采血之后	压力或压强	时间点	动脉系统	定量型
50402-7	Blood pressure systolic and diastolic--after transfusion	收缩期与舒张期血压^在输血之后	压力或压强	时间点	动脉系统	定量型
50403-5	Blood pressure systolic and diastolic--before transfusion	收缩期与舒张期血压^在输血之前	压力或压强	时间点	动脉系统	定量型

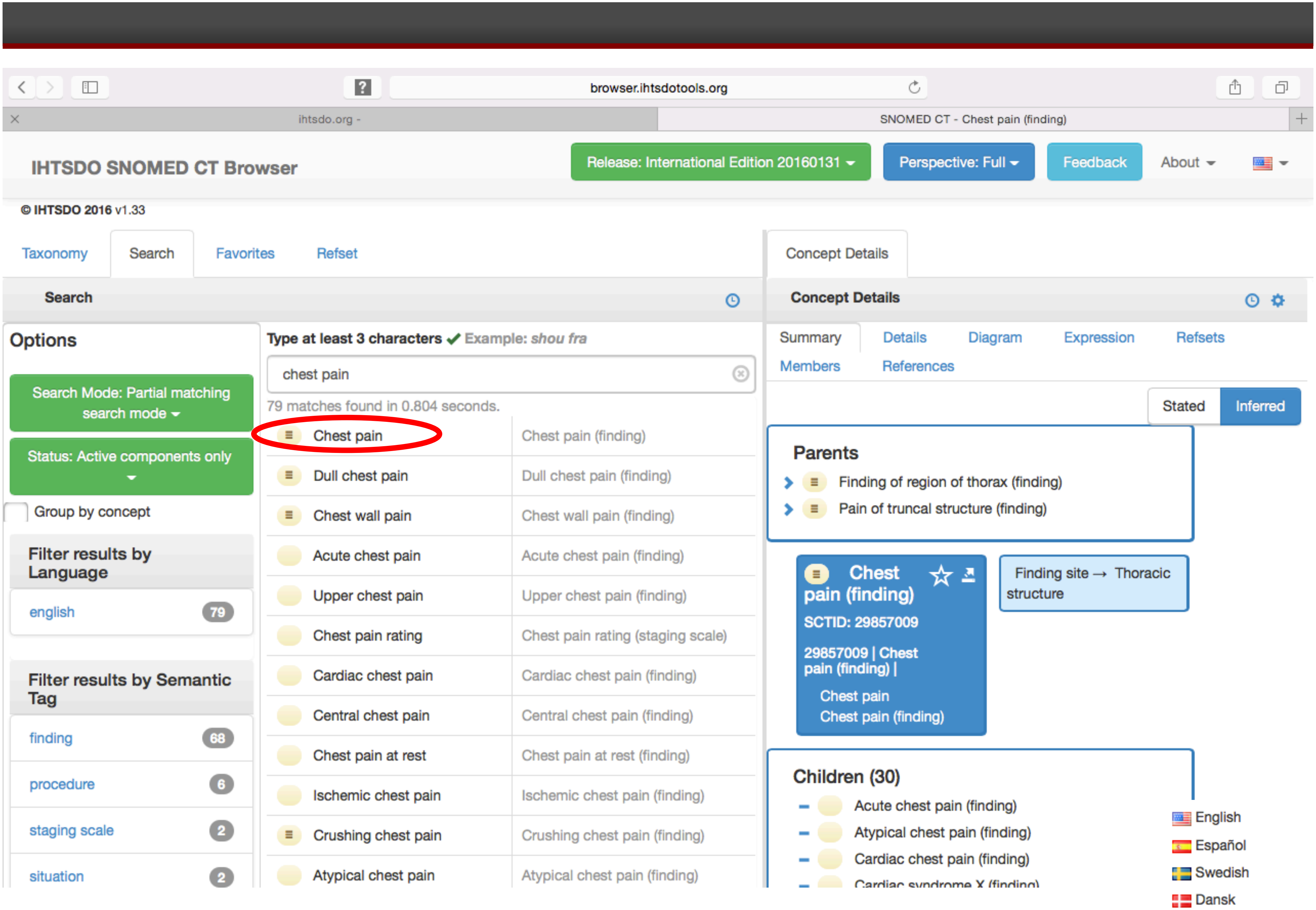
Search generated 102 hits in 0.239 secs.

版权® 2017 Regenstrief Institute Inc.

https://search.loinc.org

SNOMED CT

- **Systematized Nomenclature of Medicine – Clinical Terms**
- *Organization*: International Health Terminology Standards Development Organisation (**IHTSDO**)
  - SNOMED Terminology Solutions - College of American Pathologists
- *Purpose*: Encoding of multiple clinical domains
- *Content*: Comprehensive (diseases, signs, symptoms, living organisms, chemicals, body parts, morphology, occupations, modifiers, etc.)
- Information:
  - https://www.snomed.org



http://browser.ihtsdotools.org

Many others (incomplete list)

- **RxNorm**: clinical drugs and drug delivery devices (NLM)
  https://www.nlm.nih.gov/research/umls/rxnorm/

- **NDF-RT**: National Drug File - Reference Terminology (VA)
  http://evs.nci.nih.gov/ftp1/NDF-RT/

- **IIS**: Vaccination code sets (CDC)
  http://www.cdc.gov/vaccines/programs/iis/code-sets.html

- **HL7 Vocabulary domains** (messaging, documents, services)
  http://www.hl7.org/documentcenter/public_temp_1A973D1B-1C23-BA17-0CCBD68843B23790/standards/vocabulary/vocabulary_tables/infrastructure/vocabulary/vocabulary.html

Document standards

- **Clinical Document Architecture (CDA)**
- Organization: HL7 International
- Purpose: document markup standard that specifies the **structure** and **semantics** of "clinical documents" for the purpose of exchange between healthcare providers and patients
- Content:
  - Continuity of care, procedure note, patient assessments, etc.
  - Clinical oncology treatment plan, PHR plans, genetic testing reports, public cancer registries, etc.
  - Data Provenance
- Information:
  - http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7

Information models

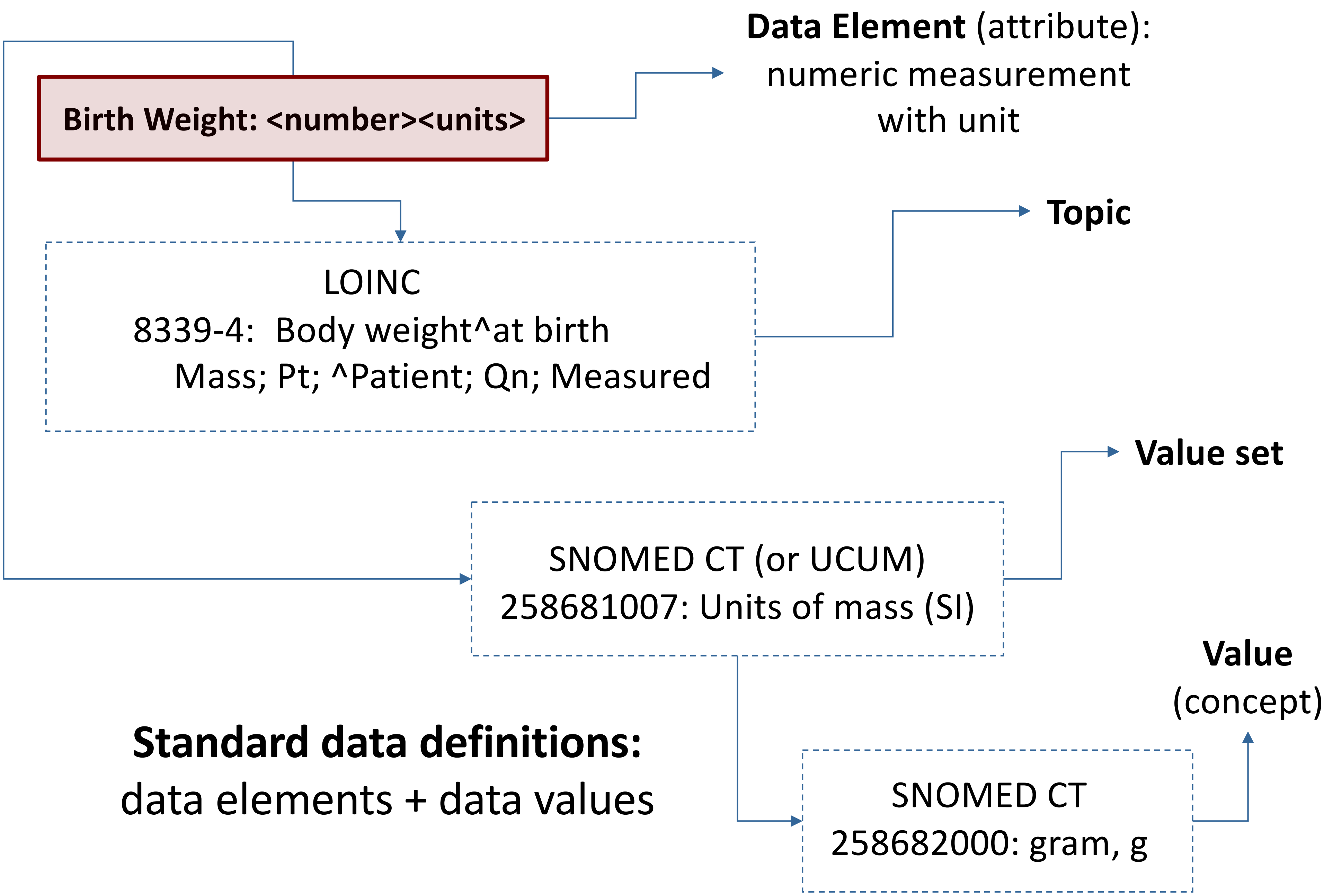
- **Clinical Information Modeling Initiative (CIMI)**
- Organization: HL7 International
- Purpose: Improve the interoperability of healthcare systems through shared implementable clinical information models - a **single curated collection** with bindings to reference **terminologies**
- Content: laboratory test results, vital signs, diagnoses, procedures, patient measures, etc.
- Information:
  - http://www.hl7.org/Special/Committees/cimi/index.cfm
  - http://www.opencimi.org

INFORMATION MODELS

Data Element	Data Type	Value Set List
SET: Pain Assessment (ALWAYS INCLUDE)		
Pain Episode Duration	Numerical	NA
Pain Episode Duration Time unit	Time with unit	Seconds | Minutes | Hours | Days | Weeks | Months | Years
Pain Location	Category	abdomen | achilles | ankle | arm | axilla | back | breast | buttocks | calf | chest | chin | coccyx | ear | elbow | eye | face | finger | flank | foot | forehead | groin | gum | hand | head | heel | hip | iliac crest | ischial tuberosity | jaw | knee | labia | leg | lip | lumbar | malleolus | mouth | mucous membrane | nail | nare | neck | nose | occipital region | parietal region | pelvic region | penis | perineum | peri-rectum | rectum | sacrum | scalp | scapula | scrotum | shoulder | sternum | suprapubic region | temporal region | thigh | toe | tongue | trochanter | umbilicus | vagina | wrist | Bladder| Clavicle| Epigastrium| Generalized| Gluteal| Mediastinum| Orbital region| Rib cage |Teeth |Thoracic spine |Throat | Uterine
Pain Location Qualifier	Category	Right | Left | Dorsal | Ventral | Anterior | Posterior | Bilateral | Distal | Proximal | Lower | Upper
Pain Quality	Category	Ache| Bloated| Colicky| Cramping| Crushing| Cutting| Deep| Diffuse| Dull | Electrical | Gnawing | Heavy pressure |Heightened sensitivity | Incisional |Itching |Numbness |Phantom Pain |Piercing |Pinching | Pins and needles sensation| Pleuritic | Pressure | Prickling| Pulsing | Sharp |Shifting |Shooting |Sore | Spasmodic | Squeezing | Stabbing | Stinging | Throbbing | Tingling| Burning | Jabbing | Patient unable to describe |Pounding |Tender |Tightness
Relative Temporal Context	Category	Post-operative/procedure | Pre-operative/procedure | During procedure | During activity | At rest
Pain Assessment Severity Scale Selection for Cascade	Category	**List of validated scales:** Numeric 0-10 Pain Scale | Baker-Wong Scale | Verbal Descriptor Pain Scale | Functional Pain Scale | CCPOT Scale | Adult NonVerbal Pain Scale | PAINAD Scale | rFLACC Scale | NPASS Scale | NIPS Scale | PIPP Scale | FPS-R Scale | Nociceptive Score | Simple Descriptive
Pain Severity Score [using validated scale]	Category	**Use Numeric 0-10 Scale Scores:** 0|1|2|3|4|5|6|7|8|9|10 **unless other scale from list below is required for patient population:** [Baker-Wong Scale | Verbal Descriptor Pain Scale | Functional Pain Scale | CCPOT Scale | Adult NonVerbal Pain Scale | PAINAD Scale | rFLACC Scale | NPASS Scale | NIPS Scale | PIPP Scale | FPS-R Scale | Nociceptive Score | Simple Descriptive]
Is Pain Relief Acceptable?	Boolean	Yes | No

Collins SA, Bavuso K, Swenson M, Suchecki C, Mar P, Rocha RA. **Evolution of an Implementation-Ready Interprofessional Pain Assessment Reference Model**. *AMIA Annu Symp Proc*. 2017:605-14.

Standard data definitions



Coded data: variation

Attribute	Value
Problem	**Severe Pain**

Attribute	Value
Pain	**Severe**

Attribute	Value
Severe Pain	Yes

Attribute	Value
Severe Pain	02-01-2001

Attribute	Value
Finding	**Elevated Sys BP**

Attribute	Value
Sys BP	**180 mmHg**

Attribute	Value
Sys BP	**Elevated**

Attribute	Value
Sys BP	**Abnormal**

What information needs to be modeled?

- <u>All</u> clinical information within an EHR:
  - Allergies
  - Problems
  - Orders
  - Test results
  - Medication administration
  - **Physical exam** and **clinical measurements**
    - Signs, symptoms, diagnoses
  - **Procedures**
  - **Family history**, **medical history**, and **review of systems**
  - **Clinical documents**

Focus on relevant clinical topics

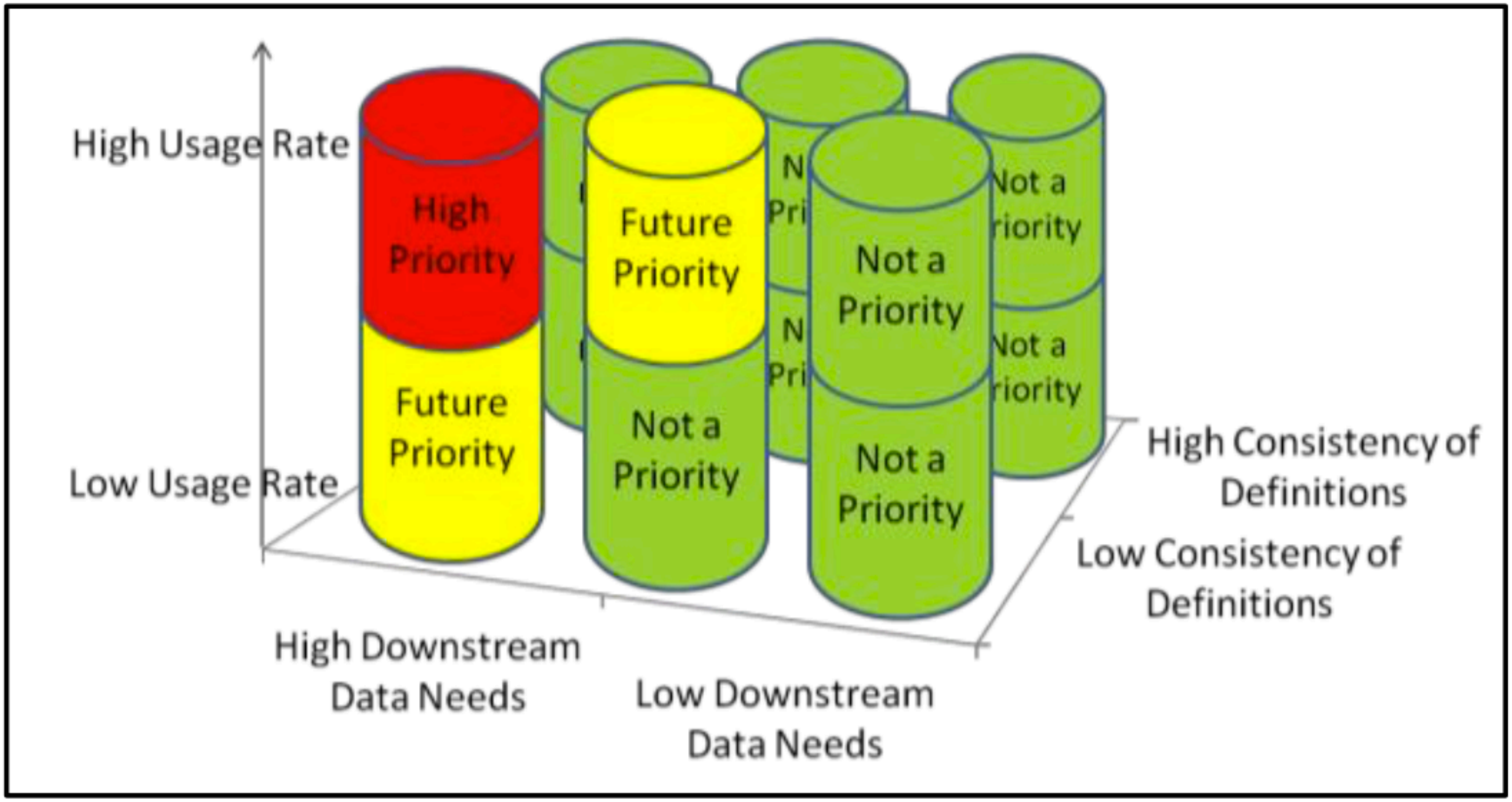


Figure 1. Criteria to Prioritize Clinical Topic Refinement

Collins SA, Gesner E, Mar PL, Colburn DM, Rocha RA. **Prioritization and Refinement of Clinical Data Elements within EHR Systems**. *AMIA Annu Symp Proc*. 2016:421-430

Acute Care Documentation (1/3)

- ***Publication***:
  - Collins SA, Bavuso K, Zuccotti G, Rocha RA. **Lessons learned for collaborative clinical content development**. *Appl Clin Inform*. 2013 Jun 26;4(2):304-16
- ***Context***:
  - Large strategic initiative back in 2007 to develop **standardized acute care documentation** (*bedside*) across two major academic medical centers: Brigham and Women's Hospital and Massachusetts General Hospital

Acute Care Documentation (2/3)

- ***Goals***:
  - **Highly structured documentation** to fulfill clinical needs, regulatory reporting, and data reuse
  - All **clinical disciplines** (e.g. nursing, medicine, social work, physical therapy, nutrition, occupational therapy)
  - Proactive **data standardization** in an effort to avoid ambiguity and duplication – e.g. naming convention for data elements, reuse of value sets, etc.
- ***Results***:
  - Over **11,000 data elements** defined, used in over **1,000 documentation templates** – e.g. initial patient assessments, progress notes, procedure and perioperative notes, event notes, transfer notes, discharge notes, assessment scales, flowsheets, etc.
  - Bedside documentation system was <u>not</u> implemented

Acute Care Documentation (3/3)

- ***Challenges***:
  - **Clinical** requirements **well understood** by stakeholder groups - easily gained traction when cited as a rationale for content development requirements
  - **Data engineering** requirements **not well understood** – formal processes to garner support and adherence
  - Limited resources, **expertise**, and competing priorities
- ***Lessons learned***:
  - Assess **knowledge needs** and set **expectations** at the start of the project
  - Define an accountable **decision-making process**
  - Increase team **meeting moderation** skills
  - Ensure adequate **resources** and **competency training** with online collaborative tools
  - Develop **goal-oriented** teams and consultative **service-based** teams

Large-scale EHR implementation (1/4)

- ***Publications***:

1. Collins SA, Gesner E, Morgan S, Mar P, Maviglia S, Colburn D, Tierney D, Rocha R. A **Practical Approach to Governance and Optimization of Structured Data Elements**. *Stud Health Technol Inform*. 2015;216:7-11
2. Gesner E, Collins SA, Rocha R. **Pain Documentation: Validation of a Reference Model**. *Stud Health Technol Inform*. 2015;216:805-9
3. Collins SA, Gesner E, Mar PL, Colburn DM, Rocha RA. **Prioritization and Refinement of Clinical Data Elements within EHR Systems**. *AMIA Annu Symp Proc*. 2016:421-430
4. Bavuso KM, Mar PL, Rocha RA, Collins SA. **Gap Analysis and Refinement Recommendations of Skin Alteration and Pressure Ulcer Enterprise Reference Models against Nursing Flowsheet Data Elements**. *AMIA Annu Symp Proc*. 2017:421-9
5. Collins SA, Bavuso K, Swenson M, Suchecki C, Mar P, Rocha RA. **Evolution of an Implementation-Ready Interprofessional Pain Assessment Reference Model**. *AMIA Annu Symp Proc*. 2017:605-14

6. Zhou L, Collins S, Morgan SJ, Zafar N, Gesner EJ, Fehrenbach M, Rocha RA. **A Decade of Experience in Creating and Maintaining Data Elements for Structured Clinical Documentation in EHRs**. *AMIA Annu Symp Proc*.2016:1293-1302

Large-scale EHR implementation (2/4)

- ***Context:***
  - **System-wide** vendor EHR implementation (2012-2017) – replace existing clinical systems
- ***Goals***:
  - Minimize (*resolve*) **inconsistent data definitions** across EHR applications and clinical settings, enabling and promoting **data reuse** and **interoperability**
  - **Practical** (*pragmatic*) approach to **governance** and **implementation** of structured data elements and reference models
    - Factors: resource allocation, implementation timeline, content refactoring, vendor best-practices, EHR limitations, etc.

Large-scale EHR implementation (3/4)

- ***Process*:**
  1. **Identify clinical topics** – align with strategic goals of the organization
  2. **Create draft reference model(s)** – find/consolidate/reuse models
  3. **Quantify downstream data needs** – reporting, regulatory requirements, clinical decision support, accurate billing, etc.
  4. **Prioritize clinical topics** – focus on high-value topics

  5. **Validate reference model(s)** – clinically accurate and complete
  6. **Quantify gap with EHR content** – prioritize revision/refactoring
  7. **Disseminate validated model(s)** – guide new content or revisions
  8. **Request revisions to EHR content** – change management process
  9. **Assess reference model utilization** – implementation and compliance

  10. **Monitor for new evidence** - revisions to reference model (*evergreen*)

Large-scale EHR implementation (4/4)

- ***Results***:
  - Data elements: **+15,000** (forms) and **+45,000** (flowsheets)
  - Dedicated workgroup: **+5** reference models (*discontinued*)
    - Pain Assessment: **47 data elements** organized into **9 data groups**
  - EHR system successfully implemented at all sites
- ***Challenges***:
  - Implementation timeline **incompatible** with the development of detailed reference models
  - EHR **processes** and **tools** not designed to promote detailed, consistent, and reusable data definitions across applications and modules
  - EHR content & data refactoring is an **iterative** process that requires **expertise** and **motivated** individuals

CONCLUSIONS

Challenges (1/3)

- Cost-effective **semantic interoperability**
  - Existing standards make data exchange possible, but not simple or efficient (projects take *months* or *years*)
  - Data exchanged in a structured and coded format still represents a small portion of the electronic record

Challenges (2/3)

- Clinical systems that can seamlessly represent and process a **complete electronic patient care record**
  - Current systems frequently rely on legacy data architectures that limit the use of clinical models
  - Slow adoption of new technologies that can overcome the current data representation limitations

Challenges (3/3)

- Clinical models with proper **domain coverage** and **extensibility**
  - Existing methods and tools to use clinical models and ontologies are not accessible to typical clinicians

Opportunities

- Government providing **exceptional incentives** for Healthcare IT adoption
  - IT identified as a key enabler of a more effective healthcare system
- Proposed healthcare delivery models require high levels of **integration** within and across institutions
  - Moving towards **seamless collaboration** where patients are active contributors
- Opportunity for a **new generation of clinical systems** beyond efficient record storage and communication
  - New paradigm with **pervasive** computerized **data analysis** and **decision support**
  - Widespread use of interoperable services and data, with advanced functions that enable **team-based care**

Conclusions: implementation

- Early engagement of clinical leaders to set expectations of technical **process**, **dependencies**, and **requirements**
- Provision of formal training about **informatics standards** and **governance processes**
- Establish a **data engineering team** with proper **authority** and robust **toolset** – guide implementation and ensure compliance with processes and standards

Conclusions: data engineering

- Establish **governance** for essential clinical domains
- Seek alignment with **standards**, maximizing interoperability and external collaborations
- Define and optimize **processes** (*lifecycle*)
  - Implement software **platform** integrated with knowledge **sources** and **consumers**
- **Monitor** & **evaluate** processes and resulting models
- **Collaborate** with other institutions to help amortize operational **costs** and promote **innovation**

Participate!

- Understand the scope and applicability of existing standards
  - Gain **access** to available standards
  - Confirm how each standard **applies** to your organization

- Contribute to and influence the development of standards
  - Bring your **specific needs** and discuss implementation options
  - Seek information from other stakeholders to make **informed decisions**
  - Most SDOs welcome open and **broad** participation
    - Healthcare providers, government stakeholders, payers, pharmaceutical companies, system vendors, consultants

- Achieve industry leadership by demonstrating interoperability
  - Learn about industry **best practices**
  - Understand implementation **timeline** and **costs**
  - Improve the **quality** and **sustainability** of your local systems

Acknowledgements

Saverio Maviglia

Sarah Collins

Karen Bavuso

Perry Mar

Li Zhou

Margarita Sordo

Members of the **Semedy** team

Beatriz Rocha

Stanley Huff

Thank you!

Roberto A. Rocha, MD, PhD
rarocha@bwh.harvard.edu
http://scholar.harvard.edu/rarocha

This work by Roberto A. Rocha is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License